Human-Computer Interaction -- INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 801-804

Vision-Speech System Becoming Efficient and Friendly through Experience

Yoshiori Kuno, Mitsutoshi Yoshizaki & Akio Nakamura

Saitama University, 255 Shimo-okubo, Sakura-ku, Saitama 338-8570, Japan

kuno@cv.ics.saitama-u.ac.jp

Abstract: This paper presents a vision-speech system for service robots that can learn the user's customs and objects fixed in the environment while helping the user, and can perform their tasks more efficiently with less user's burden. We are working on a service robot that brings objects ordered by the user through speech. The robot needs vision to recognize the objects. It asks the user for help by speech if its vision fails. In early stages, it asks the user for help many times and vision takes time to detect the objects. In later stages, however, the user does not need to say details because the robot knows where the objects usually are through experience. Moreover, the vision processing time is greatly reduced because it knows what operations can work in such cases. Experiments using a robot system show the usefulness of the proposed system.

Keywords: service robot, computer vision, speech interface, object recognition

1 Introduction

As the number of senior citizens increases, more research efforts have been made to develop service robots used in the welfare service (Ehrenmann et al, 2002; Hans et al, 2002). Speech interfaces are considered proper interface means for such robots. Thus we are developing a service robot that brings things ordered by the user through voice and gestures (Takahashi et al, 1998). In addition to gesture recognition, such robots need to have a vision system that can recognize objects mentioned in speech. We have proposed the mutual assistance between vision and speech to realize a reliable user-friendly robot system (Yoshizaki et al, 2002). The vision module helps the speech understanding module by finding the objects mentioned in speech. The latter assists the former through the interaction with the user 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 the robot's vision system.

This paper presents our extended work under this vision-speech cooperation framework. It is acceptable during some initial period from the arrival of the robot to the user's home even if the robot takes much time to find an ordered object by vision and the user needs to make several instructions by speech. However, the user may not use the robot if this situation continues. Humans expect that the same or similar orders can be carried out faster and easier at later times. We propose a vision-speech system to meet this expectation.

In everyday life, objects that the user wants to ask the robot to bring, such as books, fruits and tissue papers, 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. We cannot expect that a single fixed vision algorithm can work all the time at all locations. Thus, our vision system has various vision routines with several parameters and tries every possible combinations of the routines with every possible parameters to detect ordered objects when no additional information is available. If it knows that the current target object is often found at a certain place and remember the combinations and parameters effective there, it tries them first to improve efficiency. 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.

This paper describes how we realize the system mentioned above and shows experimental results.

2 Overview

In this paper, we mainly describe how the system reduces the user's speech burden through experience. Before that, this section gives the overview of the total system.

We use ViaVoice by IBM for speech recognition. We divide the speech recognition output into morphemes and parse them using the software developed at Nara Institute of Advanced Science and Technology (Matsumoto et al, 2001).

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 including names of people, 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.

The objects registered in the dictionary are classified into two groups: things that move frequently and easily and things that do not often move. We call the former detached objects, and the latter attached objects by adopting Gibson's terminology (Gibson, 1979). He has classified things to be perceived by vision into five categories: places, attached objects, detached objects, persisting substances, and events. Attached objects are fixed to places and become the features of them.

Attached objects include such as 'bookshelf' and 'desk.' These objects do not often move. Thus once their positions are obtained, the system can use the position data as default values in future. However, these objects are generally large and cannot be represented by simple image features. Vision is not good at recognizing these objects. Thus, for these objects, we use the assistance by speech from the beginning if their positions have not been obtained (Yoshizaki et al, 2002).

The robot will reach attached objects guided by the user's spoken commands at the first time. The system needs to remember their positions so that it can move there without the user's assistance later on. We do this by using 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, registering the position on the map.

When the attached object is later mentioned 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.

Detached objects include such as 'book', 'apple' and 'names of people.' Since objects represented by these words can move easily, the system needs to initiate the vision processes to find them when the user refers to them. Detached objects are defined by attributes such as color and shape. The system finds these objects from the results of image processing modules to detect the related attributes. It tries all possible combinations of image processing modules with all possible parameter changes. Even though these trials, if the target object cannot be detected by vision, the system calls for the assistance through speech. When the target object has been found, the system puts the object name, the successful combinations of image processing modules with the parameters, and the location into the memory as history data.

The history data is used to select effective image processing modules quickly at later times. If no history information about the target object is stored in the system, all the combinations are tried as mentioned above. If any, however, the efficiency can be greatly improved. The system searches the history data for similar experiences in terms of the object name and the attached object and tries the combinations in these cases first. For example, if 'an apple on the table' is requested and there have been several such cases before, the successful combinations in these cases are first applied.

The history data is also used to reduce the user's burden in speech. The system calculates the probability where the object exists from the past history, and chooses the action that is most comfortable for the user. Details of this part is described in the next section.

3 Friendly Interaction with the User

Attached objects can restrict the search area for a target detached object. If any attached object is not mentioned, the robot must search all area. Thus, during some period from the beginning of the use,

the robot asks the user to give a related attached object if he/she does not mention it. For example, if the user says, “Get me an apple,” the robot asks, “Where is it?” expecting an answer such as “That’s on the table.” We believe that such interaction is acceptable at earlier times but desirable to be avoided after some period. In our conversation, we tend to omit the parts that the partner can understand even if we do not mention them. 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. On the other hand, we cannot exclude the possibility that the user may omit the attached object just by forgetting to mention it. If a certain detached object usually exists on one of a few particular attached objects, it might be more comfortable for the user if the robot mentions these attached objects and asks which one, rather than asking where it is. We propose a dialog system that can deal with these situations by considering the history and the cost of action and communication.

For the present target detached object, the system examines the history data, calculating the association probability that it is found on (at, in) each attached object. The probability is defined by,

(the number of cases associated with the attached object) / (the total number of cases for the detached object).

In the current implementation, we consider only two attached objects A1, A2 that give the highest probabilities: p, the highest, and q, the second.

We consider the following costs for the user.

A: Cost for correcting the action that the system mistakenly performed.

B: Cost for correcting the dialog that the system mistakenly performed.

C: Cost of answering to a description type question such as “Where is it?”

D: Cost of answering to a selection type question such as “Is it on X or Y?”

E: Cost of answering to a yes-no type question such as “It is on Z, isn’t it?”

When the robot needs to find a particular detached object, it calculates the expected cost for each possible action. Possible actions considered here are as follows.

Action 1. The robot assumes that the object exists on the attached object with the highest probability, moving there automatically. In this case, the user does not need to do any additional thing if the assumption is correct. Otherwise, however, he/she must pay Cost A. Thus, the expected cost in this case is, $(A+C)(1-p)$. (1)

Action 2. The robot also assumes the same as in Action 1. However, it asks the user for confirmation by saying, " <the object> exists on (in, at) <the attached object>, isn’t it?" The expected cost is,

$$Ep+B(1-p). \quad (2)$$

Action 3. The robot assumes that the object exists either on the highest-probability attached-object or the second. Then, it asks the user, " Is <the object> on (in, at) <attached object A1> or on (in, at) <attached object A2>?" The expected cost is,

$$D(p+q)+B(1-p-q) \quad (3)$$

Action 4. The robot does not use the history data and asks the user, "Where is it?" In this case, the cost is,

$$C \quad (4)$$

The robot selects the action with the lowest expected cost. It is difficult to determine the absolute values for Costs A-E, and they may vary depending on the user. However, we can assume the descending order, A>B>C>D>E. In the experiments described in the next section, we use the following values: A=8, B=6, C=3, D=2, E=1.

4 Experiments

4.1 Dialog Selection

We observed a person’s life and performed simulation experiments based on the observation results to examine the effectiveness of the dialog simplification. We asked a subject to keep records of the positions of the TV remote controller, a tissue-paper box, and the alarm clock when he wanted to get them. The observation results are, for example, a tissue paper box was under the table, on the TV, and anywhere else, 73%, 17%, and 10%, respectively.

Then, we performed simulation experiments. We assume that the user asks the robot to fetch these three detached objects many times. The simulation program makes these objects happen to exist at places according to the probabilities obtained in the observation. Figure 1 shows the running-average cost. This clearly demonstrates that our method reduces the user’s cost.

4.2 Robot Experiments

We performed real-world experiments although in small scale. We use Pioneer 2 by ActivMEDIA as

a robot. The current system does not have a robot arm. Thus, we consider it success if the robot reaches the attached object and finds the detached object ordered by the user.

We asked the robot to bring a pair of gloves. The gloves were found on the table three times, on the carton box twice, and on the desk once before this experiment. Then, we asked the same order again.

The robot calculated the action costs: 5.5, 3.5, 2.7, 3.0 from Actions 1 to 4, respectively in this case. Thus, it took Action 3, asking, "Are the gloves on the table or the carton box?" The user answered, “On the table.” Then, the robot moved to the table and performed object recognition. Figure 2 shows the input image. Figure 3 shows the detection result. When the robot first tried to recognize gloves on the table, it examined 1,152 combinations, detecting candidates 104 times, among which six were glove cases. In this seventh trial, it executed 16 sequences of operations, detecting a candidate 14 times. The gloves were detected 13 times. The processing time for the first trial was about 3 minutes whereas about 5 seconds for the seventh. The experimental result shows the efficiency of the proposed method using the history data.

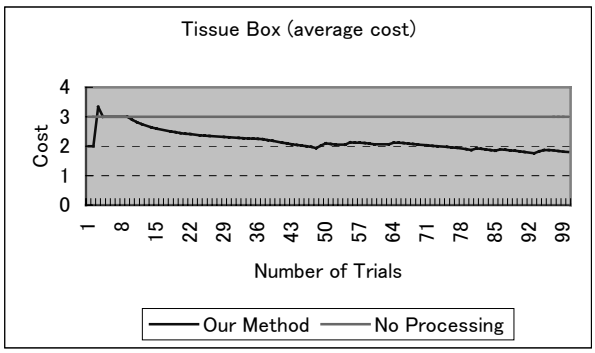

Figure 1: Average cost.

5 Conclusion

We have proposed a vision-speech system for service robots that can learn the user’s customs and objects fixed in the environment while helping the user, and can perform their tasks more efficiently with less user’s burden. Experiments using a robot system show the usefulness of the proposed system.

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 14019012, 14350127).

Figure 2: Original image.

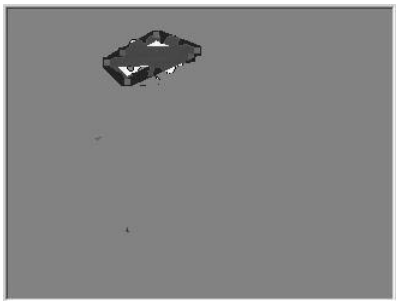

Figure 3: Detection result.

References

Ehrenmann, M., Zollner, R., Rogalla, O. & Dillmann, R.(2002), Programming Service Tasks in Household Environments by Human Demonstration, *Proc. ROMAN 2002*, pp.460-467.

Gibson, J.J.(1979), *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston. .

Hans, M., Graf, B. & Schraft, R.D.(2002), Robotics Home Assistant Care-O-bot: Past - Present – Future, *Proc. ROMAN '02*, pp. 380-385.

Matsumoto, M., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K. & Asahara, M.(2001), *Japanese Morphological Analysis System ChaSen version 2.2.4 Manual*, Nara Institute of Science and Technology (in Japanese).

Takahashi, T., Nakanishi, S., Kuno, Y. & Shirai, Y.(1998), Human-Robot Interface by Verbal and Nonverbal Communication, *Proc. IROS 1998*, pp.924-929.

Yoshizaki, M., Kuno, Y. & Nakamura, A.(2002), Mutual Assistance between Speech and Vision for Human-Robot Interface, *Proc. IROS 2002*, pp.1308-1313.