Interactive Reference Resolution for Service Robots

Tajin Rukhsana Tarannum and Yoshinori Kuno

Saitama University, Japan

{tajin,kuno}@cv.ics.saitama-u.ac.jp

Abstract Carrying out user commands entails target object detection for service robots. When the robot system suffers from a limited object detection capability, effective communication between the user and the robot facilitates the reference resolution. How human users, being aware of the poor object detection capability of their robot partner, describe objects in 2D images is of primary interest. A survey is conducted in this regard. Results show that color and spatial relation among objects are the mostly used attribute by the participants. Among several aspects of communication yielded from the experiment results, we have chosen “Feedback generation” to implement into our robot system. An algorithm has been constructed for this purpose. The role of robot and human in discourse are also modified from our previous work. To implement the system, first we detect foreground objects (blob). Object recognition takes place simultaneously. Then, the user initiates communication through speech. Based on information provided by the user, our robot system continues to generate feedback and queries, which lead to detection of target object.

1 Introduction

We aim to develop a service robot, assisting people inside home or small office environments where, most of the user requests are directly or indirectly linked to some objects in the scene. To carry out the user command, the robot has to locate the objects first. Object searching has been addressed in previous research. A global image representation that provides relevant information for place recognition and categorization is presented in [1]. Such contextual information simplifies object recognition. Global image features have been shown to benefit object search mechanisms while providing an efficient shortcut for object detection in natural scenes [2]. In [3] the authors have built up a multi-class and multi-view object detection mechanism by sharing the common features across classes.

While the aforementioned research uses statistical approaches to detect objects, we are interested in discourse-based object detection. Verbal input and natural language understanding are described as indispensable parts of human-robot interface in [4]. Communication between human and robot through dialogue is also used in [5]. The motivation behind our research is to generate an interactive object detection model in the presence of some known (can be detected by robot) objects. Our primary interest is to observe how humans describe a target object to a robot that can recognize very few objects in the scene. Previous research addresses various issues of referring behavior of humans. Mutual responsibility of participants in the making of a “definite reference” is addressed in [6], whereas [7] describes how “Grounding” gets shaped. But these works refer to general referring behavior in conversation. We could not locate any previous work specifically in our interest area, i.e. object description using attributes.

In need of observing linguistic preference of humans, a survey has been carried out. The result suggests that color and spatial relation among objects are preferred almost equally to describe objects in 2D images. Besides this, allowing the user to provide information about the target object at the beginning reduces the length of conversation considerably. This finding helps modify the conversation approach taken in our previous work. In [8][9][10], we assumed the scene to consist of simple objects in simple background. The user response was restricted to replies of "what" and "yes-no" type questions. Moreover, no kind of object detection effort was made. In [11][12] object detection technology has been appended. A detailed range of user input has also been allowed. In [12] 4 cases of interaction have been described in the model. The first 3 cases, however, consider that object model of the target is found in the database. In this research, we take case 4 into account. Moreover, we have extracted another two important features of conversation, "Feedback generation" and "Error correction". In this paper, we suggest an algorithm to incorporate the two features (both referred to as "Feedback generation" as a single identity) into the framework of dialog generation.

2 Experiment Basics

The participants are non-native speakers of English and graduate students of Saitama University, Japan. To receive user input, a Visual Basic program is developed. All input is saved as text. A pair of participants sits in front of the screen. One plays the role of a robot (referred to as "Robot") and the partner plays the role of a human user (referred to as "Human"). With a shared view of an image where there are several objects, the challenge of the Human is to describe a target object to the Robot partner providing efficient and effective hints. Through conversation, the Robot attempts to detect the target object as soon as possible, using those hints. Both participants only input text in English. No oral input is allowed. The only gesture input permitted for the Robot is to point to an object on the screen, which he thinks the target. Since the experiment was not videotaped, the object pointed at was recorded by name.

As mentioned earlier, in our research a Robot system is assumed to have limited object recognition capability. To help both participants restrict their choice of words during the conversation, they are informed beforehand that the Robot will pretend to have a very limited knowledge about the objects in the environment. So in a 2D image, the Robot is supposed to be able to recognize only two types of objects. The cognitive and linguistic database of the Robot is limited to names of these known objects and basic properties of objects i.e. color, shape, size and positional relationship among objects in an image. For a total of 15 images, 45 pairs of participants engaged in conversation i.e. 3 pairs for each image. To obtain a description of objects in different situations and different orientation of objects, 15 images were chosen. The number of participants was 20 and some of them were involved in trial more than one time with the same or a different partner. For each image we decided the known and target objects. Known objects were reported to both users so that only these objects are used for reference during conversation. The identity of the target object was shown only to the Human in written form.

We made a list (Table 1) of exemplary descriptive words for object properties and positional relationship. Both participants were shown this list before starting the conversation. This was done not to bias their word choice, but to help them have a clearer view of the range of possible descriptions.

Table 1: Example of candidate words for description

Category	Examples
Color	Red, Yellow and Green, Light Blue etc.
Shape	Rectangular, Round, Cylindrical, Square etc.
Positional relation	Left, Front, Near, Middle, Front-right etc.
Size	Big, Small etc.
Superlatives, Comparatives	Rightmost, Nearer, Smallest etc.

3 Rules for Encoding Dialogs

To analyze text input of users, we did not use any state-of-the-art language-processing tool. Instead, our own simple strategy was followed. At first we encoded the dialogs with symbols used for four primary categories of object attributes; color (C), shape (Sh), posiotnal relation (P) and size (Si). Initial reporting of known objects by the Robot is excluded from encoded data. No other word except those already mentioned, was included in encoding. To count the frequency of mentioning any specific category (Color, Shape, Size, Positional relation) in a conversation, we just calculated the frequency of symbol used for that category.

4 Results

The conversations between 45 pair of Robot and Human can be discussed from various viewpoints.

4.1 Proportions and Vocabulary

A total of 198 descriptors were mentioned by the participants; 137 by the Human and 61 by the Robot. This is not surprising because the Human contributed to the conversation describing the target object, notably more than the Robot. Among all words, color was the one mentioned most often. The following is positional relation. Percentage usage of attributes is color (38%), shape (20%), position (35%) and size (7%). Both Human and Robot, across all trials, were consistent in using color and positional relations almost equally.

Along with standard colors such as red, green, yellow etc., some modifiers like light, deep, almost, -ish etc. were used. For multicolor objects, users mentioned all prominent colors when needed. Words used for size and shape across all trials are shown below:

Property	Words used
Size	Big, small, medium, short, long, half, large, thin
Shape	Rectangular, round/circular, flat, square, cylindrical, triangular

Superlatives and comparative forms of size were also used. In response to a Robot question, “Is it the highest one?”, the Human answered, “Half of the highest”. The word “half” was used only in this case among 45 trials. Users found it easier to use rectangular and round than other shapes. For an object having two parts of two different shapes and colors, one user mentioned the shapes separately. This also was the only instance of using the word ”triangular”. For positional relations, there were a wide variety of words used (Table 2).

Table 2 Vocabulary for different positional relations

Type of reference	Vocabulary
Relative, Pr	Left, right, in front of, behind, near/close/nearby, far, middle/ between,
Group based, Pg	Middle/center, leftmost, rightmost
Directional, Pd	South-west
Image plane as a reference, Ps	Left upper corner, leftmost, center/middle

We found one instance of using the modifier “little” with “far”. If we compare the use of relative terms of three groups, left/right (29), near/far (19) and front/behind (11), the first group was the mostly used one. The proportions of using different types of reference are Pr 80%, Pg 12%, Pd 2% and Ps 6%.

4.2 Necessary information provided by Human at the beginning

In [8][9][10], robot leaded the conversation with user, generating efficient queries. In order to do this, huge primary manipulation was required for scene understanding. In our experiment, however, we have instructed the Human to provide as much information about the target object as they think useful at the start of the conversation. We have seen throughout the experiment that Human users could describe attributes of target objects avoiding unnecessary details. Using the information obtained from the Human at the very beginning, the Robot has been able to squeeze the horizon of candidate objects in an efficient manner. Thus, not only was the Robot relieved from the primary manipulation, but also the generation of subsequent queries was easier. Here we give an example. In image 11 (Fig. 1), the Robot is supposed to be able to recognize “teapot” and “cup” (denoted by O). The target is “marker pen” (denoted by X). The dialogs to detect the target in this image are depicted below.

Robot: I can see teapot and cups
Human: The object is in front of cup.
Robot: Which cup?
Human: Rightmost cup.
Robot: Nearer to that cup?
Human: Yes

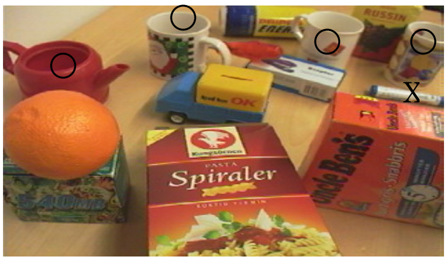


Figure 1: Known and target object in Image 11

The Robot then pointed to the intended target object. Here we see that the Robot did not have to decide which way he should ask for information about the target. Rather, the Human provided positional information, which he regarded as the best strategy to describe the target.

4.3 Error correcting strategy

This experiment reveals that in a conversation mutual agreement between partners about any physical property of objects, plays an important role in identifying the target. Some examples can be given in this regard. In one trial, the Human described a target object as “purple”. But, the Robot did not find any purple object in the scene and he reported it to the Human. Consequently, the Human inferred that what he thinks of as “purple” was not actually the same for his partner. He then relied on another property to describe the object. In another trial, the Robot agreed to all other properties except color for the target object. He considered it to be black, which was described as gray by his partner. To remind the partner of the possible mistake, the Robot made a query as follows:

Robot: Is it gray or black?
Human: Deep gray, not black.

Although the Human still sticks to “gray” rejecting the possibility of “black”, the word “deep” helps the Robot infer that the black object in his mind is being referred to as “deep gray”. During a trial, an object was described as at “southwest” position. From other indicators it was clear to the Robot that “southeast” should be the description. But instead of making his partner aware of the mistake, he continued to look for the objects at southwest position and made subsequent queries. The trial was not, as can be easily understood, an efficient one in terms of length of conversation. After the trial, the participant who played the role of Robot told that, he could detect the target object at the very first moment. We see here that, correcting an error or at least asking the partner about alternative choices, as soon as the error is revealed, makes the reference resolution efficient and reduces further complexity.

4.4 Feedback from partner

Throughout the 45 trials, there are various occurrences of feedback from both users. We summarize below the strategies, adopted to provide feedback.

Strategy 1: The robot finds several candidate objects based on the description and generates group-based query to elicit the target.

Strategy 2: Robot reports that no object of a given description is found.

Strategy 3: Expressing inability to describe a specific property of an object or to resolve a given property.

Strategy 4: Asking about specific choices for a property in order to reduce the number of candidates.

Strategy 5: Asking for more information when the robot finds the given description insufficient

5 Methodology for Interactive Target Object Localization

We propose here a methodology for interactive object detection (Fig. 2). The input image is first processed to detect blobs [13]. Some blobs may be recognized as objects using methods described in [12]. After the user's command input, the system decides the target object. If the target object is among the recognized objects, the robot asks for confirmation. There may be more than one instance of the object in the scene. So, the robot tries to confirm which one the target is among the group of objects using color, size, position or other information. Answer may be the biggest, middle one, the nearest, the leftmost etc. If the target object is unknown, the robot reports to user what it can see in the image. Then, robot asks the user for a description of the target. Based on the description it then removes candidate objects through dialog generation and finally confirms the detected object as target, given user feedback. This is a proposed model and we have not yet implemented all modules of it, except steps 1 to 4. The result of blob detection and object recognition (Fig. 3) is shown. Manipulation of spatial relations has been implemented, but not on an interactive mode.

Step 6 in Fig. 2 reflects our findings about "Feedback generation" from the survey. All consisting steps of it are described below in details:

6.1. Receive user input and decide number of attributes used in it.

6.2. If one attribute is used, find whether it is positional relation or not. Otherwise, go to step 6.4.

<u>6.2.1 Positional relation, P</u>

If Pr is used, find reference object. Otherwise, decide the region where objects will be searched later. Then, find objects matched with given positional information.

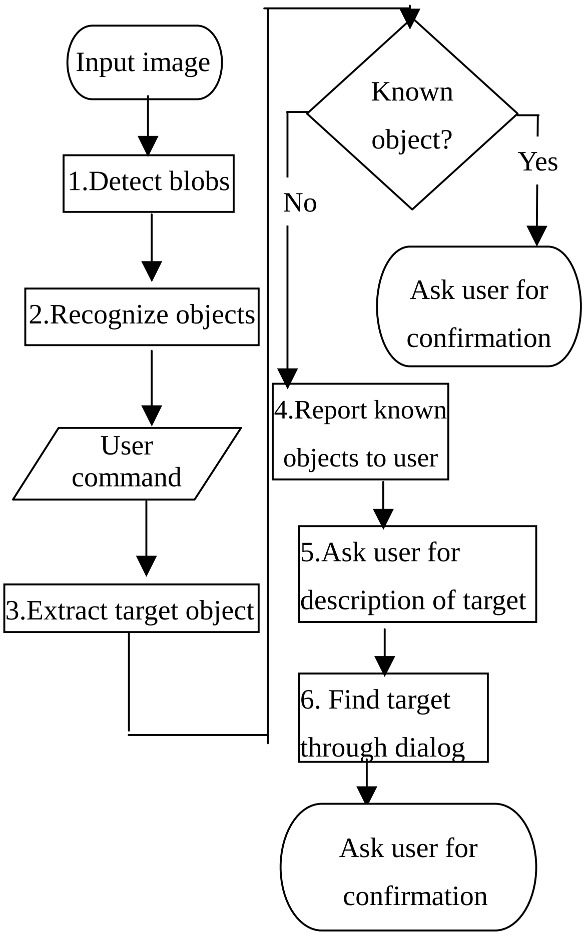
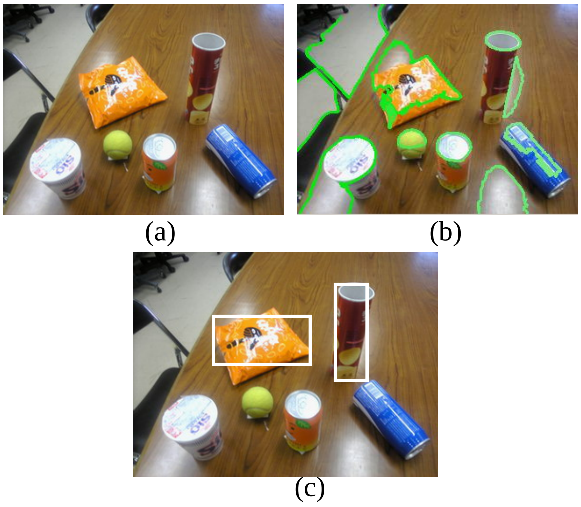

Input image
1.Detect blobs
2.Recognize objects
User command
3.Extract target object
Known object?
No
Yes
Ask user for confirmation
4.Report known objects to user
5.Ask user for description of target
6. Find target through dialog
Ask user for confirmation

Figure 2 Overview of the system


(a)

(b)

(c)

Figure 3: (a) Original image (b) Blob detection separates regions (c) Objects recognized using [12]

<u>6.2.2 Attribute other than P</u>

Find objects of the given attribute.

6.3. Decide how many objects found which match the given attribute.

6.3.1 One object found

Show the user and asks for confirmation.

6.3.2 More than one object found

Consider the objects that match given attribute as a group and ask user to tell which one the target is, among these objects. Upon user input, find the target and asks user for confirmation.

6.3.3 No object found

Report to user as “Not found” and request for more information. Repeat from step 6.1.

6.4. Decide whether the attributes include positional information or only are the combination of color, shape and size. If they include positional information go to step 6.6.

6.5. Decide whether any object satisfying all given attributes is found.

6.5.1 Objects found matching all given attributes

If there is one such object, execute step 6.3.1. Otherwise, execute step 6.3.2.

6.5.2 No object found that match all attributes

6.5.2.1 Report the attribute, which is not matched, to the user.

6.5.2.2 Keep a record of all objects that satisfy some attributes.

6.5.2.3 Then ask the user to describe position of the target in relation to any of the known objects.

6.5.2.4 Find objects using both 6.5.2.2 and 6.5.2.3. Execute step 6.3.

6.6. Execute step 6.2.1. Use attributes other than positional relation to reduce number of candidates from the result of step 6.2.1. Then execute step 6.3.

6 Conclusion

In this paper we have described the results of a survey that finds linguistic choice of humans in describing objects. A method has been proposed to detect objects in images through interaction with robot. We leave some modules of the method to implement in future. We also plan to carry out a similar survey in which the user, who plays the role of Robot, will lead the conversation by asking about attributes of target objects. The reason behind this is that, we want to find which attributes are chosen, to use in a query, in which situation.

References

[1] Torralba A. Murphy K.P., Freeman W.T., Rubin M.A.: Context-based Vision System for Place and Object Recognition, Proc. of IEEE Intl. Conference on Computer Vision, vol.1, pp.273-280, 2003.

[2] Torralba A., Oliva A., Castelhano M. S., Henderson J. M.: Contextual Guidance of Eye Movements and Attention in Real-world Scenes: The role of global features on object search, Psychological Review, 113, pp.766-786, 2006.

[3] Torralba A., Murphy K. P., Freeman W.T.: "Sharing Visual Features for Multiclass and Multiview Object Detection," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.29, no.5, pp.854-869, 2007.

[4] Fong T., Kunz C., Hiatt L. M., Bugajska M.: Using Vision, Acoustics, and Natural Language for Disambiguation, Proc. of ACM/IEEE International Conference on HRI, pp.73 – 80, 2007.

[5] Fransen B., Morariu V., Martinson E., Blisard S., Marge M., Thomas S., Schultz A., Perzanowski D.: The Human-Robot Interaction Operating System, Proc. of ACM conference on HRI, pp.41-48, 2006.

[6] Clark H. H., Wilkes-Gibbs D.: Referring As a Collaborative Process, Cognition, 22, pp.1-39, 1986.

[7] Clark H., Brennan S.: Grounding in Comm Perspectives on socially shared cognition, pp.127-149, 1991.

[8] Kurnia, R., Hossain, M.A., Nakamura, A., Kuno, Y.: Generation of Efficient and User-friendly Queries for Helper Robots to Detect Target Objects. Advanced Robotics 20, pp.499‐517, 2006.

[9] Hossain, M.A., Kurnia, R., Nakamura, A., Kuno, Y.: Interactive Object Recognition through Hypothesis Generation and Confirmation. IEICE Trans. on Info. and Sys. E89-D, pp.2197-2206, 2006.

[10] Kurnia, R., Hossain, M.A., Nakamura, A., Kuno, Y.: "Use of Spatial Reference Systems in Interactive Object Recognition," Proc. Of CRV, p.62, 2006.

[11] Mansur, A., Kuno, Y: Integration of Multiple Methods for Robust Object Recognition. SICE Conference, ‘07.

[12] Mansur, A., Sakata K., Kuno, Y.: Recognition of Household Objects by Service Robots Through Interactive and Autonomous Methods, Proc. of ISVC(2), pp.140-151, 2007.

[13] Lindeberg T.: Feature Detection with Automatic Scale Selection, International Journal of Computer Vision, 30 (2): pp.77-116, 1998.