Three-Layer Model for Generation and Recognition of Attention-Drawing Behavior

Osamu Sugiyama[1,2], Takayuki Kanda[1], Michita Imai[1,2], Hiroshi Ishiguro[1,3], Norihiro Hagita[1]

[1]ATR Intelligent Robotics and Communication Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto, Japan

[2]Graduate School, Keio University
Kouhokuku, Yokohama City, Kanagawa, Japan

[3]Graduate School of Engineering, Osaka University
Suita City, Osaka, Japan

sugiyama@atr.jp, kanda@atr.jp, michita@ayu.ics.keio.ac.jp, ishiguro@ams.eng.osaka-u.ac.jp, hagita@atr.jp

Abstract - This paper presents a three-layer model for generation and recognition of attention-drawing behavior. The model enables a robot to recognize people's attention-drawing behavior as well as to perform attention-drawing behavior to people. It consists of three layers: the PSM (Pointing Space Model), the RTM (Reference Term Model), and the OPM (Object Property Model). The PSM associates the pointing gesture with a reference term, the RTM associates positional relationships with a reference term, and the OPM associates other supplemental verbal cues with a reference term. We implemented the model in a humanoid robot, Robovie, and verified its effectiveness through an experiment.

Index Terms - Human Robot Interface; Human Robot Interaction; Deictic Gestures, Attention Drawing

I. INTRODUCTION

The aim of our research is to develop "communication robots" that naturally interact with humans and support daily human activities based on advanced interaction capabilities with their human-like bodies. Since the target audience of a communication robot is ordinary people who do not have specialized computing and engineering knowledge, a conversational interface using both verbal and non-verbal expressions is becoming more important. Previous studies in robotics have emphasized the merits of robots' embodiment. For example, they have shown the effectiveness of facial expressions [1], eye-gaze [2], and gestures [3].

We particularly focus on mutual attention-drawing behavior as shown in **Fig. 1**, which is based on pointing gestures with reference terms. When we talk about objects in an environment, we indicate to a listener which object is currently under consideration by using pointing gestures and such reference terms as "this" and "that." Humans use pointing gestures in communication even when they are infants, which is a phenomenon widely known in developmental psychology as joint attention [4]. Reference terms also play an important role in human interaction by quickly and naturally informing the listener of an indicated object's location. In casual conversation among adults, people often use reference terms in combination with pointing, such as saying "look at *this*" while pointing at an object in order to draw others attention to it.

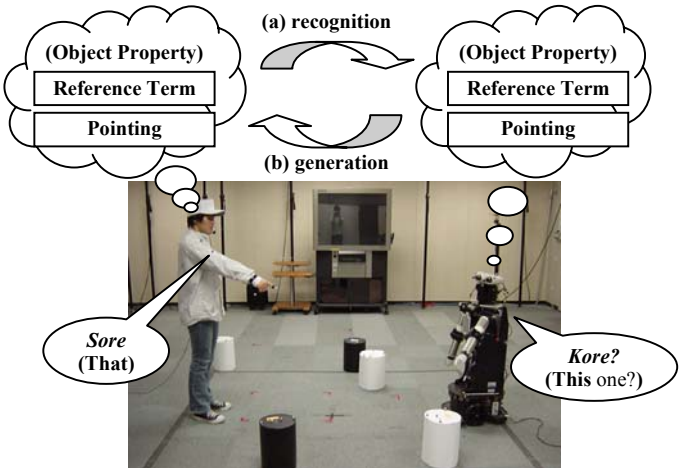


Fig. 1: Attention-drawing behavior

We believe that it is important for a communication robot to be capable of both understanding people's attention-drawing behavior and performing attention-drawing behavior for people, so that a robot can talk with a person in such a way that a person says "take *that*," with the robot answering "*this* one?" This mutual attention-drawing capability will be the essential part of a natural conversation that refers to an object.

We have established a three-layer model for generation and recognition of attention-drawing behavior by observing inter-human communication (**Fig. 2**). The first layer, the PSM, associates the pointing gesture with the reference term. The second layer, the RTM, associates positional relationships with reference terms, and the third layer, the OPM, associates other additional verbal cues used supplementary with the reference term. The most important requirement for the model is that usage of reference terms depends on physical relationships among the robot, person, and objects; thus, the appropriate reference terms for a particular object by the robot and by the person are sometimes equal but sometimes different. This is typically complicated in Japanese, because there are three types of reference terms (details are explained in Section III, **Fig. 4**).

The generation part of the model was reported in our previous work [5], while this paper addresses the recognition part. We implement the model to a humanoid robot, Robovie, for both generation and recognition of attention-drawing

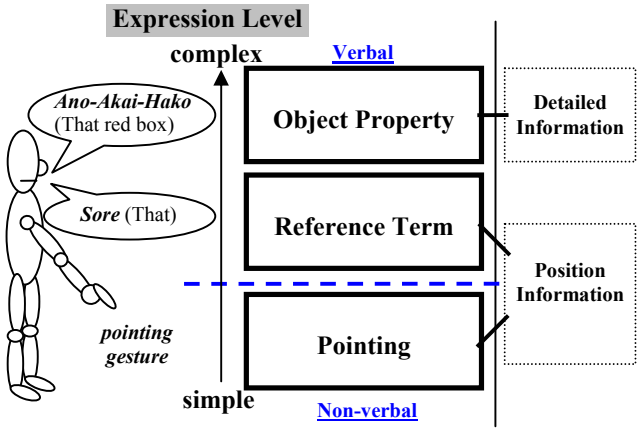


Fig. 2: The Three-layer Model of Attention-Drawing Behavior

behavior. Its effectiveness will be demonstrated through an experiment.

II. Related Work

There are many related research studies on attention-drawing interaction in robotics. One type of research focuses on joint-attention in relation to children's development [6] [7]; also, there are studies on mutual non-verbal behavior, such as research toward human-robot teamwork [8]. However, these investigations do not address a method of using reference terms appropriately in various situations.

Several previous studies focused on the recognition method for a robot of a human's pointing gesture or utterance with reference terms. For instance, Mizuno et al. proposed a method that informs robots of an object's location with both hand gestures and verbal cues [9], while Haasch et al. proposed a method to recognize the object by using a pointing gesture, a human utterance, and stored object information [10]. Hanafiah at el. made a robot recognize the object using inexplicit utterances including reference terms and pointing gestures [11]. Other studies have focused on a recognition method that uses utterance contents such as reference terms and information on the object in question. Inamura et al. proposed a probabilistic method of recognizing the object indicated by the human using object information such as color or size in utterances with reference terms [12]. In these studies, recognition is, however, investigated separately from the attention-drawing mechanism of a robot.

On the other hand, many robots are equipped with an attention-drawing mechanism [13] [14], but they cannot dynamically handle environments where the locations of objects and people change; that is, they only performed pre-implemented gestures and utterances.

To tackle these problems we have developed a three-layer attention-drawing model for a robot to use pointing gestures and reference terms [5], and develop it further to one for both generation and recognition of attention-drawing behavior.

III. A Three-layer Model For Generation and Recognition of Attention-Drawing Behavior

A. *Modeling of attention-drawing behavior*

To refer to an the object in the environment, we point at the object and use a word like "this" and "that," and limit the space where the object can be. We define this behavior as "attention-drawing behavior." In addition, we add adjectives describing properties of the object, such as color or shape to narrow down the candidates to one specific object in case we cannot identify the object only by using a pointing gesture and verbal cues. In order to refer to an object in the environment, we have developed a three-layer model for generation and recognition of attention-drawing behavior (**Fig. 2**), having the same concept as the model for generating attention-drawing behavior proposed in [5]. The three-layer model consists of three sub-models, the PSM (Pointing Space Model), the RTM (Reference Term Model) and the OPM (Object Property Model). The PSM limits the objects' possible location with a pointing gesture. The RTM limits the possible location of the object with a reference term, and the OPM narrows down the number of possible objects to one certain object by using the properties of objects. A pointing gesture limits where an object can be based in the pointing direction. On the other hand, reference terms limit where an object can be based with respect to the positions of the speaker, the listener, and other objects. Using these two different types of spatial limitation, we can identify the object space more clearly. In addition, by including an adjective describing a property of the object, we can effectively communicate the object's position to the listener.

B. *PSM: Pointing Space Model*

The PSM limits where in the environment an object can be located by using a pointing gesture. When somebody points out the place close to her, her indication is clear so that we can narrow down the object existing space. On the other hand, when she points out a place in the distance, pointing may not be clear so that the existing space of the object narrowed down by the pointing gesture will be still wide. In this research, we define the space in which the object exists to be limited by the scope of θ_p from the pointing direction as shown in **Fig. 3**. Thus, the space where the object can be located becomes narrow when the distance from speaker is short, and conversely wide if that distance is long.

C. *RTM: Reference Term Model*

The RTM limits where in the environment the object can be located by using Japanese reference terms. In Japanese, there are three reference terms, *kore*, *sore*, and *are*, which are comparable to "this" and "that" in English. There have been precedent studies about the usage of these reference terms, but they did not have clear parameters to divide the usage of them, such as distance between the speaker and objects [15]. In previous work [5], we conducted an experiment to observe a conversation between humans, where the speaker indicated the position of one object to the listener, to investigate the spatial factors that determine which reference term should be

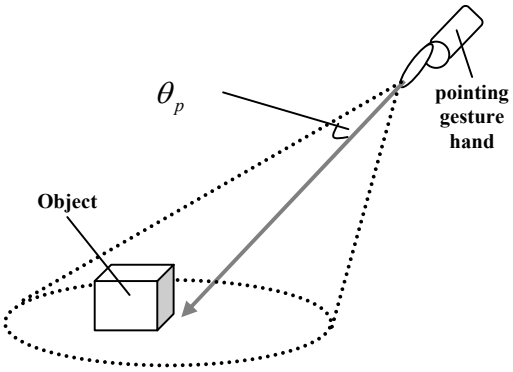


Fig. 3: PSM: Pointing Space Model

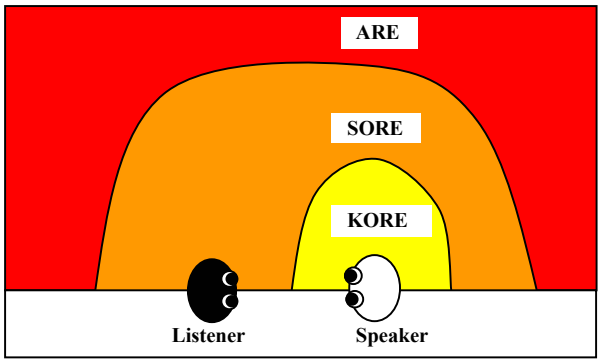


Fig. 4: RTM: Reference Term Model

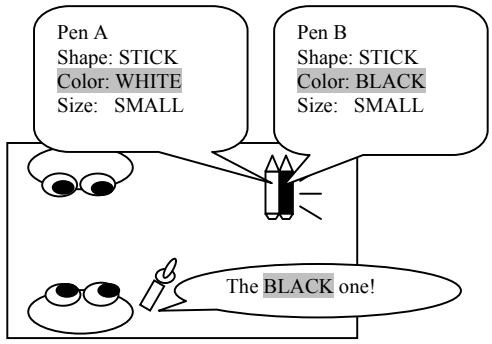


Fig. 5: OPM: Object Property Model.

used. Consequently the rough image of the reference terms' usage regions is confirmed as shown in **Fig. 4**. *Kore* refers to an object close to the speaker, *sore* refers to an object close to the listener, and between individuals, while *are* refers to an object neither close to the speaker nor to the listener. We made approximate curve functions for each border between reference terms based on the experimental results. With these functions, we determine the location of the indicated object.

D. OPM: Object Property Model

The OPM limits the number of possible objects by using properties of the object in question such as color, shape, and size when objects are too close together to be distinguished by pointing and a reference term. We often identify an object by adding adjectives describing its properties in an utterance, such as "That red box." To identify the indicated object, the property chosen should differentiate the chosen object from others. The speaker may identify several properties in one object, such as shape, color, and size, and by using a property or a property set that differentiates the object he can indicate to the listener which object is under consideration. Figure 5 shows an example of object property selection, where there are two pens in the environment. Because they are close together, the listener cannot identify one from the other simply by using a pointing gesture and a reference term. Pen A has the following properties. Shape: STICK; Color: WHITE; and Size: SMALL. For Pen B, Shape: STICK; Color: BLACK; and Size: SMALL. Here the property that divides the pens is color. If the speaker wants to indicate Pen B to the listener, he identifies it using: "That BLACK one."

IV. The Robot System for Generating and Recognizing an Attention-Drawing Behavior

Based on the three-layer model, we developed a system that recognizes a person's attention-drawing behavior and generates the robot's attention-drawing behavior to confirm the recognized object (**Fig. 6**). The hardware and software configuration are as follows.

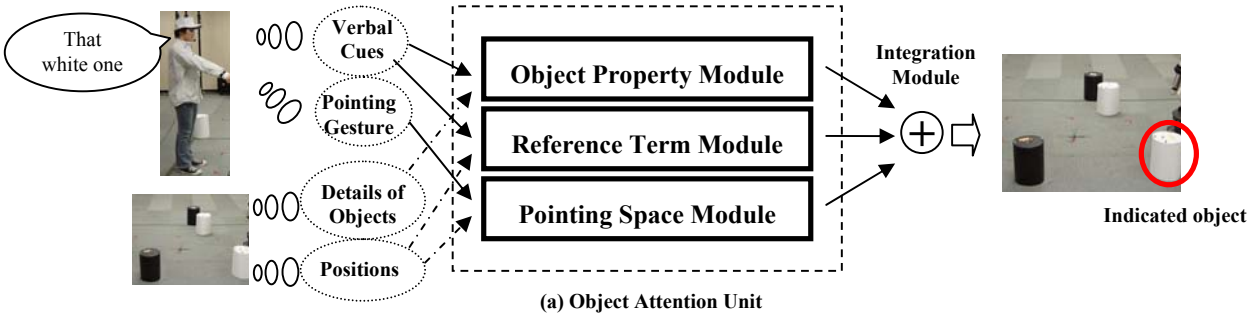
A. Hardware Configuration

A system based on our three-layer model was developed using a communication robot "Robovie" [16], a motion capturing system (VICON*), and a microphone. "Robovie" (**Fig. 8**) is 1.2 m tall with a 0.5 m radius and a human-like upper body designed for communicating with humans. It has a head (3 DOF), eyes (2*2 DOF), and arms (4*2 DOF). With a speaker in its head, it can produce output. With its 4-DOF arms, it can point with a gesture similar to that of humans. With a motion-capturing system (**Fig. 7**) we can obtain a 3D position from markers attached to the person, Robovie, and objects, and using an Ethernet, the system obtains their 3D positions as input and calculates their locations. In addition, the speaker wears a microphone as shown in **Fig. 9** to avoid the difficulty of speech recognition.

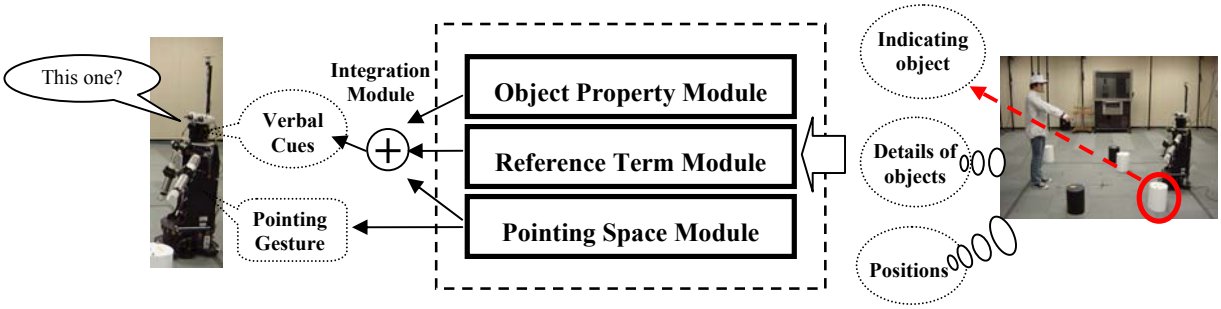
B. Software Configuration

Figure 6 shows the system configuration of our proposed system. The system consists of two main units, the Object Recognition Unit and the Object Referring Unit. The Object Recognition Unit determines which object is indicated by a person based on the three-layer model using the result of the speech recognition and positions of the person, the robot, and the objects in the environment. The Object Referring Unit, meanwhile, determines an appropriate verbal cue and generates the robot's pointing gesture. Coordinating the selected verbal cues and a pointing gesture, the Object Referring Unit enables the robot to confirm the object determined by the Object Recognition Unit based on the three-layer model. By using the same three-layer model, the system can smoothly recognize the indicated object and make the robot confirm its understanding with the human.

* http://www.vicon.com/



(a) Object Attention Unit



(b) Object Referring Unit

Fig. 6: System configuration

Fig. 7: Vicon motion-capturing system

Fig. 8: Robovie

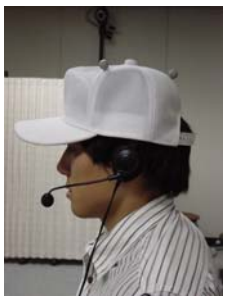


Fig. 9: The microphone used in system

C. Object Recognition Unit

The Object Recognition Unit (**Fig. 6 (a)**) determines which object the person indicated by using the result of speech recognition and subjects' 3D positions. This unit consists of the PSM, RTM, and OPM: the modules represent the functions of the three-layer model and the Integration Module. The PSM calculates the probability that the person's pointing gesture is indicating each object in the environment. The RTM calculates the probability that the reference term uttered by the person is indicating each object in the environment, and the OPM calculates the probability that the properties of the object uttered by the person indeed represent the objects. With these three probabilities calculated by the three-layer modules, the Integration Module determines the object indicated by the person. However, these probabilities are not equivalent. For example, the usage region of reference terms varies between individuals so that the certainty of the probability is low. On the other hand, the certainty of the pointing gesture is higher than that of the reference term. Thus, the module uses weighted averaging toward the three probabilities and based on the result, it determines which object the human is indicating.

PSM: A module representing the Pointing Space Model

The PSM calculates the probability, $P_{PSM}(\theta_p)$, that the person's pointing gesture indicates each object in the environment based on the Pointing Space Model and by using the 3D positions of each object from the motion-capturing system. The probability function $P_{PSM}(\theta_p)$ is based on the normal distribution function and it changes according to the angle from the person's pointing direction θ_p (See **Fig. 3**) as shown in **Fig. 10**. The smaller the angle is, the higher the probability. The PSM outputs this probability to the Integration Module.

RTM: A module representing the Reference Term Model

The RTM calculates the probability that the reference term uttered by the person indicates each object in the environment based on the Reference Term Model, by using 3D positions of each subject from the motion-capturing system. In this model, $P_{kore}(x)$ is the probability that the object is indicated by the reference term *kore*, $P_{sore}(x)$ is the probability that the object

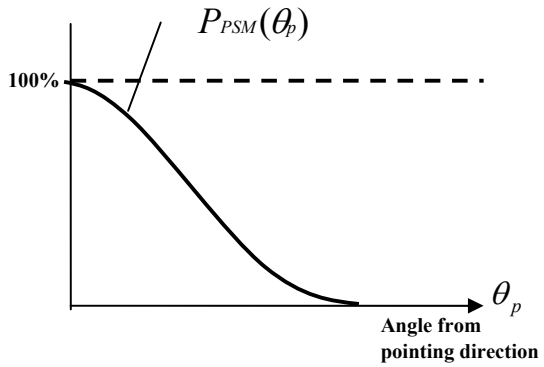


Fig. 10: PSM Probability Distribution

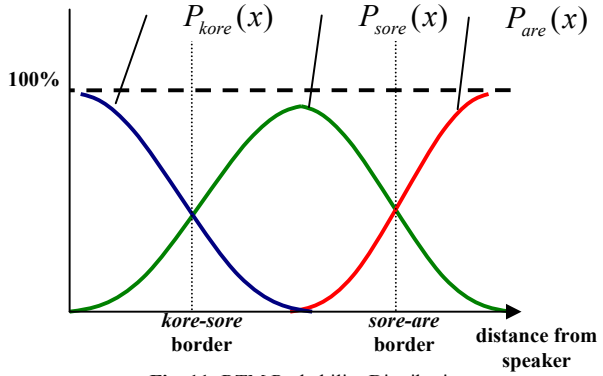


Fig. 11: RTM Probability Distribution

is indicated by *sore*, and $P_{are}(x)$ is the probability that it is indicated by *are*. These probabilities are calculated based on the normal distribution function and they change, for example, depending on the distance from the speaker, as shown in **Fig. 11**. RTM outputs $P_{kore}(x)$ if the reference term recognized in speech is *kore*, $P_{sore}(x)$ if it is *sore*, and $P_{are}(x)$ if it is *are*.

OPM: A module representing the Object Property Model

The OPM calculates the probability that the properties of the object uttered by the person indicate each object in the environment based on the Object Property Model. In this system, there are only two color properties, white and black, attached to the object, thus the property is the alternative of 100% or 0%. However, adjectives such as long and short are relative expressions depending on the objects in the environment. For example, comparing the longest object and the medium-length object in the environment, the former's probability will be higher than that of the latter when the person says, "That long one." This case applies to many adjectives such as big and small e.g. It will be future work to calculate the probabilities of each adjective.

Integration Module

The Integration Module combines the inputs from sub-model modules and determines the object indicated by the person. As described before, the probabilities from sub-model modules are not equivalent, thus we apply weighted averaging to them and calculate the probability that a pointing gesture and verbal cues indicate each object in the environment. The probability, P_{O_i}, is given by the following equation:

$$P_{O_i} = w_1 P_{PSM} + w_2 P_{RTM} + w_3 P_{OPM}$$

$$\begin{cases} w_1 + w_2 + w_3 = 1.00 \\ w_1 = 0.50, w_2 = 0.17, w_3 = 0.33. \\ w_1 > w_3 > w_2 \end{cases}$$

D. Object Referring Unit

The Object Referring Unit (**Fig. 6 (b)**) is based on the system proposed and explained in the previous works [5] in more detail. This unit selects an appropriate verbal cue and generates an attention-drawing behavior to confirm the identity of the object recognized. The unit consists of three-layer modules, the PSM, RTM, OPM, and the Integration Module. The PSM determines whether to use the object property with a reference term by a pointing gesture. (The PSM was described as LDM in [5].)

The PSM first calculates a Limit Distance based on θ_p (See **Fig. 3**) and the size of the indicated object. The Limit Distance is the distance, if other objects are present, within which the object cannot be identified only by using a pointing gesture and a reference term. If the distances between the indicated object and the other objects are within the Limit Distance, the PSM outputs its decision that the system should use the property of the object with a reference term to the Integration Module. However, if the distances between the indicated object and the other objects are all outside the Limit Distance, the PSM outputs its decision that the system can identify the object only by using a pointing gesture and a reference term.

The RTM determines an appropriate reference term to identify the object based on the border functions that divide the usage of reference terms, as shown in **Fig. 4**, made in the previous study [5], and it outputs its decision to the Integration Module.

The OPM is only used when the PSM determines it is necessary to use a property of the object. In that case, the OPM determines an appropriate property for identifying the object by comparing the properties of different objects with each other. However, at present this system only features color information as a property of the objects. If it is impossible to identify the object with color information, the OPM selects Japanese characters attached to the objects to identify the object.

The Integration Module selects an appropriate verbal cue based on the output from three-layer modules and generates an attention-drawing behavior with which the robot can confirm the indicated object with the person.

V. Experiment

[Hypotheses] To verify the system's effectiveness, we conducted an experiment in which we verified the following two hypotheses:

Hypothesis 1: The robot can behave as if it understands which object the human refers to in the environment by using the system.

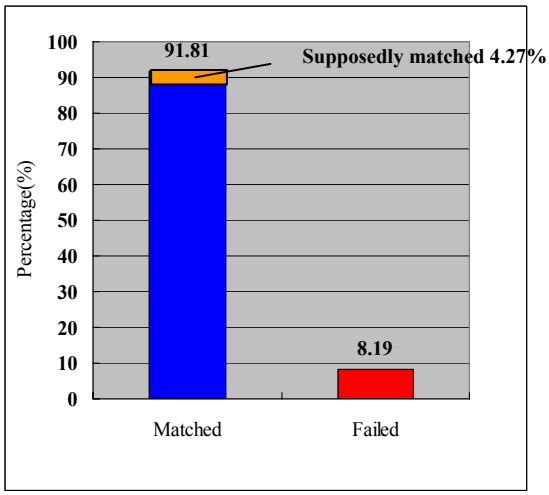
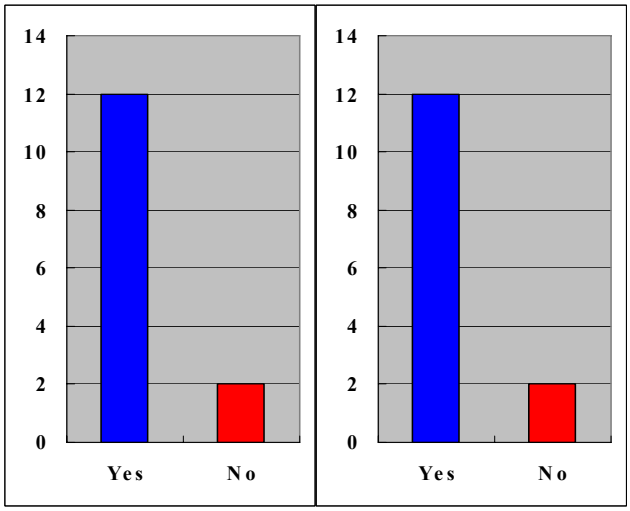


Fig. 12: Object match rate in the experiment



a) Can robot understand what you indicated?

b) Was interaction with robot natural?

Fig. 13: Data from questionnaire following experiment

Table I: Utterance selection in the experiment

R.T. only	***Kore***	***Sore***	***Are***
White + R.T.	***Kono-Shiroino***	***Sono-Shiroino***	***Ano-Shiroino***
Black + R.T.	***Kono-Kuroino***	***Sono-Kuroino***	***Ano-Kuroino***

R.T.: Reference Term, White: ***Shiroi***, Black: ***Kuroi***

Hypothesis 2: Subjects feel the conversation with the robot is natural.

[Brief Overview] This experiment was conducted to verify whether the constructed system could recognize the object indicated by a person and could interact naturally with him or her. There were five round objects in the environment, three were white and two were black. The subjects could arrange these objects freely. After arranging the objects, each would point out one of the objects to the robot by using a pointing gesture and verbal cues. The robot recognized the human indication and confirmed the object also by using a pointing gesture and verbal cues (**Fig. 1**). In every trial, subjects evaluated whether the robot seemed to understand the indicated object.

[Subjects] Fourteen of our colleagues at ATR participated as subjects.

[Experimental Procedures] The experiment consisted of four sessions. In every session, the subject followed the procedure as follows:

1. Arrange the objects freely in the environment except in positions that the robot cannot point out.
2. Point out one of the objects using a pointing gesture and one of the verbal cues given in Table I.
3. Evaluate whether the robot seems to understand the object to which you are pointing by selecting one of three entries, Matched, Supposedly matched, or Failed.
4. Repeat steps 2 and 3 for all objects in the environment.

[Verification Method] To verify Hypothesis 1, we calculated the rate at which the subjects evaluated whether the robot could understand the indicated objects. We also compiled a questionnaire for the subjects to complete after the experiment. The questionnaire contained the two following items:

Item 1: Does the robot seem to understand which object you indicated?

Item 2: Do you think the interaction with the robot was natural, as in inter-human communication?

Subjects answered with either a Yes, or a No. We verified Hypothesis 1 with Item 1 and Hypothesis 2 with Item 2.

[Experimental results] Figure 12 illustrates the "matched" and "failed rates" as evaluated by the subjects. While, The questionnaire data is shown in **Fig. 13**.

[Verification of Hypothesis 1] As **Fig. 12** shows, the rate at which the subjects evaluated the robot as understanding the indicated object was 91.81% of the time. With this result we can confirm that this system enabled the robot to behave as if it understands which object was the indicated one. Furthermore, a chi-squared test for the number of subjects who answered with Yes in Item 1 of the questionnaire was significantly high ($x^2_{(1)} = 7.142, p < .01$). With this result also, we can confirm Hypothesis 1.

[Verification of Hypothesis 2] Through a chi-squared test, we found that the number of subjects who answered with Yes in Item 2 of the questionnaire was also significantly high ($x^2_{(1)} = 7.142, p < .01$). With this result, we can confirm that the subjects feel conversation is natural when this system is used.

[Discussion of Experimental results] In this experiment, we let the subjects arrange the objects freely in the environment.

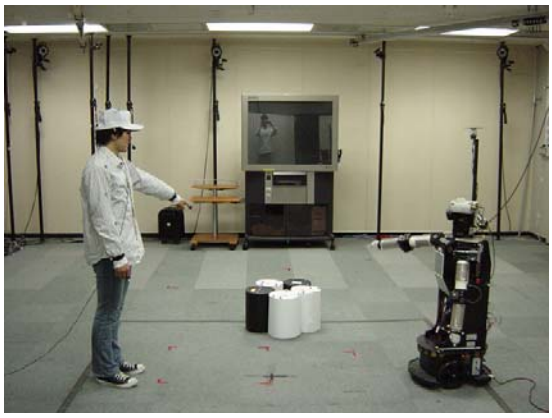


Fig. 14: An impossible situation in the experiment
(The "matched" rate was low in such a case)

Thus the "matched" rate varied from high to low depending on the objects' locations. Figure 14 illustrates one of the worst cases, where the robot misunderstood some objects in the object group. However, the subject who arranged the objects in **Fig. 14** answered in the questionnaire that he felt that it is natural that the robot should make mistakes in a situation that is supposed to be difficult even for humans. This answer shows that, in the case that a human cannot understand the indicated object either, the robot's error is not critical.

The other remaining problem for recognizing which is the indicated object is the timing required for capturing a person's pointing gesture. In this system, we capture the posture of the human body as a pointing gesture when the verbal cues are recognized by speech recognition. However, there were some cases of incorrect timing. To communicate with people in a living environment, this problem is highly critical for humanoid robots and will be a future work to measure this timing. Finally, in the questionnaire the subjects answered how eye contact and the pause until the robot indicates the object influence the naturalness of interaction. Future work will also involve finding and adjusting natural eye contact and pauses.

VI. Discussion

Summary of the results

The experimental results revealed that the robot successfully recognized the attention-drawing behavior of the subjects, and that it correctly drew their attention to the indicated object as well. Thus, the effectiveness of the three-layer model for generation and recognition of an attention-drawing behavior was verified.

Potential advantage of shared models for generation and recognition

We believe that our method holds great potential. Since the Object Recognition Unit and the Object Referring Unit share the same models (PSM, RTM and OPM), adjustment of the models will contribute to better performance in both recognition and referral. Furthermore, it may be possible to adjust the models through interaction with people. In particular, RTM will require a large amount of adjustment, since the use of reference terms changes depending on the size of the environment, objects to be referred to, and individuals. When we prepared the RTM, we had 20 subjects perform about 200 trials of using reference terms [1]. It was partly due to individual differences among subjects: at certain positions, one person might say *kore* to refer to an object, but another might say *sore*. Thus, it is better to perform such a number of adjustments through interaction. This should be included in our future work.

Advantage of having the PSM layer

Of course, there are individual differences among people in the accuracy of their pointing gestures. Some people accurately point at an object, while others do so roughly. Thus, this layer allows us to adjust the PSM parameters for each person.

Individual difference among robots will also affect our model. Because appearance probably affects the robot's pointing gesture, we may need to adjust the Pointing Space Model (PSM). The Robovie-type robot had a relatively weak effect in its pointing gesture, leading us to adjust the limit distance in the PSM to 10 degrees. Of course, this parameter is hardware dependent; Robovie only has one pointing finger on a spherical hand. Several current humanoid robots feature five-fingered hands, and there are also android robots equipped with very human-like hands with five fingers. We believe that such a sophisticated hand will have a more powerful effect in pointing gestures. Consequently, we will be able to use a smaller angle as θ_p (See **Fig. 3**) in the PSM.

Thus, by having the separate PSM layer, we can independently handle the individual differences among people and robots in the PSM.

Advantage of having the RTM layer

The Reference Term Model (RTM) is language-dependent. In Japanese, there are three reference terms: *kore*, *sore*, and *are*. The usage borders between these reference terms are associated with the positions of the speaker, the listener, and the object. In English, for example, there are two main reference terms, "this" and "that" (as well as "over there"), whose usage border is only associated with the position of the speaker and the object. We believe that such a dependency can be implemented in the RTM. In other words, the developed attention-drawing model is probably capable of operating in other languages by switching the current RTM to one for a different language.

Advantage of having the OPM layer

It is not our intention to emphasize the detailed mechanism in the OPM. Rather, traditional research in artificial intelligence and linguistics addressed this issue; that is, how to identify the target object without using reference terms or gestures. Our model is based on the idea to use minimum necessary words in addition to the reference terms. In other words, the OPM allows us to focus on the robot-oriented research within the PSM and RTM.

VII. Conclusion

We have implemented a three-layer model for generation and recognition of attention-drawing behavior into the humanoid robot, Robovie. It enabled the robot to recognize people's attention-drawing behavior with pointing gestures and reference terms. Also, the robot is capable of drawing people's attention. The effectiveness of the three-layer model was verified through an experiment, demonstrating that the system could identify attention-drawing behavior from people except in some impossible situations. The reaction of the robot was also evaluated as mostly natural. One of this system's most promising features is that the recognition and generation parts share the same model. For example, there is the potential to dynamically adjust precise parameters in the model through interaction with people.

Acknowledgment

This research was supported by the Ministry of Internal Affairs and Communications of Japan.

References

[1] C. Breazeal and B. Scassellati: "Infant-like social interactions between a robot and a human caretaker," Adaptive Behavior, 8(1), 2000.

[2] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano: "Real-Time Auditory and Visual Multiple-Object Tracking for Robots," Proc. Int. Joint Conf. on Artificial Intelligence, pp.1425-1432, 2001.

[3] M. Kamashima, T. Kanda, M. Imai, T. Ono, D. Sakamoto, H. Ishiguro, and Y. Anzai: "Embodied Cooperative Behaviors by an Autonomous Humanoid Robot," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2004), pp. 2506-2513, 2004.

[4] C. Moore and Philip J. Dunham eds: "Joint Attention: Its Origins and Role in Development," Lawrence Erlbaum Associates, 1995.

[5] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita: "Three-layered Draw-Attention Model for Humanoid Robots with Gestures and Verbal Cues," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2005), pp. 2140-2145, 2005.

[6] H. Kozima and E. Vatikiotis-Bateso: "Communicative criteria for processing time/space-varying information," Proc. 10th IEEE International Workshop on Robot and Human Communication, IEEE, pp. 377-382.

[7] Y. Nagai: "Learning to Comprehend Deictic Gestures in Robots and Human Infants," In Proceedings of the 14th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN'05), pp. 217-222, August 2005.

[8] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin: "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2005), pp. 383-388, 2005.

[9] T. Mizuno, Y. Takeuchi, H. Kudo, T. Matsumoto, N. Onishi, and T. Yamamura: "Informing a Robot of Object Location with Both Hand-Gesture and Verbal Cues," IEEJ Trans. EIS, Vol. 123, No. 12, 2003 (Japanese)

[10] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer: "A multi-modal object attention system for a mobile robot," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2005), pp. 1499-1504, 2005.

[11] Z. M. Hanafiah, C. Yamazaki, A. Nakamura and Y. Kuno: "Understanding Inexplicit Utterances Using Vision for Helper Robots," Proceedings of the 17th International Conference on Pattern Recognition, /CD-ROM V44_2_04.pdf, Cambridge, UK, August 23-26, 2004.

[12] T. Inamura, M. Inaba, and H. Inoue: "PEXIS: Probabilistic Experience Representation Based Adaptive Interaction System for Personal Robots," Systems and Computers in Japan, Vol. 35, No. 6, pp. 98--109, 2004.

[13] B. Scassellati: "Investigating Models of Social Development Using a Humanoid Robot," Biorobotics, MIT Press, 2000.

[14] M. Imai, T. Ono, and H. Ishiguro: "Physical Relation and Expression: Joint Attention for Human-Robot Interaction," IEEE Transactions on Industrial Electronics, Vol. 50, No. 4, ITIED 6, pp. 636-643, 2003

[15] S. Kuno: "The Structure of the Japanese Language," MIT Press, 1974.

[16] T. Kanda, H. Ishiguro, M. Imai, and T. Ono: "Development and Evaluation of Interactive Humanoid Robots," Proceedings of the IEEE Vol. 92, No. 11, pp. 1839-1850, 2004.