Three-Layered Draw-Attention Model for Humanoid Robots with Gestures and Verbal Cues

Osamu Sugiyama, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro and Norihiro Hagita
ATR Intelligent Robotics and Communication Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, JAPAN
sugiyama@atr.jp

***Abstract*— When we talk about objects in an environment, we indicate to a listener which object is currently under consideration by using pointing gesture and such reference terms as "this" and "that". Such reference terms play an important role in human interaction by quickly informing the listener of an indicated object's location. In this research, we propose a three-layered draw-attention model for humanoid robots with gestures and verbal cues. Our proposed three-layered model consists of three sub models: Reference Term Model (RTM), Limit Distance Model (LDM) and Object Property Model (OPM). RTM decides an appropriate reference term using functions constructed by an analysis of human behavior. LDM decides whether to use the object's property with a reference term. OPM decides the appropriate property for indicating the object by comparing object properties with each other. We developed an attention drawing system in a communication robot named "Robovie" based on the three layered model. We confirmed its effectiveness through the experiments.**

***Index Terms*— human-robot interface, human-robot interaction, Humanoid robot**

I. INTRODUCTION

Recently, many robots have begun working in such living and societal environments as guide robots in museums [1] and nursing robots for the elderly [2]. What such robots share is that their targets are people without specialized computing and engineering knowledge. For these robots, a conversational interface using physical expressions and verbal cues is becoming more important as a universal human-robot interface. Our research aims to develop this interface for the interaction between humans and robots by focusing on a method that enables robots to interact with humans and objects in the environment.

Pointing is one social cue collectively called the mechanism of Joint Attention that indicates to the listener which object is currently under consideration[3]. Reference terms are words that inform the listener of the object's location in the environment [4]. We focus attention to use such pointing gesture and reference terms in robotic interface to draw human attention.

There are precedent researches for a robotic interface using pointing and reference terms. Imai et al. implemented conversation robots that indicate objects using gestures and reference terms [5]. Mizuno et al. proposed a method that enable robots to figure out object location with both hand gestures and verbal cues [6]. However, these researches did not adequately solve three problems: First, choosing verbal cues depends on the locations of the robot, the listener and the objects; second, the definition of the object when objects get close together; and third, identifying objects that have similar properties with others. Concerning the first problem, since Mizuno et al. proposed method only identified the area where verbal cues refer an object depending on the distance from robot or human to the object, the method is inadequate to decide verbal cues when the speaker and listener get close together. For the second problem, neither research sufficiently treated a situation and only offered limited solutions to solve it. About the third problem, neither research treated such a situation, so there are no solutions for it.

Our research aims to enable robots to indicate to humans which object is under consideration even if these three problems occur. To realize our aim, we outline a three-layered draw-attention model that has several variations of verbal cues, including reference terms, which can select cues depending on conditions. To select appropriate cues, we set up 3 sub models, RTM, LDM, and OPM based on observation of human behavior. We developed an attention drawing system for humanoid robots based on the model.

II. CONVERSATION USING REFERENCE TERMS

First, we observe human behavior, especially a conversation using reference terms, to define a draw-attention model.

A. Reference terms in Japanese

In Japanese, we use three reference terms to inform people of an object's location: "KORE", "SORE", and "ARE". They correspond to "this" and "that" in English. "KORE" refers to an object close to the speaker, and "SORE" refers to an object close to the listener or in the middle between the two individuals. "ARE" refers to an object that is neither close to the speaker nor to the listener. However, these definitions are so vague that we have to set up some border to divide the usage of reference terms.

B. Analysis of Human Behavior

1) Conversation using reference terms:

To investigate the spatial factors that decide which reference term should be used, we conducted an experiment for observing a conversation between humans, as shown in Fig. 1.

Fig. 1. Analysis of human behavior

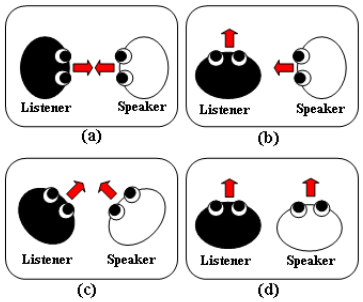

Fig. 2. Direction of speaker and listener in analysis of human behavior

[Brief Overview] We set up an environment where a speaker, a listener and an object exist. In the experiment, a speaker asked a listener to take an object using one of the following words, "KORE-TOTTE (take this)","SORE-TOTTE (take that)" and "ARE-TOTTE (take that)" in Japanese. We changed the object's location and recorded which reference term the speaker used in each trial.

[Subjects] The subjects were 20 male and female university students. They were divided into 10 pairs: one played the speaker role and the other did the listener role.

[Experiment Procedure] Experiments consisted of six sessions. In each session, we changed the distance between speaker and listener and their direction. The distance had three variations: 50 cm, 1 m and 2 m. There are 4 variations of direction about speaker and listener as shown in Fig. 2.

[Analysis Method] We calculated the probability of each reference term used in each object's location. Based on these probabilities, for each angle we calculated the border distance where the probabilities of both reference terms became 50%.

[Analysis Results] Fig. 3 (white and black circles stand for speaker and listener, respectively) shows the borders among three regions: "KORE" region, "SORE" region and "ARE" region. When objects are in "KORE" region, speaker calls the objects "KORE". Same applies to "SORE" region and "ARE" region. Both borders are elliptical shaped and have either the speaker or the speaker and listener at their center.

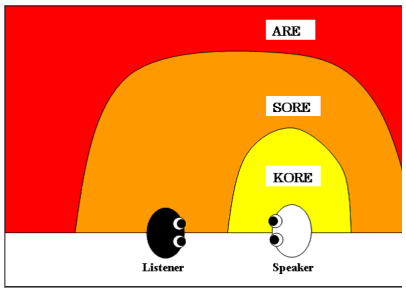

Fig. 3. Reference Term Model

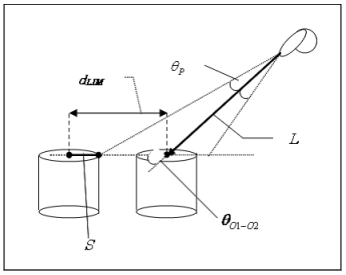

Fig. 4. Limit Distance Model

2) Conversation when two or more objects are close together:

When two or more objects in the environment are close together, it is difficult for listener to identify the object to only from pointing gesture and reference term. We define Limit Distance as the distance in which we cannot distinguish the indicated object from the other objects only using pointing gesture and reference term.

The definition of Limit Distance is shown in Fig. 4. The definition shows that listener cannot distinguish the indicated object, if the edge of the other object intrudes into the scope of θ_P from the indicated direction. In other words, Limit Distance d_{LIM} is the distance that includes scope θ_P and distance S from the object's center to its edge.

The difficulty to distinguish the indicated object from the other objects is dependent on the distance between one object to another object, the distance from objects to the speaker, and the size of the object.

3) Additional adjectives to identify the indicated object:

When objects get close together, we identify the indicated object using its property as an adjective with a reference term. In this research, we adopted this method to identify indicated objects when objects get within the Limit Distance. To identify the indicated object, its property should be different from the others. Speaker can find several properties in one object, such as shape, color, and amplitude. Using a property or a property set different from the others, he can indicate to the listener which object is under consideration. An example of object property selection is shown in Fig. 5. There are two pens in the environment. Because they are close together, listener cannot identify one from the other simply by pointing gesture and a reference term. Pen A has the following

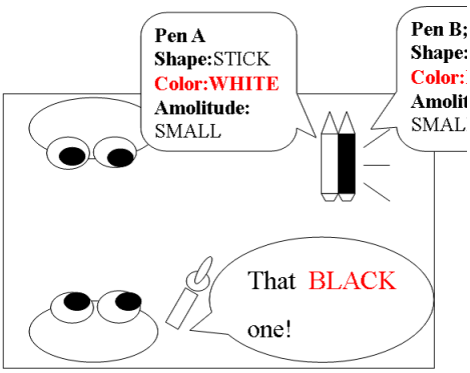

Fig. 5. Object Property Model

properties: Shape: STICK, Color: WHITE, and Amplitude: SMALL. Pen B has the following properties: Shape: STICK, Color: BLACK, Amplitude: 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".

III. MODEL'S PROPOSAL TO DRAW HUMAN ATTENTION

A. An overview of the model

Based on the analysis of human behavior, we propose a three-layered model for humanoid robots using pointing gesture and verbal cues as shown in Fig. 6. The model consists of three sub models: RTM, LDM and OPM. Selecting appropriate verbal cues based on three sub models, the model can restrict an indicated object. The sub model roles are as follows:

- **[RTM]** Select an appropriate reference term.
- **[LDM]** Estimate whether listener can identify object by pointing gesture and a reference term.
- **[OPM]** If the indicated object cannot be restricted, select an appropriate object's property to use with the reference term to identify the indicated object.

B. RTM: Reference Term Model

RTM determines an appropriate reference term to identify the indicated object based on the approximate curve functions of Fig. 3. The functions are as follows:

$$f_{KS}(d_{SL}, d_{SO}, \theta_L) =$$
$$\begin{cases} 1.3 - max\_subtract \times |cos\theta_{SO}|^{curve\_adjust} (\theta_{SO} \leq 90) \\ 1.3 - max\_subtract \times |cos(\theta_{SO} - 22.5)|^{curve\_adjust} \\ (\theta_{LO} \geq 90) \end{cases}$$
$$max\_subtract = f(d_{SL})$$
$$curve\_adjust = f(d_{SL}, \theta_L)$$

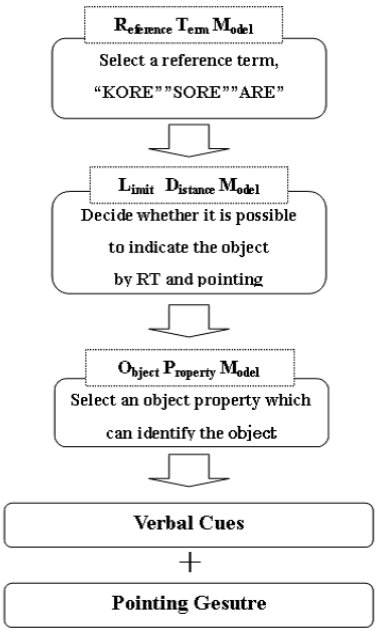

Fig. 6. Flowchart of 3-layered Draw Attention Model

$$f_{SA}(d_{SL}, d_{SO}, d_{LO}, \theta_{LO}, \theta_L, \theta_S) =$$
$$\begin{cases} 2.0 + sin(90 - \theta_{SO}) (\theta_{SO} \leq 90) \\ 2.0 + 1.0 \times cos^2(\theta_{SO} - 22.5) (\theta \geq 90) \end{cases}$$
$$(d_{SO} \leq d_{LO})$$
$$\begin{cases} r + max\_addition + sin(90 - \theta_{LO}) (\theta_{LO} \leq 90) \\ r + max\_addition \times sin^{curve\_adjust}\theta_{LO} \\ (\theta_{LO} \geq 90) \end{cases}$$
$$(d_{SO} \geq d_{LO})$$
$$max\_addition = f(d_{SL}, \theta_L)$$
$$curve\_adjust = f(d_{SL}, \theta_S)$$

where $max\_subtract$ is the maximum subtractive value and $max\_addition$ is the maximum additional value that makes the curve elliptically, and $curve\_adjust$ is the value that adjusts the curve change rate. These parameters are not static but variable based on the parameters defined in Fig. 7. $f_{KS}(d_{SL}, d_{SO}, \theta_L)$ are the approximate curve functions that formularize the elliptical shape around speaker in Fig. 3. While, $f_{SA}(d_{SL}, d_{SO}, d_{LO}, \theta_{LO}, \theta_L, \theta_S)$ are the approximate curve functions that formularize the outer elliptical shape around speaker and listener. RTM determines which reference term to use by the following rules:

- If $d_{SO} \leq d_{LO}$
  - Use "KORE" when $d_{SO} \leq f_{KS}(d_{SL}, d_{SO}, \theta_L)$
  - Use "SORE" when $f_{KS}(d_{SL}, d_{SO}, \theta_L) \leq d_{SO}$ and $d_{SO} \leq f_{SA}(d_{SL}, d_{SO}, d_{LO}, \theta_{LO}, \theta_L, \theta_S)$
  - Use "ARE" when $d_{SO} \geq f_{SA}(d_{SL}, d_{SO}, d_{LO}, \theta_{LO}, \theta_L, \theta_S)$
- If $d_{SO} \geq d_{LO}$
  - Use "SORE" when $d_{LO} \leq f_{SA}(d_{SL}, d_{SO}, d_{LO}, \theta_{LO}, \theta_L, \theta_S)$

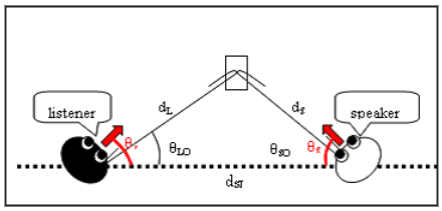

Fig. 7. Parameters of approximate curves

– Use "ARE" when
$d_{LO} \geq f_{SA}(d_{SL}, d_{SO}, d_{LO}, \theta_{LO}, \theta_L, \theta_S)$

C. LDM: Limit Distance Model

LDM estimates whether or not the listener can identify the indicated object with a pointing gesture and a reference term based on Limit Distance. Limit Distance d_{LIM} is given by

$$d_{LIM} = f(S, L, \theta_P) = \frac{tan\theta_P(L + cos\theta_{O1-O2})}{sin\theta_{O1-O2}}$$

where the parameters are described in Fig. 4. LDM decides whether the model uses an object's property by the following rules:

- Use object's properties, when $d_{LIM} \leq f(S, L, \theta_P)$
- Don't use, when $d_{LIM} \geq f(S, L, \theta_P)$

D. OPM: Object Property Model

OPM chooses an indicated object's property different from the other objects, which are all in the Limit Distance. OPM has a list of the object properties of each object, and by comparing each property among objects, the model finds the appropriate property. In this research, OPM, however, has only color as an object property. A system to get object properties and an algorithm to compare them are under consideration.

IV. DEVELOPMENT OF THE THREE-LAYERED DRAW-ATTENTION MODEL

A. Hardware Configuration

An attention drawing system based on our three-layered model is developed in a communication robot "Robovie"[7] with a motion capture system [1].

"Robovie" (Fig. 8) is 1.2 m tall with a 0.5 m radius and a human-like upper body to communicate with humans. It has a head (3DOF), 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 in a gesture similar to humans.

A motion capture system (Fig. 9) can get a 3D position from markers attached to the targets, Robovie, a subject, and objects. Using 100 Mbps Ethernet, the developed system gets the 3D positions of subjects as input from the motion capture system and calculates their location.

B. System Configuration

Figure 10 shows the configuration of the developed system. The roles of each module in Fig. 10 are as follows:

[1]http://www.vicon.com/

Fig. 8. Robovie

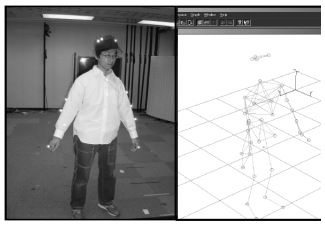

Fig. 9. Motion Capture System

1) RTM module: This module receives the 3D position of Robovie, a subject, and objects from a motion capture system. Then it decides which reference term to use based on RTM and outputs it to the Integration module

2) LDM module: This module receives the 3D positions of Robovie and the objects from the motion capture system as input and decides whether to use object property to identify the indicated object based on LDM. It output the decision to the OPM module and Integration Module.

3) OPM module: If the LDM module decides to use an object's property to restrict the object, it selects the appropriate object's property and identifies the object based on OPM. It output the decision to the Integration module.

4) Integration Module: This module chooses verbal cues from the Language database to identify the object based on input from the RTM, LDM and OPM modules. It output the result to the actuated module. As a result, Robovie orally gives the verbal cue.

V. EXPERIMENTS

To confirm the effectiveness of the developed system, we conducted the following three experiments:

1) Verification of the effectiveness of RTM.
2) Verification of the effectiveness of LDM.
3) Verification of the effectiveness of the system.

A. Experiments to verify the effectiveness of RTM

[Brief Overview] We verified RTM effectiveness in this experiment. Every three minutes Robovie repeated a reference term selected by RTM that indicated the moving object. Subjects evaluated each spoken reference term.

[Subjects] Subjects were 13 male and female university students.

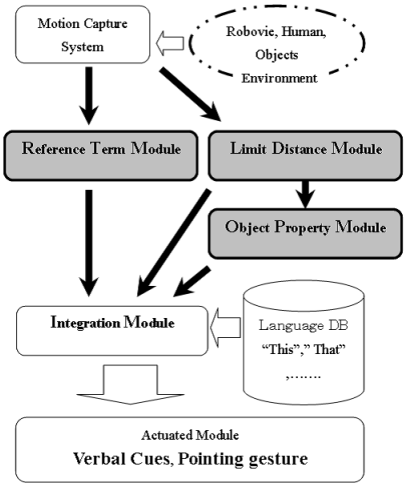

Fig. 10. System Configuration

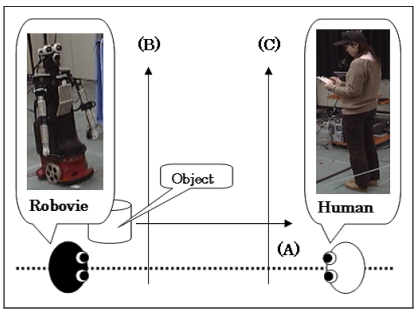

Fig. 11. Directions an object moved in experiment to verify RTM effectiveness

[Experiment Procedure] The experiments included Robovie, who repeated a reference term every 3 seconds, a subject, and an object moved in one direction. The experiment consisted of three sessions, and in every session we changed the direction from (A) to (C) (Fig. 11) to which the object was moved. Subjects evaluated the acceptability of the reference term and completed the questionnaire at that time.

[Hypothesis] We tested whether a hypothesis that with RTM the system can choose an appropriate reference term for participants to understand the object's location.

[Verification Method] Subjects evaluated the acceptability of the reference term by tri-level evaluation.

[Verification of hypothesis] The questionnaire data are shown in Table I. As Table I shows, the total probability of "No problem" and "Acceptable" is 96.02%. Consequentially, results clearly show that with RTM the system can choose an appropriate reference term for participants to understand the object's location.

B. Experiments to verify the effectiveness of LDM

[Brief Overview] We verified LDM effectiveness in this experiment, in which Robovie indicated to a subject one of two objects decided by LDM. Subjects answered which object the robot indicated.

TABLE I
QUESTIONNAIRE DATA IN EXPERIMENTS TO VERIFY RTM EFFECTIVENESS

Entries	Recognition
No problem	81.95%
Acceptable	14.07%
Unacceptable	3.97%

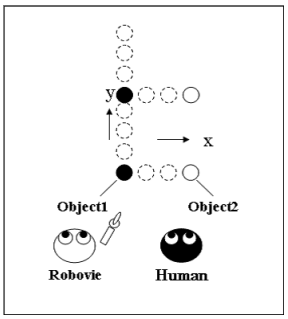

Fig. 12. Position of two objects in experiment to verify LDM effectiveness

[Subjects] Subjects were 10 male and female university students.

[Experiment Procedure] Robovie indicated to a subject one of two objects decided by LDM. The indicated object was decided randomly. In each session the position of the two objects were moved, as shown in Fig. 12.

[Hypothesis] We tested a hypothesis that with LDM the system can identify the area where people can define the indicated object with pointing gesture and verbal cues.

[Verifying Method] We verified the hypothesis by the recognized rate of the indicated object.

[Hypothesis Verification] As a result, the recognized rate in the certified area by LDM was 83.3%. The rate out of the certified area was 56.6%. Consequently, with LDM the system can identify the area where people can identify the indicated object with pointing gesture and verbal cues.

C. Experiments to verify the system effectiveness

[Brief Overview] In this experiment shown in Fig. 13, we verified the effectiveness of the attention drawing system. Robovie indicated to subjects one of five objects in the environment. The verbal cues that the system used have two variations.

- Verbal cues with reference term, which are decided by our proposed three-layered model.
- Verbal cues with three kinds of symbol written on the object (number, alphabet and Japanese symbol) that can identify each object

The 2nd cues are set up for comparing the effectiveness of reference terms and other verbal cues. Subjects answered the object they thought the system indicated.

[Subjects] Subjects were 21 male and female university students.

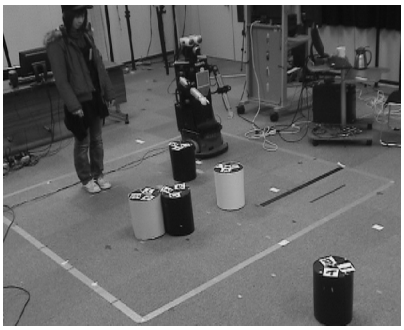

Fig. 13. Experiments to verify system effectiveness

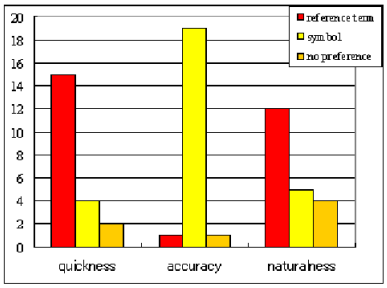

Fig. 14. Questionnaire data in experiments to verify system effectiveness

[Hypothesis] In this experiment, we tested the following two hypotheses.

Hypothesis1: The system can correctly draw listener's attention to the indicated object among objects in the environment

Hypothesis2: Using reference terms, the system can quickly draw listener's attention to the indicated object.

[Verifying method] We verified Hypothesis1 by the recognition rate of the indicated object and the Hypothesis2 by questionnaire data compiled after the experiment. The questionnaire entries were as follows:

- With reference terms or symbols, which can you quickly recognize the indicated object?
- With reference terms or symbols, which can you correctly recognize the indicated object?
- With reference terms or symbols, which seems to have more natural expressions?

Subjects choose from the following answers: (1) Reference term, (2) Symbol, and (3) no preference.

[Experimental Results] The recognized rate of the indicated object reached 93.33% if we used the verbal cues selected by the model. The other reached 92.38% if we used symbols to identify the indicated object. Questionnaire data are shown in Fig. V-C. Through a chi-square test, entries of quickness ($\chi 2(2) = 14.000, p < 0.01$) and accuracy ($\chi 2(2) = 30.853, p < 0.01$) differ significantly. For quickness, the number of subjects who answered with reference term was significantly high. On the other hand, for accuracy, the number of subjects who answered with symbols was significantly high.

[Verification of hypothesis1] The recognized rate shows that both a symbol with pointing and a reference term with pointing could indicate the object in the environment.

[Verification of hypothesis2] The results of the chi-square test for the questionnaire data clearly shows that the system can quickly draw listener's attention to the indicated object by using reference terms.

VI. CONCLUSION

In this research, we proposed a three-layered draw-attention model for a humanoid robot to indicate to listeners which object is under consideration using an appropriate verbal cue with pointing gesture. The three-layered model consists of three sub models: RTM (Reference Term Model), LDM (Limit Distance Model) and OPM (Object Property Model), with which the robot can select an appropriate cue to use. We developed an attention drawing system for Robovie based on this model. Experiments were conducted for verifying the effectiveness of these sub models and the developed system. As a result, it revealed the following issues:

- With RTM, the system can choose an appropriate reference term for participants to understand the object's location.
- With LDM, the system can identify the area where people can define the indicated object with pointing gesture and verbal cues.
- The system can correctly draw listener's attention to the indicated object among objects in the environment
- By using reference terms, the system can quickly draw listener's attention to the indicated object.

ACKNOWLEDGMENT

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

REFERENCES

[1] Wolfram Burgard and Armin B. Cremers and Dieter Fox and Dirk Ahnel and Gerhard Lakemeyery and Dirk Schulz and Walter Steiner and Sebastian Thrun, "The Interactive Museum Tour-Guide Robot", Proc. of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), 1998

[2] M. Montemerlo and J. Pineau and N. Roy S. Thrun and V. Verma, "Experiences with a Mobile Robotic Guide for the Eldery", 18th National Conference on Artificial Intelligence, pp. 587-592, 1999

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

[4] Michita Imai, Testuo Ono, Hiroshi 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, (2003-8).

[5] Michita Imai, Kazuo Hiraki, Tsutomu Miyasato., "Physical Constraints on Human Roboto Interaction", Proceedings of 16th International Joint Conference on Artificial Intelligence (IJCAI99), Vol.2, PP.1124–1130, 1999

[6] Tomoyasu Mizuno and Yoshinori Takeuchi and Hiroaki Kudo and Tetsuya Matsumoto and Noboru Onishi and Tsuyoshi Yamamura, "Informing a Robot of Object Location with Both Hand-Gesture and Verbal Cues", IEEJ Trans. EIS, Vol123, No.12, 2003

[7] Takayuki Kanda and Hiroshi Ishiguro and Tetsuo Ono and Michita Imai and Ryohei Nakatsu, "Development and Evaluation of an Interactive Humanoid Robot "Robovie"", IEEE International Conference on Robotics and Automation (ICRA 2002), 1848-1855, 2002