MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan

8-23

Object Recognition Using Environmental Cues Mentioned Explicitly or Implicitly in Speech

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

Abstract

The service robot that carries out tasks ordered by the users through speech needs a vision system to recognize the objects appearing in the orders and a speech interface for natural communication with the user. The user's order may be explicit or implicit. The speech interfaces should have a capability of dealing with explicit as well as implicit utterances. In this paper, we present 'how environmental cues help in understanding user's orders.' We assume that humans usually put a particular object on a small number of places in the environment. Using the environmental knowledge, the robot can efficiently understand and accomplish the user's demand with less vision task and user burden.

1. Introduction

Helper robots or service robots in welfare domain have attracted much attention of researchers for the coming aged society [1][2]. Multimodal interfaces [3][4][5] are considered good interface means for such robots. Thus, we are developing a helper robot that carries out tasks ordered by the user through voice and/or gestures [6]. The 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 [7][8] for a reliable user-friendly robot system. The speech understanding module of the robot system assists the vision module through the interaction with the user. 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. This mutual assistance between vision and speech helps to make object recognition tasks more efficient with less user's burden.

However, object recognition is still a challenging task. Difficulties will diminish if the robot can restrict the search area for target objects. Environmental cues such as the existence of furniture can be used for this. For example, if the robot knows that the target object is on the table, the robot can concentrate on examining the objects on the table. It does not need to search other places. The user sometimes mentions this kind of information explicitly. However, he/she often mentions this implicitly when the environmental cues are apparent from the context of dialog, or the target objects are usually found at particular places. The robot should understand environmental cues in such implicit cases. Asking the environmental cues by speech should be the last resort.

This paper shows how the environmental cues are effectively used in object recognition and presents a robot system that can understand the environmental cues even from implicit human utterances. When it cannot obtain the cues through the presented method, the robot asks the user by speech.

2. Environmental Information

In everyday life, objects that the user wants to ask the robot to bring, such as books, fruits and facial tissues, are usually put on a small number of places for each object. If the robot learns these places, it can restrict the search area to improve efficiency. Moreover, vision processes themselves can be faster. This location information is also used to reduce the user's burden in speech. For example, if the robot knows that the ordered object is almost always at a certain place, it asks the user just to confirm the location. The user can simply say 'yes' except in irregular cases.

The objects in the environment are classified into two groups: things that may move frequently and easily such as 'book' and 'apple', and things that do not often move such as 'bookshelf' and 'table'. Gibson has classified things to be perceived into five categories: places, attached objects, detached objects, persisting substances, and events [9]. We use Gibson's terminology in this paper. The objects in the former group are detached objects, and those in the latter are attached objects.

As attached objects do not move frequently, they provide the most useful data for realizing the environment. The information of the environmental cues includes the 2D map of the room and attributes of the attached objects with their positions in the room. The room map also indicates the world directions.

For identifying the world direction and position of attached objects in the 2D room map, we use some special markers**.** We use color landmarks in the current implementation. We have attached a different color square

plate as the landmark on each upper corner of the room. When the robot gets to the object, it tries to detect multiple color plates by turning the camera. Since the positions of the plates are given in advance, it can calculate its current position easily and register the position on the map. When the attached object is later referred to by the user, it finds the color plates to know its current position, calculating the necessary movements from the current position and the object's position written on the room map.

2.1 Information of attached objects

We use the following general information of attached objects:

Name: It specifies the name of the attached object.
Position: It specifies the X, Y and Z position in the room map, where
X represents the x position of the attached object in a 2D room map of 320X240.
Y represents the y position of the attached object in a 2D room map of 320X240.
Z represents if the attached object is grounded or not in a 2D room map of 320X240.
Color: The color of the attached object is represented in HSI color space. The threshold values for H, S, I are stored.
Height Width Ratio: It is the value of Height/ Width.

However, the entire general information is not needed for all attached objects. For example, to identify 'Book Shelf" all attributes are necessary, but for 'Table', the height to width ratio is not important.

2.2 Information of detached objects

Every detached object must be associated with some attached object. The following information is required for correspondence between the detached and attached objects:

Name: Specifies the name of detached objects.
Corresponding attached object: Specifies the name of the corresponding attached object.

Every detached object in the environment is defined using the following attributes:

Color: The color of the detached object is represented in HSI color space.
Size: Size is considered relative to other objects in the scene.
Position: Position is also considered relative to other objects in the scene.
Shape: We limit the vocabulary of shape to simple ones such as 'circle' and 'rectangle' for easy interaction with the user. We assign 'others' to other complex or irregular shapes.

These attributes are used to detect the target detached object from its corresponding attached object and to make dialogue.

3. Environmental cues for object recognition

The system knows the user's order from the results of the basic speech analysis. However, some important information may be lacking to actually activate the action. Since the task of our robot is to bring the objects asked by the user, the most necessary information to carry out the task is the location information of the target and other related objects.

When the system has recognized words indicating objects, it checks whether or not the positional information related with the words necessary to carry out the action has already been obtained. If not, it starts the process to obtain the information. Once their positions are obtained, the system can use the position data as default values in future.

The robot comes up to know the position of the detached objects and associated attached objects from the following five cases:

Case 1: From explicit orders; Example:
User: Get the book on the table.

As the order is explicit, the robot gets the positional information of the target object and its corresponding attached object by analyzing the user's order.

Case 2: From implicit order; Example:
User: Get that.

In this case, the object indicated by 'that' was not mentioned before. Since the object is outstanding in the scene, the robot tries to get cues about the position of the attached and detached objects from the user. As the order is implicit, the robot must get cues from the user by observing his/her gaze and/or gesture and/ or from the manipulation or touched by the user or the robot [10]. Using these cues, the robot will confirm the positions of the attached and detached objects.

Case 3: From context; Example:
User: Go to the table.
Look right side.
There will be a green book.
Bring that book.

The robot analyses the conversation of the user and find out the position of the detached object and corresponding attached object in the environment.

Case 4: From previous experiences; Example:
User: Get the book.

In our conversation, we tend to omit the parts that the partner can understand even if we do not mention them. For example, if a particular detached object is almost always placed on a certain attached object, the user may omit the attached object in his/her instruction for the detached object. The user expects the robot to know this fact after asking it for the object several times, because the robot should build a knowledgebase from its past

experiences. On the other hand, we cannot exclude the possibility that the user may omit the attached object just by forgetting to mention it. In this example case, the robot uses the knowledge of the detached object book. From this knowledge it will come up to know that the usual place of books is the bookshelf. However, if the robot cannot obtain any information about the object in the knowledgebase, then it starts dialogue to improve its knowledgebase and vision system as in case 5.

Case 5: From the interaction with the user; Example:
User: Give me a mango.
Robot: Where is the mango?
User: It is on the table.

If any information of the attached object is not mentioned in the order and the robot is not familiar with that detached object, the robot must search all area. However, to reduce the search area, the robot must come up to know the associated attached object. Thus, the robot asks the user to give a related attached object.

For the name of the attached object associated with any detached object, the robot needs to know the answer to the question of 'where' in any case. From cases 1 to 4, the robot comes up to know the answer from the order, user's gaze and/or gesture, context and knowledgebase, respectively. However, for asking question to the user using 'where', we need to consider the following sentence pattern:

Where + be verb + Object;
Example: Where is the mango?

We use the following sentence pattern to know the answer: Subject + Be Verb + Adverbial of place. Adverbials of place are usually prepositional phrases like:

Preposition + Determiner + Adjective + Noun;
Example: The book is on the table.

Thus, from user reply, the robot will get the name of the attached object.

From the above five cases, the robot gets cues about the position of the detached object and its associated attached object. However, in real complex scenes, the vision system may detect various detached objects on the related attached object for a certain detached object. The robot must choose the target detached object among them, which is a hard problem. We have tackled this problem in [11]. The robot determines the target through a conversation with the user. If the vision fails or the robot does not find the target object, the robot asks the user for help through speech.

4. Color Segmentation

We use a robust approach of feature space method: the mean shift algorithm [12] combined with HSI (Hue, Saturation, and Intensity) color space for color image segmentation. Although the mean shift algorithm and the HSI color space can be separately used for color image segmentation, they surely fail to segment images when the illumination condition will change. We solve this problem in [13]. 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 have been used for merging regions. Finally, the Median filter has been used for smoothing the image and the region growing algorithm has been used to eliminate small regions as image post processing.

5. Experiments

We performed experiments in various cases to confirm the usefulness of our method. We show two examples here. The experimental images are shown in Figure 1.

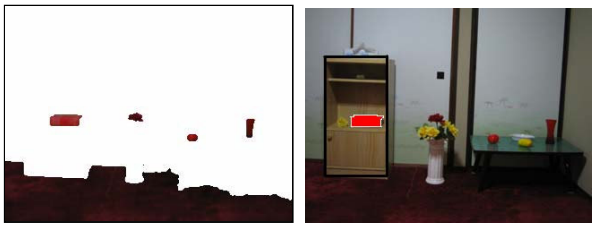
(a) Original image. (b) Segmented image.

Figure 1. Original and segmented images.

Example 1:

User: Give me the red Book.

If the robot tries to find out red objects without using environmental knowledge, it will find 5 red objects using color segmentation (Figure 2(a)). In this case, it is difficult for the robot to identify which is the book.



(a) Five red objects, (b) The red book in the bookshelf.

Figure 2. Target object among the red objects in the scene.

However, the robot has the knowledge of the detached object Book. From this knowledge it will come up to know that the usual place of a Book is Bookshelf. Then from the knowledge of the attached object Bookshelf, the robot moves to the Bookshelf and recognizes it using its attributes. After that, the robot finds out the red object in the Bookshelf (Figure 2(b)).

Example 2:

User: Bring me a Mango.
Robot: Where is the Mango?
User: It is on the table.

In this example, the robot has no knowledge of the attached object for the detached object Mango. If the robot tries to find out the Mango just as a yellow object without environmental knowledge, it may find too many candidate objects. Thus, the robot first asks the user about the associated attached object. Then, from the knowledge of the attached object Table, the robot moves to the table and recognizes it using its attributes. After that, the robot finds out the Mango on the Table based on the attribute of Mango (Figure 3).

Figure 3. Mango on the Table.

6. Conclusion

We have proposed a method that can understand user's order and accomplish the user's demand with less vision task using environmental cues. This paper shows how the environmental cues are effectively used in object recognition and presents a robot system that can understand the environmental cues even from implicit human utterances. When it cannot obtain the cues through the presented method, the robot asks the user by speech. Experiments using the robot system show the usefulness of the proposed approach.

Acknowledgmen

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," ROMAN 2002, pp. 460-467, 2002.
[2] M. Hans, B. Graf, R.D. Schraft, "Robotics home assistant Care-O-bot: Past-present-future," ROMAN 2002, pp. 380-385, 2002.
[3] G. A. Berry, V. Pavlovic, and T. S. Huang, "BattleView: A multimodal HCI research application," Workshop on Perceptual User Interfaces, pp. 67-70, 1998.
[4] I. Poddar, Y. Sethi, E. Ozyildiz, and R. Sharma, "Toward natural gesture/speech HCI: A case study of weather narration," Workshop on Perceptual User Interfaces, pp. 1-6, 1998.
[5] R. Raisamo. "A multimodal user interface for public information kiosks," Workshop on Perceptual User Interfaces, pp. 7-12, 1998.
[6] T. Takahashi, S. Nakanishi, Y. Kuno, and Y. Shirai, "Human-robot interface by verbal and nonverbal communication," IROS 1998, pp. 924-929, 1998.
[7] M. Yoshizaki, Y. Kuno, and A. Nakamura, "Mutual assistance between speech and vision for human-robot interface," IROS 2002, pp. 1308-1313, 2002.
[8] M. Yoshizaki, A. Nakamura, and Y. Kuno, "Vision-speech system adapting to the user and environment for service robots," IROS 2003, pp. 1290-1295, 2003.
[9] Gibson, J.J., "The Ecological Approach to Visual Perception," Houghton Mifflin, Boston, 1979.
[10] Z. M. Hanafiah, C. YamaZaki, A. Nakamura and Y. Kuno, "Human-Robot Speech Interface Understanding Inexplicit Utterances Using Vision," CHI 2004, pp. 1321-1324, 2004.
[11] R .Kurnia, M. A. Hossain, A. Nakamura, and Y. Kuno, "Object Recognition through Human-Robot Interaction by Speech," ROMAN 2004, pp. 619-624, 2004.
[12] D. Comaniciu and P. Meer, " Mean shift : A robust approach toward feature space analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, pp. 603 – 619, 2002.
[13] M. A. Hossain, R. Kurnia, A. Nakamura, and Y. Kuno, "Color Objects Segmentation for Helper Robot," **3**rd International Conference on Electrical & Computer Engineering (ICECE) 2004, pp. 206-209, 2004.