Human-Robot Speech Interface Understanding Inexplicit Utterances Using Vision

Zaliyana Mohd Hanafiah Chizu Yamazaki Akio Nakamura Yoshinori Kuno
Department of Information and Computer Sciences, Saitama University
255 Shimo-Okubo, Sakura-ku, Saitama-shi, Saitama 338-8570 JAPAN
{zaliyana, yamazaki, nakamura, kuno}@cv.ics.saitama-u.ac.jp

Abstract
Speech interfaces should have a capability of dealing with inexplicit utterances including such as ellipsis and deixis since they are common phenomena in our daily conversation. Their resolution using context and a priori knowledge has been investigated in the fields of natural language and speech understanding. However, there are utterances that cannot be understood by such symbol processing alone. In this paper, we consider inexplicit utterances caused from the fact that humans have vision. If we are certain that the listeners share some visual information, we often omit or mention ambiguously things about it in our utterances. We propose a method of understanding speech with such ambiguities using computer vision. It tracks the human's gaze direction, detecting objects in the direction. It also recognizes the human's actions. Based on these bits of visual information, it understands the human's inexplicit utterances. Experimental results show that the method helps to realize human-friendly speech interfaces.

Categories & Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces - *input devices and strategies, natural language*

General Terms: Design; Experimentation; Human Factors

Keywords: Speech understanding; multimodal interface; robot; gaze; ellipsis; deixis; natural language processing; computer vision

INTRODUCTION
Speech is a promising human-interface means for helper robots, which have growing needs in coming aging society. Thus, robots with a speech interface have been investigated [3,9]. Speech interfaces should have a capability of dealing with inexplicit utterances including ellipsis and deixis since they are common phenomena in our daily conversation. Their resolution using context and a priori knowledge has been investigated in the fields of natural language and speech understanding [8,10]. However, in the case of robots, we must consider inexplicit utterances that cannot be understood by such symbol processing alone. We humans have vision. Thus, in our speech, we may omit or mention ambiguously things which we think that the listeners know by vision. For example, we may say, "Get that to me," even though the object indicated by *that* was not mentioned before. Since the object is outstanding in the scene and the listener seems to look in the direction from his/her gaze, we assume that he/she is aware of the object. The robot should be able to act like this to be user-friendly. In this paper, we propose a method of understanding speech with such ambiguities using computer vision.

CHI 2004, April 24–29, 2004, Vienna, Austria.
ACM 1-58113-703-6/04/0004.

Grice proposed the conversational maxims [4]. One of them is that conversation is the cooperating work between a speaker and a listener where both will offer necessary and sufficient related information briefly and clearly. Based on this, we assume that we can get the information by vision about the things that are important but are not mentioned clearly in speech. Actually, there are various inexplicit utterance cases other than this vision-derived one. We assume that these are solved by other researches and only vision-derived ones are left in our speech input.

We are developing a helper robot that brings the object that the user asks through speech [11]. In this paper, we present a speech interface for the robot that allows the user to use utterances with such vision-derived inexplicitness.

INEXPLICIT UTTERANCES
As mentioned in Introduction, visual information shared by a speaker and a listener may cause inexplicit utterances. Among such utterances, we deal with ellipsis and deixis that may appear in the speech interface for the helper robot [11].

Human utterances can be considered mostly the requests that he/she asks the robot in this restricted application domain. Such a request consists of a verb and an object. (The subject is always the robot.) The verb indicates an action that the human wants the robot to take. The object indicates the target of the action. For each of verb and object, the human may say it definitely or ambiguously (deixis). Or he/she does not say it at all (ellipsis). Thus, utterances are classified into nine cases.

We give an utterance example for each case below. Note that we use Japanese language in our system. In the

following examples, we give direct translations of the Japanese and show the English supposed to be used in such situations in parentheses if necessary.

Case 1. Verb omitted; Object omitted. Examples are greetings such as "Hello."

Case 2. Verb omitted; Object ambiguous. "That one."

Case 3. Verb omitted; Object definite. "That apple."

Case 4. Verb ambiguous; Object omitted. To say "Make to four." while watching the television. ("Channel four.")

Case 5. Verb ambiguous; Object ambiguous. "Do that."

Case 6. Verb ambiguous; Object definite. "Do the red one." ("Red one.")

Case 7. Verb definite; Object omitted. "Get." ("Get it.")

Case 8. Verb definite; Object ambiguous. "Get that."

Case 9. Verb definite; Object definite. "Get the red book."

INFORMATION OBTAINED BY VISION

We see the environments around us. We see what the conversation partners are doing. We use such visual information to understand inexplicit utterances. First, we show what visual information can be used to guess the object parts in inexplicit utterances. We omit an object or mention it ambiguously because we think that it is apparent to the partner. We (including the partner) commit ourselves to the object in some sense. Thus, the object must be something that is related to our action. Candidates of such objects are as follows.

1. Being near (Proximity): Objects close to the human or the robot.
2. Watching (Gaze): Objects watched by the human.
3. Pointing: Objects pointed by the human's arm and hand.
4. Manipulating: Objects touched or manipulated by the human or the robot.

If the object part is uncertain after the language analysis, the system tries to detect objects in the above categories. If multiple objects have been detected, the pointed object is given the first priority, since the pointing action is intentional. In other cases, the robot asks the human which one is the target object through speech.

The objects registered in the dictionary 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 'television'. Gibson has classified things to be perceived into five categories: places, attached objects, detached objects, persisting substances, and events [2]. We use Gibson's terminology in this paper as in our previous research [11]. The objects in the former group are detached objects, and those in the latter are attached objects.

The system guesses the verb parts based on the object parts, which are either mentioned definitely or guessed by the method described above.

The possible verbs for each object are registered in the dictionary in advance. The default for detached objects is to get. If detached objects can have other possible verbs, they are written for each object. For example, we may ask the robot to do various operations with a TV remote controller: turning on/off of the power, controlling volume, and changing the channels. For attached objects, the default is to go (to the object). When there are multiple possible verbs, we do not give priorities among them in the current implementation. In such cases, the robot asks the user through speech.

LANGUAGE PROCESSING

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 [7]. From the parsing result, we classify sentences into one of the nine cases.

Figure 1 shows several examples of parsing results. After parsing, we assign either D (definite) or A (ambiguous) to the object and the verb. If the verb or the object is omitted (missing), M is assigned. Example 1 in Figure 1 is analyzed as a perfect request, Case 9, since the result is {D, D}. Examples 2, 3, and 4 are determined as Case 4 {A, M}, Case 8{D, A}, and Case 3 {M, D}, respectively. After the utterance is classified, vision processes are initiated depending on the case.

VISION PROCESSING

We have developed a robot system with two stereo camera pairs as shown in Figure 2. The lower stereo pair of IEEE 1394 cameras (DFW-V500, Sony) are watching the user's face and hands, obtaining the face direction (rough gaze direction), recognizing pointing gestures, and detecting the objects touched or manipulated by the hands. The upper stereo pair of pan-tilt controllable cameras (EVI-D100, Sony) search for objects along the 3-D line of the face

Legend

$S \rightarrow Sentence$
$V \rightarrow Verb$
$N \rightarrow Noun$
$P \rightarrow Noun-Pronoun$
$Adj \rightarrow Adjective$
$Ot \rightarrow Others$ (interjection, number etc.)
$NP \rightarrow Noun\ Phrase$

- **Request Sentences**

A perfect request sentence {<Verb>,<Object + Attribute>}
Example 1 : Get me that red book {D,D}
$S1 \rightarrow V\ adj\ N$
Example 2 : Make to 4 {A,M}
$S2 \rightarrow V\ Ot$
Example 3 : Get that {D,A}
$S3 \rightarrow V\ P$
Example 4 : That apple {M,D}
$S4 \rightarrow P\ N$

Figure 1. Classification of inexplicit utterance cases.

direction or the pointing direction using the zero-disparity filter (ZDF) [1]. They also detect objects in front of the human and the robot.

We use the 3-D human motion recognition system MARIO developed at Kyushu University [6] to detect the face and the hands. Then, the system computes the 3-D direction of the face (or the arm). We consider the face direction as approximate gaze direction. The two pan-tilt controllable cameras rotate while their optical axes keep converging on the 3-D line indicating the face (or arm) direction. The system calculates the correlation between the central regions of the two camera images. If the correlation is high, the system judges that an object exists there. Figure 3 shows an experimental scene. The robot detected the electric pot that the human was looking at using the ZDF. Figure 4 shows a stereo pair of images (left and center) and the ZDF result (right) where the detected object is indicated by the rectangle.

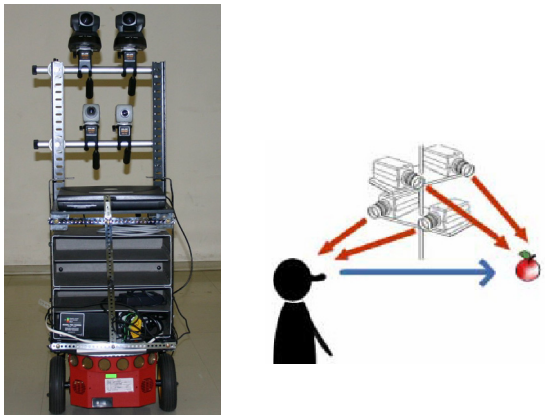


Figure 2. Robot system with two stereo camera pairs.

Figure 3. Experimental scene.

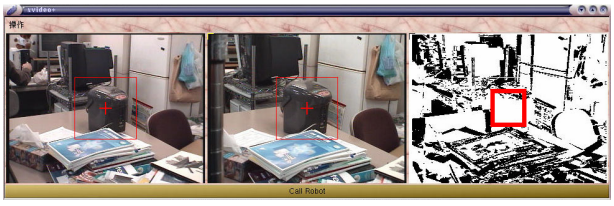


Figure 4. Object detection result.

SYNCHRONIZATION BETWEEN SPEECH AND VISION

We need to consider the synchronization between speech and vision. The vision processes to track the face and the hands are working all the time. If a hand is raised during an utterance, this hand motion is considered as a pointing gesture. Actions of being near and manipulating are not so fast that the robot can start the vision processes after the analysis of utterances. However, gaze moves fast. It may change during an utterance. The object detection based on the ZDF cannot work so fast as the eyes move. And even if it can, the robot needs to determine what is the target object if multiple objects are detected in multiple gaze directions. Thus, we performed an experiment to examine the synchronization between speech and gaze direction.

We put five objects in the scene. A subject sat close to the robot as shown in Figure 3, asking it, "Ano (that) *object-name* totte (get)." (Get that *object-name*. We use the Japanese language in our system.) The robot computed the gaze (face) direction during the period from a little before the utterance to a little after. We used three subjects; all were graduate students in our department. We asked them to change the part of *object-name* randomly and said the sentence. Each subject ordered the robot 20 times.

Figure 5 shows the result. We consider that the gaze directions holding more than five frames (0.17s) are meaningful and others are transient. The figure shows the frequency (the percentage out of 60 trials) of gaze directions during the period of each word utterance (approximately 10 frames) and during 10-frame (0.33s) periods before and after the utterance. In the figure, the directions are specified by what exists there such as "Robot" and "Object" (the target object). "Others" indicates any other objects than the target object. "Moving" means that the gaze direction was changing during the period and no meaningful gaze direction was observed. The result indicates that the subjects begin to see the target object before the utterance and see it when they utter its name. The fact of seeing the target object before utterance agrees on the findings by Kaur et al. on their gaze-speech input system [5].

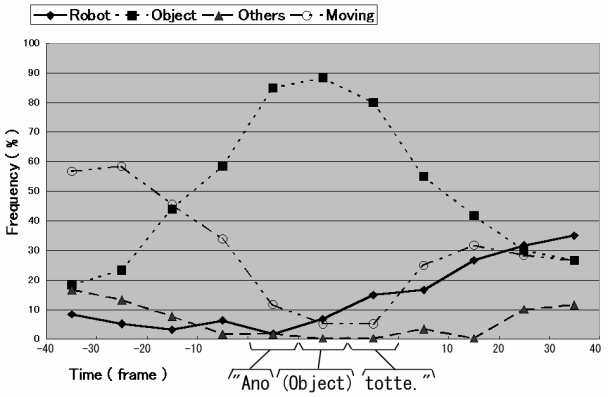


Figure 5. Gaze direction changes when giving an order.

We used complete utterances (Case 9) in the above experiment because we could not design experiments in which subjects used inexplicit utterances in a natural way. If we ask subjects to use inexplicit utterances, they tend to fix their gaze on the target object. Experiments in inexplicit utterance cases are left for future work. Still, the above experimental result suggests how the robot can determine the gaze direction where it searches for a target object. Important findings are that the main gaze directions are toward the object or the robot, and that the gaze direction tends to begin moving toward the object before the utterance. Thus, we have set up the system as follows. After the utterance, it starts searching for an object in the gaze direction observed most frequently during the period from a little before the utterance through its end except that toward the robot. If multiple stable gaze directions are observed during the period, the one around the starting time of the utterance is first examined.

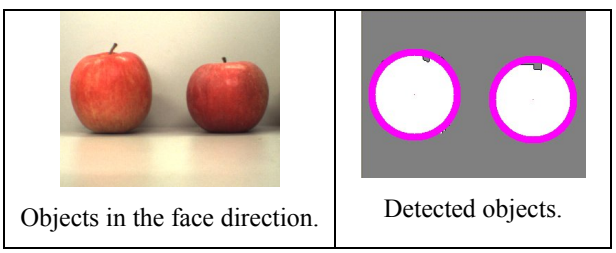
EXPERIMENTS

We performed experiments in various cases to confirm the usefulness of our approach. Here, we show an example of the dialog between the user and the robot. In this case, both of them were looking in the same direction and there were two red apples in the directions. The dialog and the robot actions in this experiment were as follows.

User: "Get that."
The utterance was classified as Case 8. The robot recognized the user's face direction, and detected two objects in the direction (Figure 6). The robot told the current understanding status (image processing result) to the user.
Robot: "I have found two red round objects. Which should I get?"
User: "Get the left apple."
Robot: "This one?"
The robot shows the user the display where the target object is placed in the center, while saying the above sentence.
User: "Yes."

Although the current system is a small experimental system with the vocabulary size of 300 words, the experimental results show that it can understand such inexplicit utterances as listed as examples in the second section.

The robot may find multiple objects associated with the human's various actions. In the current implementation, we do not give orders among them except in the case of pointing action. The robot asks the human through speech in such cases. More detailed language analysis and/or the use of other visual information may solve this issue. This is left for future work in addition to improving each part of the system.



Objects in the face direction. Detected objects.

Figure 6. Experimental example.

CONCLUSION

In our conversation, we often omit or mention ambiguously things which we think that the listener knows by vision. It is desirable to allow such simplification in speech to realize user-friendly human interfaces. In this paper, we have presented a speech interface for helper robots that allows the user to use such simplified inexplicit utterances using computer vision. Experimental results show our approach promising.

ACKNOWLEDGMENTS

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, 15017211).

REFERENCES

[1] Coombs, D. and Brown, C., Real-time binocular smooth pursuit. *IJCV 11*, 2 (1993), 147-164.

[2] Gibson, J.J., *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

[3] Graf, B. and Hagele, M., Dependable interaction with an intelligent home care robot. *Proc. ICRA 2001*, 21-26.

[4] Grice, H.P., *Logic and Conversation Syntax*. Harvard University Press, 120-150, 1975.

[5] Kaur, M., Tremaine, M., Huang, N., Wilder, J., and Zoran., "Where is "it"? Event synchronization in gaze-speech input systems. *Proc. ICMI 2003*, 151-158.

[6] MAlib development team. http://www.malib.net/.

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

[8] Schiehlen, M., Ellipsis resolution with underspecified scope. *Proc. ACL 2000*, 72-79.

[9] Seabra Lopes, L. and Teixeira, A., Human-robot interaction through spoken language dialog. *Proc. IROS 2000*, 528-534.

[10] Watanabe, M., Masui, F., Kawai, A., and Shino, T., Conversational ellipsis and its complement. *Trans. IEICE 2000 SP2000-99*, 31-36 (in Japanese).

[11] Yoshizaki, M., Nakamura, A., and Kuno, Y., Vision-speech system adapting to the user and environment for service robots. *Proc. IROS 2003*, 1290-1295.