PAPER *Special Section on Machine Vision Applications*

Interactive Object Recognition through Hypothesis Generation and Confirmation

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

SUMMARY 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 for objects recognition through speech. This paper presents a system that recognizes objects in occlusion and/or multicolor cases using geometric and photometric analysis of images. Based on the analysis results, the system makes a hypothesis of the scene. Then, it asks the user for confirmation by describing the hypothesis. If the hypothesis is not correct, the system generates another hypothesis until it correctly understands the scene. Through experiments on a real mobile robot, we have confirmed the usefulness of the system.

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

1. Introduction

Recently, helper robots or service robots which interact with humans in welfare domain have attracted much attention of researchers [1], [2]. We also have been developing such robot systems [3]–[6]. Our research goal is to realize the robots that can carry out tasks ordered by humans through verbal and nonverbal interaction. The tasks may include such as getting objects, operating appliances, and helping humans to move. Although we need research on mechanism, control and various other techniques to realize such robots, we would like to address computer vision problems in this paper.

We first consider tasks to get objects. Humans may ask the robot, "Get that book." In this case, the robot needs to detect the book in the scene by vision to carry out the task. Humans sometimes use simplified utterances such as "Get that," to ask orders. We have proposed a vision system to detect the object indicated by 'that' [7]. We assume that humans use such simplified utterances because the object is related to the actions of the speaker and/or the listener and the speaker thinks it unnecessary to mention the details. In the former case, the robot can use a priori knowledge about the object in its vision process, whereas it cannot in the latter. Even in the former case, however, the knowledge becomes useless if the robot fails to detect the object by using the knowledge. In fact, such failures often happen in vision systems. Thus, we address the problem of identifying the object without any a priori knowledge in this paper. The robot needs such a vision system as the last resort. The robot can use any other method to recognize objects. However, it may not always work. Welfare service robots should work every time even though it may take time, since their users may not be able to carry out the tasks by themselves. Thus, we are working on such a vision system as a basic component of the robots.

To compensate for failures in vision systems, we have proposed to use the human user's assistance through speech [3]–[5]. 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. The object recognition method proposed in this paper follows this approach.

There has been a great deal of research on robot systems understanding the scene or their tasks through interaction with the user [8]–[14]. These conventional systems mainly consider dialog generation at the language level. Moreover, most of them consider relatively simple scenes 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.

In the initial stage of our research [3]–[5], 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. However, in actual complex scenes, the vision system may detect various objects. The robot must choose the target object among them. We have tackled this problem in [6]. The robot determines the target through a conversation with the user. In that 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. Two most typical cases that break this assumption are occlusion and multicolor object situations. 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 have started to tackle these complex cases. We have re-

Manuscript received November 1, 2005.
Manuscript revised January 23, 2006.

[†]The authors are with the Deparment 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/e89–d.7.2197

ported our initial result in [15]. There, the robot examines a more detailed feature based on photometric invariance [16] in each ambiguous border of regions. If it can get any reliable information to judge about the border, it proceeds to the next. Otherwise, the robot asks the user whether the two regions are from the same object or different objects. We have confirmed that the system can work as we expected. However, the system has two problems. One is that it is not efficient since it examines each border, asking the user if it cannot get any definite information. It is not user-friendly to ask the user many times. The other is that the system may ignore segmentation failures. The segmentation decision on borders may not be correct even though the feature values strongly support the decision. This problem could be avoided if the robot asked the user for confirmation for all borders. However, this needs so many interactions, thus not being practical.

This paper proposes an interactive object recognition system to solve these two problems. The system considers a cluster of regions with multiple borders as a unit, generating the most probable hypothesis for the cluster. Then, the system asks the user for confirmation by describing the hypothesis. If the hypothesis is not correct, the system generates the next probable hypothesis. This process is iterated until the correct interpretation is found. Considering a cluster as a whole improves efficiency. Testing multiple hypotheses prevents from making segmentation errors. This framework of interactive object recognition is the main point of the paper. However, if the system cannot generate probable hypotheses in earlier stages, the efficiency decreases. To avoid this, we introduce another detailed feature, an intensity profile across the border, to examine ambiguous cases. This paper presents this additional feature and the object recognition system based on the hypothesis generation.

2. Previous System

This section briefly describes our previous system [15] since the system proposed in this paper is its extension and shares the basic framework.

The system first carries out image segmentation. We have proposed a robust approach of feature space method: the mean shift algorithm combined with HSI (Hue, Saturation, and Intensity) color space for color image segmentation [17]. This efficiently segments out specific color regions in different illumination conditions.

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 of the reflectance ratio.

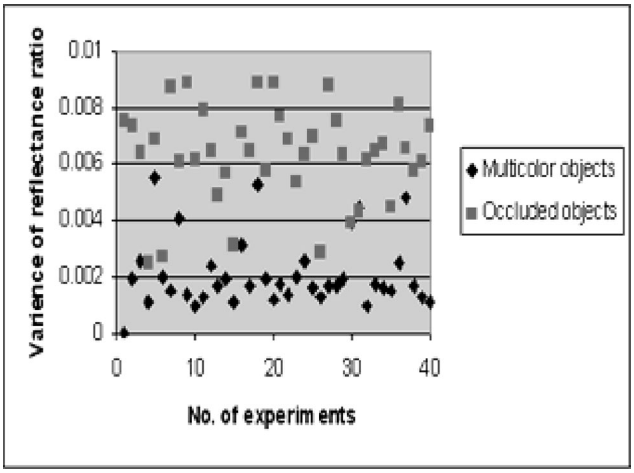


Fig. 1 Distribution of variances of reflectance ratio for multicolor and occluded objects.

We performed experiments to examine the usefulness of this measure. We measured the variance of reflectance ratio from 80 test images that were taken in different illumination conditions. The images consisted of 40 multicolor object cases and 40 occluded object cases. Figure 1 shows the result.

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

Case 1: $0.0 \leq \mathrm{Vr} \leq 0.002$: Two regions are from the same object.

Case 2: $0.002 < \mathrm{Vr} \leq 0.006$: Ambiguous case (cannot determine).

Case 3: $\mathrm{Vr} > 0.006$: Two 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. It asks the user for confirmation.

The system considers all pairs of adjacent color regions, checking all the borders. If there are many confusing borders in the image, it will ask many questions of the user. The system cannot be user-friendly. Moreover, the system proceeds to the next step without confirmation in Cases 1 and 3. Although the possibility is small, this decision may cause errors. This paper addresses these problems.

3. Intensity Profile for Geometric Shape Continuity Analysis

Before explaining the interaction method, we describe a new feature introduced in the current system.

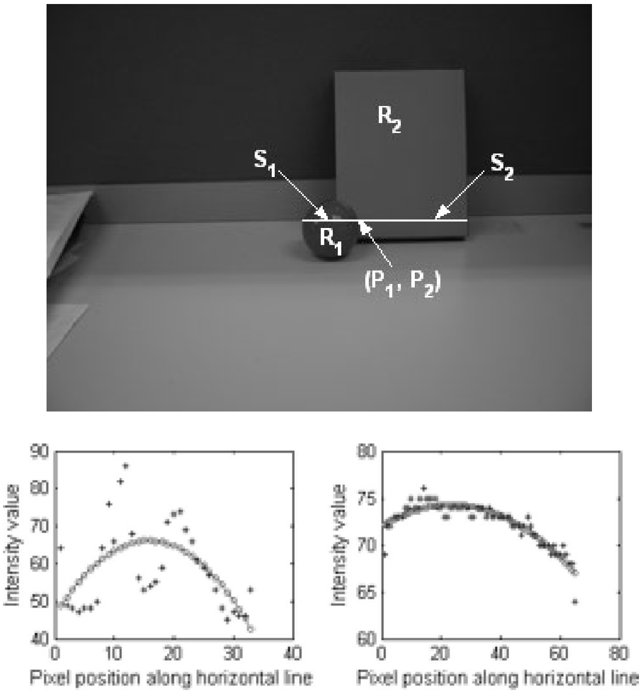


Fig. 2 Original image with the horizontal line in the middle of adjacent regions (top), the line profile of line segment S_1 (left) and the line profile of line segment S_2 (right).

As is well known in shape-from shading studies, intensity changes can give 3-D shape information. We propose to examine the intensity values over adjacent regions to determine whether or not the two regions come from a single object or different objects.

Rather than observing the intensity values over the regions, we reduce the problem to a simpler domain by analyzing the intensity values along the horizontal or vertical line crossing through both regions. In other words, we examine the intensity profile on the crossing line. To obtain the line profile for a region pair, we take the pixel (P_1, P_2) on the middle of the border of the adjacent regions R_1 and R_2 as shown in Fig. 2. We then use quadratic regression to fit straight lines or circular arcs to the intensity profiles for the line segments S_1 and S_2 that are passing through this point and crossing both regions.

For a complex scene containing non-uniform 3D objects, intensity profiles may have any degrees of complexity and their modeling is an elaborate task to do. However, for piece-wise uniform objects, we can 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). This parametric modeling is summarized as follows.

1. Straight line, $y = c$
2. Line with slope, $y = bx + c$
3. Curve, $y = ax^2 + bx + c$

where c is the constant term, b is the linear term, and a is the quadratic term. We compute parameters as:

$$c = mean(y), \quad \begin{bmatrix} b \\ c \end{bmatrix} = \frac{X_1}{Y}, \quad \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \frac{X_2}{Y} \tag{1}$$

$$X_1 = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ . & . \\ . & . \\ x_n & 1 \end{bmatrix}, \quad X_2 = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ . & . & \\ . & . & \\ x_n^2 & x_n & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \tag{2}$$

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 intensity 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. 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 the intensity profiles, we can examine the compatibility of models.

4. Hypothesis and Dialog Generation

4.1 Interpretation of 3-D World from 2-D Images

We have introduced two detailed features. The system uses them when it is not sure about the segmentation result. The system judges whether or not two adjacent regions form a continuous surface in the 3-D world from these two features obtained from 2-D image properties. If the surface is continuous, the variance of reflectance ratio may be small and the intensity profile may show a continuous line or curve. However, even if these two features are so in the 2-D image, we cannot guarantee that the surface is continuous in the 3-D world. There can be various accidental cases that the features take such values even if the surface is not continuous. On the other hand, there can be cases that the features do not show such values even if the surface is continuous. For example, if illumination condition changes across the object, the intensity profile may not be continuous. If there exist high light parts, the intensity profile cannot tell anything about the 3-D surface continuity.

The above consideration has led us to design our interactive object recognition system as follows. First, the system makes a hypothesis of a given image scene assuming that the judgment about the 3-D world made from the 2-D image features is correct. Then, the system explains the interpretation result to the user, asking him/her for confirmation. If the user's response is negative, the system modifies the hypothesis to meet the information given by the user.

Table 1 Decision by reflectance ratio.

Region 1	Region 2	Variance of reflectance ratio	Decision
R_1	R_2	small (≤ 0.002)	same object
R_1	R_2	large (> 0.006)	different objects
R_1	R_2	medium (0.002–0.006)	unknown

The same process is repeated with this modified hypothesis. Such processes are iterated until the system can get the user's confirmation. In summary, the system starts with the most probable interpretation. However, the system does not ignore possible cases that the system cannot obtain correct 3-D information from image features.

4.2 Hypothesis Generation

Our previous system deals with each border of regions independently. Since this is not efficient, the current system considers all the borders of regions forming a cluster at a time. The system generates a hypothesis of configuration of objects in the 3-D world in the following way.

Two adjacent regions can be parts of a single object or different objects. We check two adjacent regions for their compatibility of being from a single object or non-compatibility. The possible decisions about the two regions are:

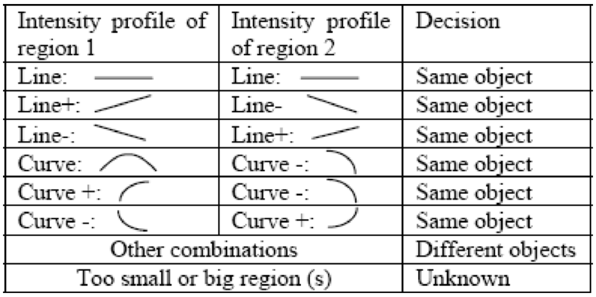
1. From the same object
2. From different objects
3. Unknown (cannot judge at this time)

We use two image features: reflectance ratio and intensity profile to judge the situation. If R_1 and R_2 are two adjacent regions, the use of reflectance ratio as a decision factor is shown in Table 1.

In unknown cases, the system further investigates the image. We use the intensity profile in addition to the reflectance ratio. We use quadratic regression to the intensity values along a horizontal or vertical line to fit a straight line or a curve. Then, we check the compatibility of being from a single object between the straight lines or curves for the adjacent regions. Figure 3 shows the rules of compatibility decision. Note that the intensity values across the border of adjacent regions can be different, not continuous even if the regions are from the same object. Thus, we check the compatibility by dealing with the intensity profiles in a symbolic way. The check of continuity here means to examine whether or not the shape patterns of profiles for both sides can satisfy one of the possible combinations that two regions from the same object can take as shown in Fig. 3. The intensity profile is used as a decision factor only in unknown cases by the reflectance ratio to reduce the calculation burden of the system.

4.3 Dialog Generation

The system, then, verifies the hypothesis through interaction with the user. The next problem should be what dialog the system will make.



Intensity profile of region 1	Intensity profile of region 2	Decision
Line: —	Line: —	Same object
Line+: ╱	Line- ╲	Same object
Line-: ╲	Line+: ╱	Same object
Curve: ⌒	Curve -: ◝	Same object
Curve +: ◜	Curve -: ◝	Same object
Curve -: ◟	Curve +: ◞	Same object
Other combinations		Different objects
Too small or big region (s)		Unknown

Fig. 3 Compatibility decision by intensity profile.

The basic way is to describe the hypothesis, that is, the current interpretation of the scene, by word. However, the user may not want to listen to a lengthy explanation of the scene. Thus, the system, first, tells the number of objects in the scene. If the user returns a positive response, the system describes the scene for confirmation. Even if the number of objects is correct, the segmentation by the system may not be correct. The system needs to change the hypothesis in this case. If the user says 'no' to the first statement about the number of objects, the system should also modify the hypothesis. Detailed explanations of these cases are given below.

4.3.1 Total Number is Wrong

If the user says 'no' to the system's question for the confirmation of the number of objects in the hypothesis, the system asks the user, "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 the image to adjust the hypothesis. The system examines the intensity profile for the region pairs that have not been examined before as they fall in Cases 1 or 3 in the analysis by the reflectance ratio. Regions of Case 1 decision are first examined in this case. This is because the variance of reflectance ratio can be small in various cases when the regions are from different objects, whereas it is relatively rare that the variance of reflectance ratio is large when the regions come from a single object.

4.3.2 Total Number Agrees but the Explanation Disagrees

Although the user says 'yes' to the number of objects, the user does not agree on the description of the scene. In this case, the system knows the number of objects. However, the initial hypothesis is wrong. The situation is the same as in the wrong number case after asking the number of objects. Thus, the system proceeds in the same way.

4.3.3 Unknown Cases

During the processes in the above cases (a) and (b), the system may find region borders where it cannot make any judgment from the intensity profile since the situation is the one shown in the last row of Fig. 3. The system may also find cases that its hypotheses were rejected by the user and that it cannot continue the recognition process. In such cases, if there are only a small number of regions, the system tells the user all possible interpretations of the scene, asking to choose the appropriate option. If the number of objects is large, the system adopts the last resort, taking the way used in the previous system. The system asks the user about each border where it cannot make a decision whether the two regions along the border comes from a single object or different objects.

In principle, if the system uses this method, the system can reach the correct interpretation although it may take much time. However, there is a serious problem in this interaction how the system let the user know the border currently under investigation. In the current implementation, we use a display to show the part. However, we would like to do this just by using speech and actions of the robot. We have started working on this problem by introducing the use of reference objects [18]. The system can show the part by mentioning the positional relation with a reference object already known or easily recognized by the user. For example, the robot may say, “I am not sure about the red and yellow parts in front of the blue object. Are they different objects?” This has not been implemented and is left for future work.

5. Experiments

5.1 Example Cases

We performed 80 experiments for various cases in different illumination conditions. Here, we show four typical example cases.

We use Pioneer 2 by ActivMEDIA as a robot (Fig. 4) in our experimental purposes. The main target objects are cups, cans, bottles, fruits, books, etc., on tables or shelves. 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.

Experiment 1: Robot’s initial hypothesis is correct

In the scene shown in Fig. 5, 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 obtains four connected regions, R_1, R_2, R_3 and R_4. To confirm which regions were parts of the single or different objects, the robot examines the value of the reflectance ratio of the adjacent regions.

According to the value of the reflectance (Fig. 5), 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 certain about the regions R_1 and R_3, because the value of the variance is in the range of Case 2. Thus, the robot examines the intensity profile along the horizontal line. From the result, the robot is certain that the regions are parts of different objects. Then, the robot asks the user for confirmation.

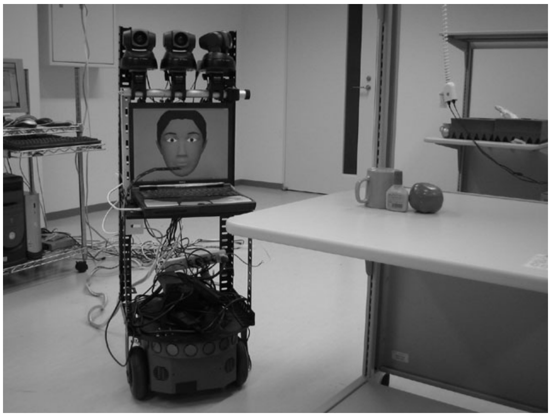


Fig. 4 Robot used in the experimental purposes.

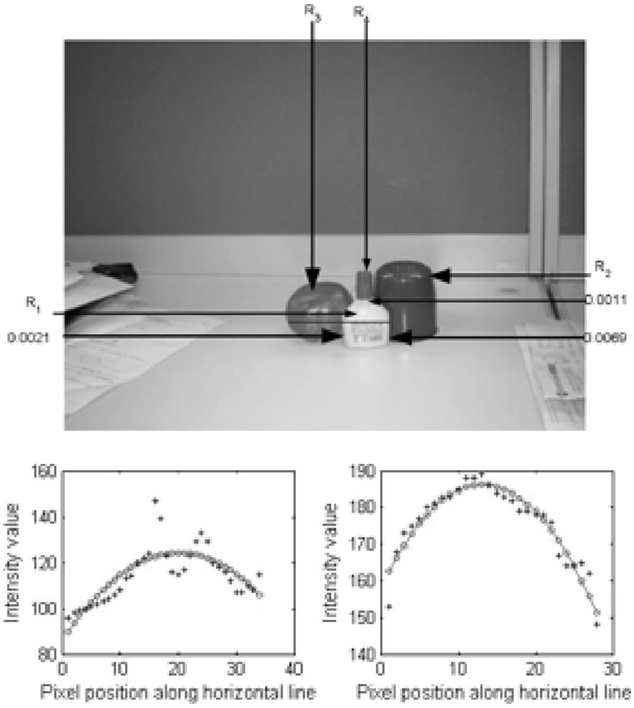


Fig. 5 Image containing single color, multicolor and occluded objects (top), Intensity profile of region R_3 and R_1 (left-bottom and right-bottom).

Robot: “Are there three objects?”
User: Yes.
Robot: “Are there one red, one blue and one multicolor object containing red and yellow parts?”
User: Yes.

Experiment 2: Robot’s initial hypothesis is wrong

Figure 6 shows six regions R_1, R_2, R_3, R_4, R_5, R_6. 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. Regions R_1 and R_2 are parts of the same

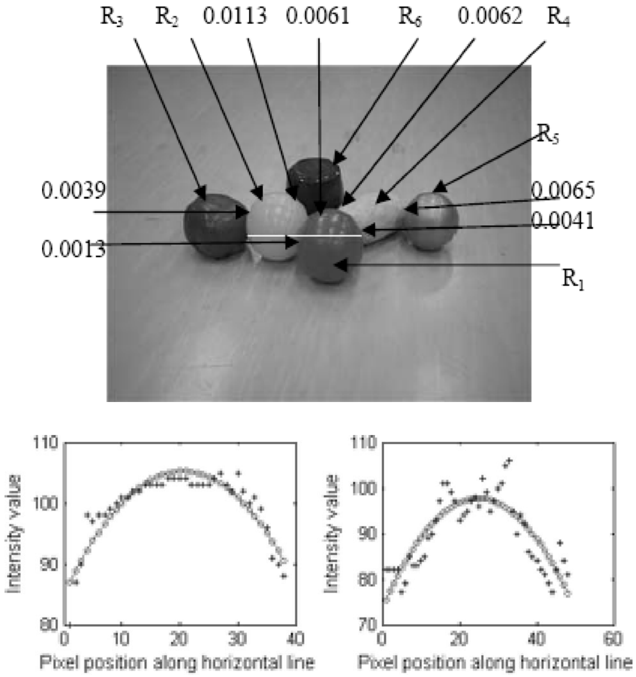


Fig. 6 Image containing single color occluded objects (top), Intensity profiles of regions R_2 and R_1 (bottom left and right).

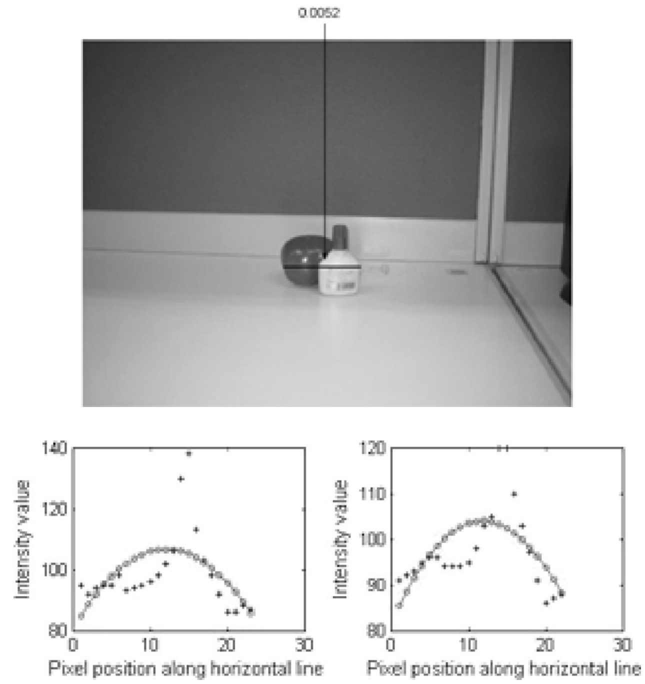


Fig. 7 Occlusion case where parts of two objects are merged into one region with intensity profile.

object. However, the robot is not sure about the region pairs (R_2, R_3), and (R_1, R_4). After investigating the intensity profile along the horizontal lines for the region pairs (R_2, R_3), and (R_1, R_4), it is sure that the regions are parts of different objects. Then, the robot asks its user for confirmation.

Robot: "Are there five objects?"
User: No.
Robot: "How many objects in the scene?"
User: Six.

As the robot's initial assumption is wrong, it should reinvestigate to adjust the result. Although the test using reflectance ratio has a strong ability to measure the compatibility of regions, it may only show which regions are definitely not compatible. Thus, all region pairs that have lower values 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), the system is sure that the regions are parts of different objects. The robot asks the user for confirmation:

Robot: "Are there two red, two yellow, one blue and one green objects?"
User: Yes.

Experiment 3: Robot's initial hypothesis about the total number is correct but segmentation is not correct
In the occluded object case shown in Fig. 7, two regions, yellow and red are found after initial segmentation. Since the variance of the reflectance ratio in the region boundary is 0.0052, the robot needs further image investigation. From the analysis result of the intensity profile, the robot makes a hypothesis that there are two single color objects. Then, the robot asks:

Robot: "Are there two objects?"
User: Yes.
Robot: "Are there one red and one yellow object?"
User: No.

The initial assumption of the system is rejected by the user. The system cannot continue the recognition process without the user's assistance. Since there are only a small number of regions (two) in the scene, the system asks the user to choose the appropriate one from all possible combinations of the regions.

Robot: Choose the combination:
A: One yellow and one multicolor.
B: One red and one multicolor.
C: Two multicolor objects.
User: B

Now the robot understands the situation. However, in the current implementation, the robot cannot extract the correct border between the red object and the red part of the multicolor object. If the task of the robot is to get the multicolor object, the robot can carry it out by grasping the yellow part. However, the robot may need to obtain the correct borders to complete some other tasks. This is left for future work.

Table 2 Comparison experiments.

Exp. No.	Total objects in scene	Previous system			Present system		
		No. of interactions	Objects found	Status	No. of interactions	Objects found	Status
1	3	1	3	ok	2	3	ok
2	6	2	5	Error	3	6	ok
3	2	1	Unknown	Error	3	2	ok
4	2	1	2	ok	2	2	ok

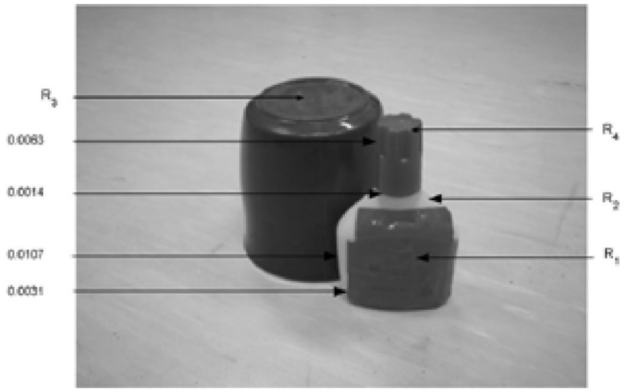


Fig. 8 Image containing single color, multicolor and occluded objects.

Experiment 4: Robot cannot conclude

The robot obtained four connected regions after applying the initial segmentation technique as shown in Fig. 8. To confirm which regions are parts of a single or different objects, the robot examines the variances of the reflectance ratio of the adjacent regions. From the result, the robot concludes that both region pairs (R_2, R_3) and (R_3, R_4) are parts of different objects. Regions R_2 and R_4 are parts of the same object. However, the robot cannot tell the relation between the regions R_1 and R_2. The robot needs further image investigation. However, the region R_2 is too small for analyzing the intensity profile because such profile data along a short segment cannot be reliable. Thus, the robot asks the user for help about the border that it cannot determine in the following way,

Robot: "Are those regions parts of the same object?"
User: Yes.
Robot: "Are there one blue and one multicolor object containing two red and one yellow parts?"
User: Yes.

From the user's answer to the first question, the robot confirms that regions R_1 and R_2 are parts of the same object. Then, the robot concludes that there are two objects, one single color and another multicolor.

However, in complex cases like the above, the user may not know which part the robot is talking about as mentioned in the last paragraph of 4.3. The robot should make this clear to the user. 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.

Table 3 Experimental results.

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

5.2 Comparison Experiments

The proposed system has been designed so that it will not make errors even in such cases that the previous system may fail. We performed comparison experiments to prove this. Table 2 shows the recognition results by the previous system and the proposed system for the four example experimental cases described above. The proposed system can work appropriately in the scenes where the previous system makes errors. The proposed system first tells the number of objects. Since we count this as another interaction, the number of interactions for the proposed system is larger in these experiments with a small number of objects.

5.3 Experiments about Efficiency

The proposed method is expected to reduce the user's burden 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 3 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.

5.4 Current Status, Limitation, and Future Work

As mentioned in the introduction, we have been working on interactive object recognition and have been extending the situations that our systems can manage. We started to deal with a few objects in the scene [3]–[5], then with multiple objects [6]. We first considered only single color objects in no occlusion scenes [6]. Then, we have started to research

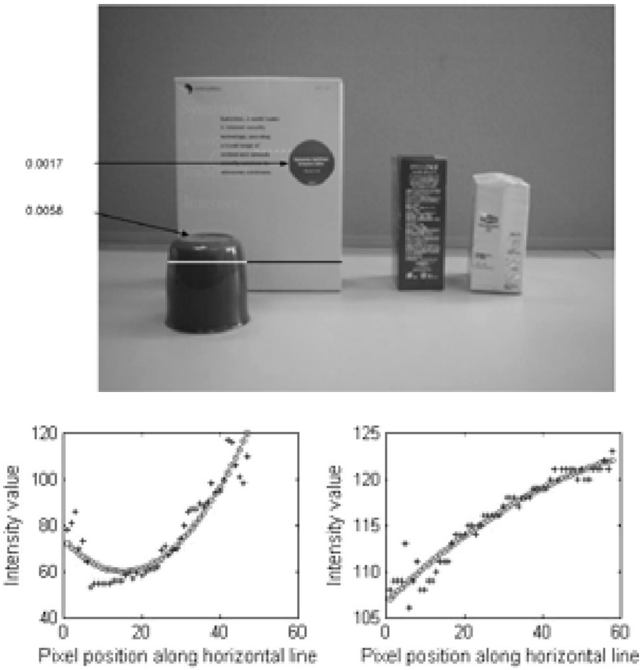


Fig. 9 Objects with small marks (top). The intensity profile along the white line on the blue object (bottom left) and that along the black line on the left yellow object (bottom right).

on multi-color objects in occlusion scenes [15]. In this paper, we have proposed a more efficient and user-friendly framework for the same level situations. Up to this research, all our systems start with color segmentation. Thus, they cannot treat textured objects. We need to introduce texture segmentation. This is left for future work.

We have designed the proposed system to deal mainly with objects composed of a small number of color parts as shown in Figs. 5 to 8. Such objects might be considered too simple for actual situations. However, various actual objects can be in this object category if the system ignores small regions. Figure 9 shows an example. After image segmentation, post processing and ignoring small regions, there remain five regions. Two are separated and three are connected. Using reflectance ratio and intensity profile, the system assumes that there are four objects. This hypothesis is confirmed by the user.

In theory, the proposed system can deal with objects even with many marks on them. In practice, however, the system cannot be usable in such cases if the initial hypothesis based on the current two feature analysis methods is not correct. The system may need a great number of interactions to reach the correct interpretation. The current status of research and the point of this paper are to devise an efficient and user-friendly interaction framework. We have prepared two feature analysis methods to show the usefulness of the framework. Although the methods can work effectively in the framework, we cannot guarantee that they can work in all occasions. In fact, it is the start point of our research that any vision method cannot work perfectly all the time. Thus, vision systems need interaction with humans. Still, the system needs to obtain a correct initial hypothesis with high probability to be practical. Both current features are based on photometric properties. To improve the capability of the system, we are planning to examine contour shape features as well as small textures.

Research on learning is also left for future work. Simple one is to adjust decision parameters. We use the decision rule shown in Table 1 obtained from the experiments in the reflectance ratio analysis. This should be modified if the hypotheses based on this are often rejected by the user. More serious one is to learn object names and to recognize them when their names are mentioned. As mentioned in the introduction, the purpose of the current research is to provide the basic function for welfare service robots. The system does not have any a priori knowledge, and the vocabulary that can be used in interaction is limited to that for attributes of objects, such as color, shape, size, and position. It is certain that interaction using object names is easier for humans. Humans may say object names while using the proposed system. The robot may be able to associate such names with the objects detected through interaction. Later, it can recognize the objects without interaction when the user mentions their names. This is an interesting next research topic.

6. Conclusion

We have proposed an interactive object recognition system for helper robots. The system makes a hypothesis of the scene and asks the user for confirmation. If the user's response is negative, the system makes another hypothesis. The system iterates this process until it can obtain the user's confirmation. The proposed system represents the scene as a set of regions each of which may correspond to an object. The interaction between the user and the system is performed at this object-region level by describing the attributes such as color and shape of the regions. It is needless to say that the most convenient way for humans is to specify objects by their names. In this sense, the proposed system, which forces the user to communicate by the attributes of objects, does not seem to be user-friendly. However, considering the fact that helper robots should work all the time, we need such a vision system that can work under various conditions even though it may take time to accomplish the task.

We are currently working on a layered object recognition architecture. The highest level layer is an object recognition module at the object-name level. The proposed system can be situated in the bottom layer and support the total system when the higher level modules fail. Our previous system [15] could serve this purpose. However, it interacts with the user at the level of the borders between object regions. As experimental results show that the proposed system is more user-friendly and efficient. Actually, the proposed system uses the previous system when necessary. The idea of the layered architecture is adopted in the current system.

Acknowledgment

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

References

[1] M. Ehrenmann, R. Zollner, O. Rogalla, and R. Dillmann, "Programming service tasks in household environments by human demonstration," 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] 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.

[4] 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, pp.1308–1313, Lausanne, Switzerland, Sept./Oct. 2002.

[5] 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, NV, Oct. 2003.

[6] 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.

[7] Z.M. Hanafiah, C. Yamazaki, A. Nakamura, and Y. Kuno, "Human-robot speech interface understanding inexplicit utterances using vision," Extended Abstracts, Conference on Human Factors in Computing Systems, pp.1321–1324, Vienna, Austria, April 2004.

[8] 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.

[9] 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 Inteligent Systems, pp.323–328, Tokyo, Japan, 2003.

[10] P. McGuire, J. Fritsch, J.J. Steil, F. Roothling, G.A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter, "Multi-modal human machine communication for instruction robot grasping tasks," Proc. International Conference on Intelligent Robots and Systems, pp.1082–1089, Laussane, Switzerland, Sept./Oct. 2002.

[11] 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.

[12] A. Cremers, Object Reference in Task-Oriented Keyboard Dialogues, Multomodal Human-Computer Communication: System, Techniques and Experiments, pp.279–293, Springer-Verlag, 1998.

[13] T. Winograd, Understanding Natural Language, Academic Press, New York, 1972.

[14] 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.

[15] M.A. Hossain, R. Kurnia, A. Nakamura, and Y. Kuno "Interactive object recognition system for a helper robot using photometric invariance," IEICE Trans. Inf. & Syst., vol.E88-D, no.11, pp.2500–2508, Nov. 2005.

[16] S.K. Nayar and R.M. Bolle, "Reflectance based object recognition," Int. J. Comput. Vis., vol.17, no.3, pp.219–240, 1996.

[17] M.A. Hossain, R. Kurnia, A. Nakamura, and Y. Kuno "Color objects segmentation for helper robot," Proc. International Conference on Electrical and Computer Engineering, pp.206–209, Dhaka, Bangladesh, Dec. 2004.

[18] R. Kurnia, M.A. Hossain, A. Nakamura, and Y. Kuno, "Using reference objects to specify position in interactive object recognition," Proc. International Conference on Instrumentation, Communications and Information Technology, pp.709–714, Bandung, Indonesia, Aug. 2005.

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., M.S. and Ph.D. degrees 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.