Noname manuscript No.
(will be inserted by the editor)

Can using pointing gestures encourage children to ask questions?

Tsuyoshi Komatsubara · **Masahiro Shiomi** · **Takayuki Kanda** · **Hiroshi Ishiguro**

Received: date / Accepted: date

Abstract Even though asking questions is fundamental for self-motivated learning, children often have difficulty verbalizing them. Hence, we hypothesized that a robot's capability to perceive pointing gestures will encourage children to ask more questions. We experimentally tested this hypothesis with the Wizard-of-Oz technique with 92 elementary-school students who interacted with our robot in a situation where it served as a guide who explains a museum exhibit. The children asked the robot significantly more questions when it could perceive pointing gestures than when it lacked such a capability. We also discuss the possibility of implementing autonomous robots based on the findings of our Wizard-of-Oz approach.

Keywords social robot · deictic interaction · robots for children

1 INTRODUCTION

Education and information-providing are promising applications for social robots. For instance, a guide robot has been developed that leads visitors and explains exhibits [1, 2]. Robots are used for such language-related activities as games and reading to help students learn language [3, 4]. In these previous works, however, interaction tended to be one way: a robot explains and people listen. This is partly because robots' perception is inadequate for recognizing human utterances and behaviors.

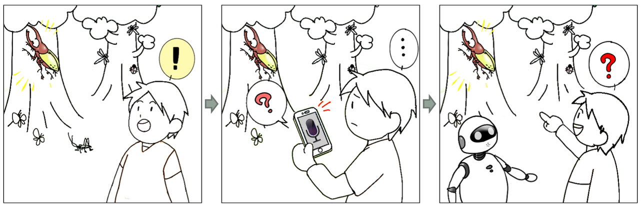

Fig. 1 By pointing, boy asks questions even when he know litle about his target of interest

On the other hand, it is also important to encourage people to actively seek information. Learning science advocates a learner-centered approach (e.g., [5]) that allows people to be self-motivated and to actively collect information by themselves. One previous HRI study investigated how to facilitate ways in which students can easily ask questions [6].

Since little is known about encouraging students to ask questions, our study addresses one such issue: deictic interaction. This type is particularly critical in situations when students lack knowledge about the target. For instance, assume a space shuttle exhibit in a museum. What kind of questions do we expect students to ask? Older, more knowledgeable students might ask:

"How fast does it fly?"

"How many times has it gone into space?"

However, younger, novice learners might only be able to ask:

"What's this?" (*pointing at its engine*).

Here, deictic interaction (pointing and reference terms, as illustrated in Fig. 1) plays an important role. For deictic interaction in HRI, past studies mainly focused on recognition methods that identified pointing gestures and the importance of using such gestures by robots for more natural interaction (Section 2); past studies focused less on the actual results

---

T. Komatsubara, M. Shiomi, T. Kanda
ATR, 2-2-2, Hikaridai, Keihanna Science City, Kyoto, 619-0288, Japan
Tel.: +81-774-95-1405
Fax: +81-774-95-1408
E-mail: {komatsubara.tsuyoshi, m-shiomi, kanda}@atr.jp

H. Ishiguro
Osaka University, 1-3, Machikaneyama, Toyonaka, Osaka, 560-8531, Japan
Tel.&Fax: +81-6-6850-6360
E-mail: ishiguro@sys.es.osaka-u.ac.jp

and the consequences of whether the robots actually did understand gestures in such education situations. Therefore, in this study, we confirmed our hypothesis that a robot that can perceive pointing encourages children to ask questions. In other words, this study investigates whether a robot's capability to perceive pointing gestures facilitates children to use them to ask questions.

2 RELATED WORK

2.1 Importance of Gestures for Children's Speech Activity

Gestures play an essential role in communication and thought because they help communicators and recipients during human-human interaction [7, 8]. Several research works reported that they facilitate lexical access and increase fluency [9,10]. Rauscher et al. concluded that "gestures function more at the level of speech production in helping the speaker to find the right words" [11].

In particular for school-age children, gestures elicit speech. For example, children's performances in counting tasks improved by pointing [12], and gestures can help both adults and children express information whose expression is difficult by speech [7, 13]. Pine et al. argued that children (6-8 years old) say the names of pictures more correctly when gestures are allowed instead of being prohibited [14]. Sauter et al. showed that children (8-10 years old) convey more detailed information about the locations of toys when they use gestures [15]. Matlen et al. suggested that the gestures of children (8-16 years old) facilitate the understanding of difficult spatial science concepts [16].

The research literature indicates the importance of gestures in speaking by school-age children, especially in situations where they lack the vocabulary to describe specific targets. Much research also reports how gestures facilitate adult communication (e.g., [17, 18]). Thus, we believe that using gestures will encourage children to communicate with robots and speculate that pointing gestures are particularly crucial for younger children because they often lack the language skills or vocabulary to discuss things they do not know well.

2.2 Deictic Interaction in HRI

Many previous studies clearly demonstrated the benefits of gestures from robots. For instance, Kuzuoka et al. developed a robot named GestureMan, which implements deictic interactions to facilitate remote interaction between co-located users and remote operators [19]. Scassellati implemented a joint-attention mechanism in a humanoid robot where a pointing capability was one important way for attention sharing [20]. A robot's pointing gesture increases the understanding of direction information [21]. Lohse et al. revealed that robot gestures improved the recall of difficult directions by users [22]. Sauppé and Mutlu investigated different types of deictic gestures and identified accurate and effective ones [23].

Moreover, several research works proposed methods that clarified which pronouns are appropriate for the deictic gestures of robots. For example, Ng-Thow-Hing et al. proposed a model that selects appropriate pronouns for a robot's gestures including a deictic gesture based on annotated data from video images [24]. Huang et al. also developed a model that selects appropriate narrative gestures including pointing gestures during storytelling based on the data of human narrators and evaluated it using a social robot [25]. Hato et al. proposed a deictic interaction model by referring to regions and analyzing how people refer to them. Their model enables a robot to appropriately refer to regions by pointing and with pronouns and highlights the importance of simulating human cognition for referring regions [26]. Bremner et al. investigated the efficiency of speech and iconic gestures in multi-modal interaction with people and argued that robot communication should be multi-modal to disambiguate its meaning and improve the quality of interaction [27]. These research works enabled social robots to appropriately use gestures and deictic words (e.g., this/that) based on the position relationships among a robot, people, and the objects/areas in interactions.

A common assumption argues that robots need the capability of understanding user gestures. For instance, Dautenhahn discussed the importance of implementing perceptions of human activity [28]. Other studies addressed the idea of enabling robots to perceive user pointing gestures. For instance, Bergh et al. developed a mobile robot that understands gesture input from a user who is giving directions [29]. Droeschel used pointing gestures to provide manipulation commands to a robot [30]. Breazeal et al. developed a robot system that communicates with users by recognizing pointing gestures and concluded that a robot's nonverbal expressions help users understand and complete tasks faster [31]. These studies demonstrated some potential applications enabled by the perception capability of robots for pointing gestures. Note, however, that since these studies realized pointing-gesture recognition under a situation where a single sensor is located in front of a user, applying such recognition would be difficult in a situation where a user freely moves because she might vacate the sensing area of a single sensor or cause occlusion with her body.

On the other hand, the influence of the capability of understanding gestures on robots has been overlooked; most studies have failed to address to what extent interaction is facilitated by the robot's capability of understanding. Sugiyama et al. experimentally compared a robot with and without a deictic interaction capability (recognizing the pointing ges-

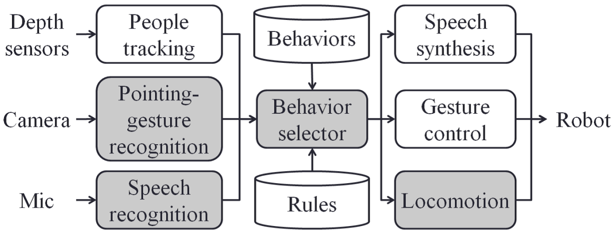

Fig. 2 System overview: gray modules controlled by operator

tures of participants and making them with its arms) and concluded that it provides better subjective impressions [32]; however, in their study participants engaged in the same task in both conditions without investigating whether such a capability helped them perform a task. In other words, even if the robot can recognize pointing gestures, it remains unknown whether that capability encourages questions from interacting partners.

Thus, little empirical evidence exists about what specifically would be enabled if robots could understand gestures. If we armed ourselves with empirical evidence, we might better comprehend whether implementing such a capability is beneficial for educational use or museum guides. From these viewpoints, our study is novel because it reveals to what extent a robot's capability of understanding pointing helps children perform their task: asking questions.

3 SYSTEM

We systematically teleoperated our system, which is designed for Wizard-of-Oz studies [33], to create an autonomous system for the near future. This approach enables us to study user interactions in ideal situations where the robot always correctly responds. At the same time, we can collect data to evaluate to what extent the system is ready for autonomous use and identify its remaining problems.

3.1 Architecture

Figure 2 illustrates the architecture of our developed, semi-autonomous system that consists of seven modules: people-tracking, pointing-gesture recognition, speech recognition, behavior selector, speech synthesis, gesture control, and locomotion. The gray modules are controlled by the operator.

3.2 Robot and Infrastructure

We used a 120-cm tall, human-like robot that has two arms (4*2 DOF) and a head (3 DOF). It is equipped with cameras, microphones, and a speaker. Its mobile base is a Pioneer 3DX. The robot was placed in a room where ten depth sensors (Microsoft Kinect) were attached to the ceiling at 2600 mm for people-tracking based on a previous algorithm [34]. Pointing-gesture recognition was performed by the operator, and we also prepared a system that recognizes pointing for the future implementation of autonomous systems (Section 6.1).

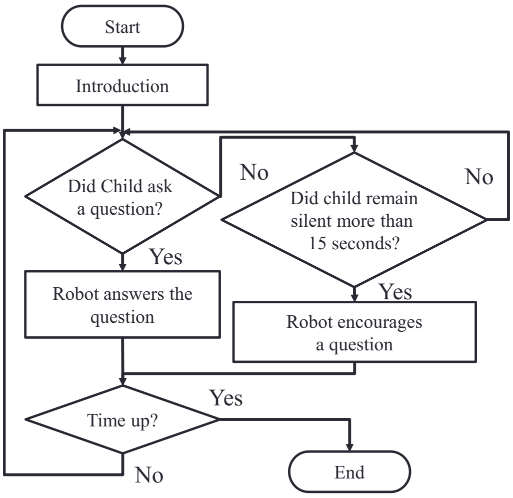

Fig. 3 Flow of robot's dialogue

3.3 Behavior Selector and Behavior Implementation

The behavior selector module selects a *behavior* (a combination of gestures and utterances) from a pre-implemented set of them by following the pre-determined rules for the dialogue flow shown in Fig. 3. The primary part is the question-answer dialogue, in which the responding behavior is selected based on speech and pointing input. When children ask relevant questions about an exhibit, the robot exhibits a behavior to provide answers. We designed each answering behavior to provide rather simple answers, including one sentence that directly answers the question, and another that offers a relevant and/or informative fact. For instance, when a person asks about the space station's purpose, it answers, "the space station conducts scientific experiments in space" and adds "people in various countries are working together on it." We consulted with elementary-school teachers beforehand to create understandable contents for children.

The robot starts its introduction when a child enters the room. If she is silent for 15 seconds, it directly encourages her to ask a question: "Do you have any questions? Don't be shy." After five minutes (the maximum time duration), the robot tells her that her time is finished: "Well, our time is up. Thanks for coming."

In all the behaviors, the robot's face is controlled using information from the people-tracking module that looks at the child's face. When the robot discusses a target (i.e., a poster), it looks at it and then faces the child to establish joint-attention. When the robot finishes speaking, it looks at the child's face again. When a gesture is defined in a behavior, the *gesture control* module manages each of the robotfs joints to perform it. Utterances are sent in text form to the *speech synthesis* module.

3.4 Operator Involvement

As shown in Fig. 2, the operator's management is related to four of the seven modules, not the whole system. For this purpose, we developed a graphical user interface that has several buttons from which to choose a behavior or a speech recognition result; the interface also sets the facing target ID and has a text box for text-to-speech. We prepared 126 rules to answer questions, i.e., by selecting an appropriate one from 77 robot behaviors to maintain consistent operator actions. For mapping questions from children to operator actions, we focused on unique features to identify such targets as name, appearance, position relationship, and so on.

First, in *locomotion* control, the operator directs the robot to face the child. Once the operator sets a facing target ID, the robot autonomously faces the child using the position information from the position-tracking system. Second, to substitute for *speech recognition*, the operator interprets the user utterances to select a robot's corresponding behavior from the *behavior selector* to be executed. Third, for *pointing-gesture recognition*, the operator monitors video images from the cameras and provides the target of a pointed-at object.

Finally, for the *behavior selector*, as mentioned above, the operator chooses from a set of pre-implemented behaviors following the pre-defined rules. After selecting the robot's behavior, it autonomously controls its motors and voices based on pre-defined sets of gestures and utterances. If there is no corresponding behavior, the operator types the utterances, follows the behavior design, and provides a fact for simple answers and one additional relevant fact. In many cases (as discussed below), the operator merely edits existing utterances, which can be done quickly. If the operator must look up the answer, such operations typically take just 20 to 30 seconds. The operator can stall using conversational fillers, like "well, let's see. . . ," which are pre-defined on the interface, similar to other behaviors.

4 EXPERIMENT

We investigated the effects of a robot's deictic interaction capability by measuring the number of questions asked by children using a robot with/without the capability of understanding pointing through a between-participant design experiment.

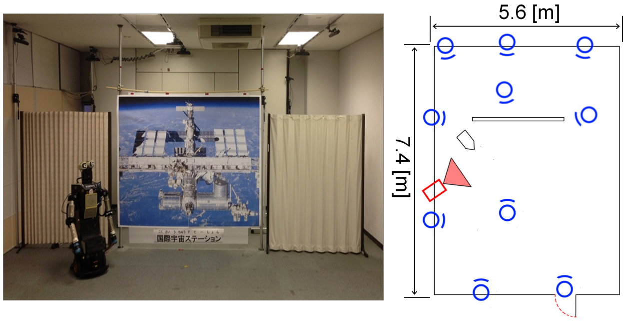

Fig. 4 Experiment room with one of two posters: blue circles indicate sensor positions and red rectangle indicates camera position.

4.1 Hypothesis

Since our exhibits are probably rather new to most of the children, we expected them to have difficulty articulating questions just with language. According to the literature, school-age children point during learning. Hence, we expect that they will communicate more easily and ask more questions about the exhibits if they are allowed to point. Based on the above discussion, we made the following hypothesis: **Hypothesis**: Children will ask a robot more questions in contexts when they can point than in contexts where they cannot.

4.2 Participants

Our participants were 92 (55 boys and 37 girls) elementary-school children: 16 1st graders, 28 2nd graders, 10 3rd graders, 12 4th graders, 16 5th graders, and 10 6th graders. We recruited from a nearby elementary school 28 volunteer students who were encouraged by the school to join as an educational activity. We complied with the school's request that the children not be paid. The parents of 64 children, who were found from mailing lists from other elementary schools, received 1000 yen for their children's participation. Both groups were equally distributed in each condition, and we did not find any differences based on the recruiting methods. We first explained the experiment information to the children's parents; if both parents and the children agreed with the participation conditions, they were invited to join.

4.3 Environment

Figure 4 shows our experiment's 7.4 by 5.6 m room that resembles a museum exhibit. It has depth sensors and a camera for recognizing the children's locations and their pointing gestures. Participants were randomly assigned to observe one of two big posters, a space shuttle or a space station. For these posters we prepared 77 robot behaviors (Section 3.3) to answer their questions. We chose these posters because they probably contain new information for elementary-school students as well as many conspicuous parts at which they can point. Since we wanted to confirm that the results are consistent across different contents, we prepared two different posters. We did not find any differences in our measurements between them.

4.4 Procedure

An assistant escorted the children to the experiment room, which they entered to learn about an exhibit. They were told to pretend that this room is in a museum and that the robot is a guide that will answer their questions about the exhibit. They were told to leave the room when they finished asking questions. Finally, we conducted a brief interview. This research was approved by our institution's ethics committee for studies involving human participants. Written, informed consent was obtained from the participants and their parents.

4.5 Conditions

The study had a between-participant design with the following two conditions. Participants were randomly assigned to the conditions while balancing the different elementary-school grades.

Deictic condition: In this condition, since the robot uses all the modules in Fig. 2 to explain the exhibits, it can perceive the pointing gestures (Table 1). Its capability was explicitly communicated to the children. Before the children entered the room, the assistant demonstrated a pointing gesture to them and explained that they can ask the robot questions by pointing and that such characteristics as an exhibit's name can be used by asking questions. The robot also introduced the following capability: "if you have any questions about the exhibit, please ask me or just point at something in the poster." The robot also added this encouragement: "If you have any questions, please ask me. You can point like this," and then it pointed at the exhibit.

Speech-only condition: In this condition, we did not use the pointing-gesture recognition module (Section **??**). Since the operator did not use any visual information for speech recognition, he could not distinguish between such reference terms as this and that. Its capability was explicitly communicated to the children (Table 1). Before they entered the room, the assistant just explained that such characteristics as an exhibit's name can be used in verbal questions. In this instruction, the assistant did not use any pointing gestures himself because the robot has no capability to recognize them. The robot mentioned that its eyes do not function: "if you have any questions about the exhibit, please ask me verbally." The robot also offered encouragement: "If you have any questions, feel free to ask me."

In both conditions, the robot used pointing gestures when it first mentioned the exhibit's name. During the question-answer dialogue, we did not implement any of them because they are typically used when interactors do not share the attention target. It seemed unnatural when the answering robot pointed at the part that was being asked about (Table 1).

For safety purposes, the operator monitored the robot/child interactions by video in both conditions, but when interpreting questions, the operator just used it in the speech-only condition.

We prepared reaction behaviors for invalid questions (described in the next subsection) based on each condition. For example, if a child asked a question just using deictic words and pointing gestures in the speech-only condition, the robot said, "Excuse me, would you ask your question verbally again?" Or if the operator could not understand a question, the robot used a different expression like "would you repeat your question?" However, if the children asked non-related questions about the exhibits, the robot reminded them to ask about the exhibits to discourage such questions.

4.6 Measurement

We evaluated the interactions based on the following criteria:

Number of questions asked: We counted the number of questions asked by each child about the exhibit.

For this measurement, we excluded questions whose meaning the operator was unable to recognize or whose target the operator failed to determine. For instance, in the speech-only condition, a child pointed and asked "what's this?," but the operator couldn't identify the target. If two parts were round, and a child asks, "what's that round part?" without pointing, the operator was unable to identify the target (Table 1). We classified such cases as invalid questions.

During the experiment, the operator judged whether a question was valid/invalid in real-time, and the robot only answered valid questions. After the experiment, two people who did not know the study design (e.g., its hypothesis and conditions) separately coded the transcribed text offline to count the number of children questions using recorded videos.

5 RESULTS

5.1 Observations

In the typical interaction pattern of both conditions, the children looked at the poster after it was introduced by the robot, asked questions, and listened to the answers while looking

Table 1 Summary of each condition

	Deictic	**Speech-only**
Robot’s pointing gesture (ex., robot explains space shuttle’s wing and points at it.)	Not used in experiment (only once in instruction)	Not used in experiment (only once in instruction)
Child’s pointing gesture (ex., pointing at a part of exhibit with/without utterance (Figs. 5 and 6))	Robot recognized	Robot did not recognize
Child’s reference term with reference movement (ex., “What’s this?” with pointing gesture)	Robot recognized	Robot did not recognize
Child’s reference term without reference movement (ex., “What’s this?” without pointing gesture)	Robot did not recognize	Robot did not recognize
Noun that can identify a target (ex., part marked by Japanese flag)	Robot recognized	Robot recognized
Noun that cannot identify a target (ex., a round part)	Robot did not recognize	Robot did not recognize

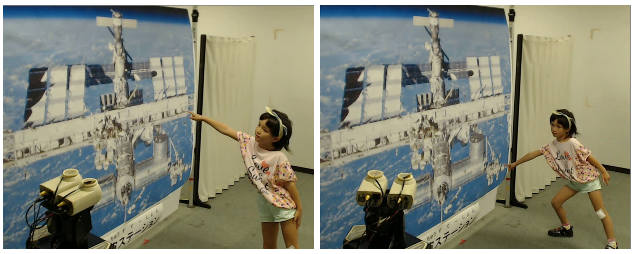

Fig. 5 Lower elementary-school girl asking questions by pointing

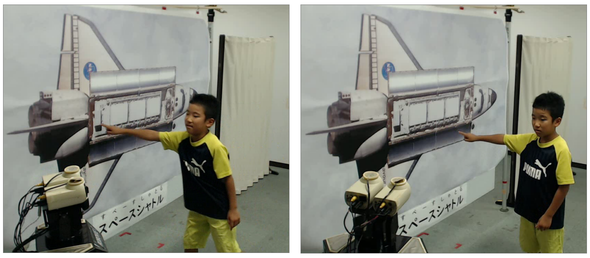

Fig. 6 Upper elementary-grade boy asking questions by pointing

at the poster. In the deictic condition, children also typically asked questions by pointing. Fig. 5 shows one typical scene with a lower elementary-school girl:

Girl: What’s this? (*pointing*)

Robot: That upper, middle part is currently an unused place. It’s the oldest part of the space station.

Girl: What’s this? (*pointing*)

Robot: That bottom, right part controls the robot’s arm, which is used to repair the space station.

She repeated such questions a couple of times and listened to the robot’s answers.

Figure 6 is another example in the deictic condition with an older boy who repeatedly asked questions and sometimes pointed or just used language:

Boy: What’s this engine? (*pointing*)

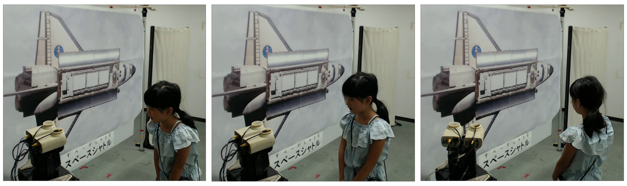

Fig. 7 Girl asking questions in speech-only condition

(a) t = 6 sec (b) t = 65 sec (c) t = 115 sec

Fig. 8 First grader who could not ask questions without pointing

Robot: That engine is near the rear of the space shuttle. It’s about 100 times faster than a bullet train.

Boy: What’s inside it? (*pointing*)

Robot: Cargoes carried by the space shuttle. They are components for the space station.

Boy: How high can it travel?

Robot: It travels between 200 to 1000 km above the earth, where it’s so high that there’s no air.

In the speech-only condition, children usually asked questions without pointing, as in Fig. 7. Here the girl asked a couple of questions:

Girl: What does the space shuttle do?

Robot: It conducts experiments and takes cargo into space.

Girl: What kind of people ride in it?

Robot: Astronauts. Japanese astronauts have often ridden in this space shuttle.

However, the children in lower elementary-school grades sometimes had difficulty asking questions. Fig. 8 shows a girl who did not ask any questions for almost two minutes.

She entered the room and listened to the robot's introduction (Fig. 8a). She stood in the same place until the end and stared at the poster (Fig. 8b). When the robot encouraged her to ask a question, she just silently stared at it. About two minutes later she left the room without speaking (Fig. 8c). We counted the number of children who did not ask any questions. 2 of 46 children (one 1st grader and one 2nd grader) in the deictic condition, and 11 of 46 (six 1st graders and five 2nd graders) in the speech-only condition did not ask any valid or invalid questions.

Children sometimes asked invalid questions. For example, in the speech-only condition, they asked questions by pointing, even though they were told to just use language. In such cases, the robot admitted that its eyes were so bad that it couldn't see the target and requested the child to repeat the question with language. Children typically obeyed. Two children in the deictic condition and 17 in the speech-only condition asked invalid questions (in our experiment, invalid questions were asked 2 and 32 times, respectively), The majority followed that request, except for two children in the speech-only condition who left the room without asking a single question after that.

We also observed cases where the children asked questions after the robot offered encouragement. For instance, one child did not speak in the beginning, but six seconds after the robot encouraged him he blurted out, "How does its jet (engine) work?" and then asked two more questions. 12 children asked their first question after being encouraged (three in the deictic and nine in the speech-only condition).

Three children asked overly easy questions, such as "What is the color of the earth?" (Since the earth is shown on the poster, its color is clearly visible) and "What's this picture?" (Its name was written at the bottom of the poster and had already been mentioned by the robot). They seemed to be testing/teasing the robot. In contrast, seven children seemed to have a fair amount of knowledge about the exhibits and asked difficult questions that sought deeper understanding: "Where in the space station does the *Kounotori* satellite dock?" and "How does the space station get its electricity?"

A majority of the children seemed to enjoy interacting with the robot. For instance, after the experiment one girl said, "That was the first time I ever talked with a robot, and even though I didn't ask many questions, it was fun." Many children in the older grades also reported they enjoyed both conditions. However, some in the younger grades said that they did not enjoy it because they were unable to ask very many questions.

5.2 Verification of Hypothesis

Next we analyzed the numbers of questions that were asked. We performed a Kolmogorov-Smirnov test to check the normality of the number asked in both conditions; they were not normally distributed. Therefore, we conducted a Mann-Whitney U test instead of an ANOVA and learned that the number of questions asked (Fig. 9) was greater for the deictic condition (median=4) than for the speech-only condition (median=2) ($U = 601.000$, $p < .001$, $r = .038$). Thus, our hypothesis was supported.

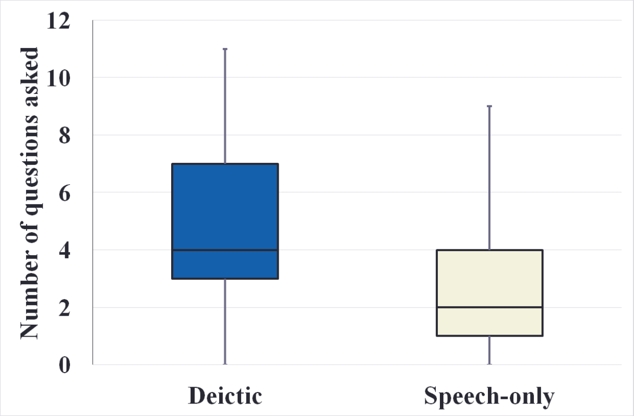

Fig. 9 Number of questions asked in each condition

Fig. 10 Ratio of questions with pointing gestures in deictic condition

If we also include the number of invalid questions, the analysis still indicates that the number of valid and invalid questions is greater for the deictic condition (median=4) than for the speech-only condition (median=3) ($U = 736.500$, $p = .011$, $r = .026$).

5.3 Additional Analysis

Since we expected that the younger children would have difficulty asking questions and would benefit from pointing, we investigated how the data were distributed in relation to the ages/grades of the children by determining the ratio of questions asked with pointing gestures in the deictic condition (Fig. 10). The ratio seemed to decrease as their grade level increased. For instance, 1st graders always asked questions

Table 2 Classification of types of questions

	Factual		Conceptual	
	with pointing	w/o pointing	with pointing	w/o pointing
Deictic	163	35	11	8
Speech-only	13	86	0	16

Table 3 Classification of conceptual questions

	Immediate	After factual
Deictic	6	13
Speech-only	5	11

by pointing, but 6th graders just asked 53.3% of their questions by pointing.

We also analyzed the details of the age effects in each condition. For each condition, we ran a Kruskal-Wallis H test to compare the grades of the children and identified significant differences in the number of questions asked within the speech-only condition ($H = 17.261$, $p = .004$). Children in 1st and 2nd grades asked fewer questions than 3rd, 4th, and 5th graders. No significance was found in the deictic condition ($H = 5.563$, $p = .351$). If we compare conditions within each grade, only the Mann-Whitney U test showed significance for 2nd graders ($U = 37.500$, $p = .010$), who in the speech-only condition asked fewer questions than in the deictic condition.

One might expect children in the deictic condition to ask easier questions (e.g., “what’s this?”) than children in the speech-only condition. We addressed this potential concern by scrutinizing the contents of their valid questions and classifying them based on the knowledge that was provided as answers. Different dimensions of knowledge and cognitive processes exist [35]. Within the range of our study, the children started to *remember factual* knowledge (about specific details and elements), which requires low order cognitive skills, and they later learned to *understand conceptual* knowledge (e.g., principles, theories, and models), which requires rather advanced cognitive skills. In our cases, most of their questions sought *factual* knowledge (“what is this?”), and some were seeking such *conceptual* knowledge as “Why is the shuttle’s nose so sharp?” and “How does the space shuttle move in space?” Thus, we categorized the children’s questions into two types: *factual* and *conceptual*. Two independent coders, who did not know our research hypothesis, classified them. Their judgment matched reasonably well and yielded a Cohen’s kappa coefficient of 0.842. We further analyzed whether the questions were accompanied by pointing.

Table 2 shows the analysis result. We found no difference in the total number of *conceptual* questions between the conditions. This suggests that the robot’s deictic interaction capability encouraged more questions about *factual* knowledge without discouraging them from acquiring *conceptual* knowledge. As expected, note that questions with pointing were mostly found in the deictic condition. Some valid questions in the speech-only condition were found by pointing, although the meaning was successfully conveyed without acknowledging the pointing.

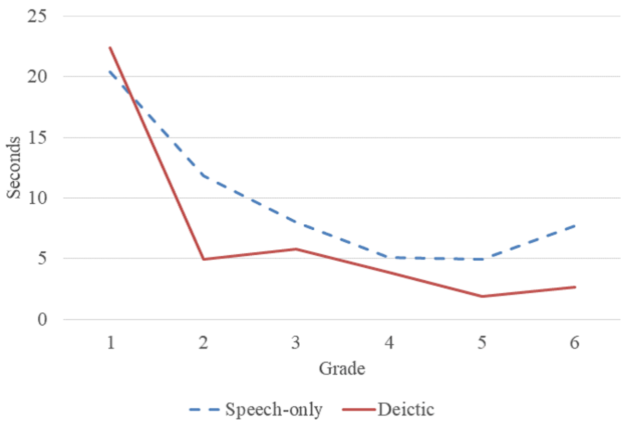

Fig. 11 Times until first questions

We also analyzed whether their utterances were related to questions or just chatting. Two independent coders, who did not know our research hypothesis, classified all 459 utterances (268 in the deictic condition and 191 in the speech-only condition). The Cohen’s kappa coefficient from their classifications was 0.923, indicating strong agreement between them. As a result, the utterances were mainly questions (219/268 = 81.7% in the deictic condition and 147/191=77.0% in the speech-only condition) or backchannels, e.g., “I see” and “ok” (39/268 = 14.6% and 36/191=18.9%), and there were very few other utterances, such as “hello” or “bye” (10/268 =3.7% and 8/191=4.2%). Many children apparently regarded the robot as more of an adult than a peer, since they spoke to it politely like with adults. We did not find evidence that the children were more/less talkative in either condition.

Moreover, we measured the number of follow-up (conceptual) questions to investigate whether the pointing gestures encouraged more follow-up questions than vagueness without the pointing gestures (Table 3). Within 35 conceptual questions, 68.6% ((13+11)/35) were asked just after the factual questions. The remaining 31.4% ((6+5)/35) were not related to the immediate-precedent questions or the first questions from the children.

The numbers of conceptual questions, which were related to whether the precedent questions were with/without pointing gestures (see right side of Table 2), were 31.4% ((11+0)/35) and 68.6% ((8+16)/35). We conducted a chi-squared test but found no significant differences between them.

5.4 Time Until First Question From Children

We analyzed the time between the end of the robot's instruction and a child's first question to investigate the ease with which the children asked questions. The median time until the first valid question was 4.0 seconds in the deictic condition and 19.0 seconds in the speech-only condition. A Mann-Whitney U test showed a significance ($U = 350.000$, $p < .001$). Including both valid and invalid questions, it was 4.0 seconds in the deictic condition and 6.4 seconds in the speech-only condition, indicating a significant trend ($U = 611.000$, $p = .054$).

We also investigated the relationships between the children's ages and the time until the first question. Fig. 11 shows the median time until the first question including both valid and invalid questions in both conditions. Except for the 1st graders, the median time until the first question was lower in the deictic condition than in the speech-only condition. If we just focused on the valid questions, the time of the 1st graders is also lower in the deictic condition than in the speech-only condition.

5.5 Response Times of Robot

One might expect that the difference of the operator's response times between the two conditions caused the difference in the children's intention to ask questions. We analyzed the response times from the end of the children's questions until the start of the robot's answers to investigate the effect on their intention to ask questions. The operator's median response times were 8.6 and 8.1 seconds for the deictic and speech-only conditions. We conducted a Mann-Whitney U test, which was not significantly different ($U = 11746.000$, $p = .576$).

5.6 Instruction Effects

The brief instructions that we gave to the lower elementary-school children might have created an unintended bias and suppressed their motivation to ask questions (Section 4.5). To investigate the instruction effects, we analyzed what children said about the robot during their interviews. Two independent coders, who were unaware of our research hypothesis, classified all 91 transcribed interview results into the following five categories: 46 in the deictic condition and 45 in the speech-only condition; 1 dataset was lost due to a recording failure. The Cohen's kappa coefficient from their classification was 0.846, indicating moderate agreement between them.

1. Praise for robot's explanation capability (e.g., "The robot knows many things.")
2. Criticism of robot's explanation capability (e.g., "It didn't explain very much.")
3. Praise for robot interaction (e.g., "I've never been with such a robot before.")
4. Criticism of robot interaction (e.g., "I was a little nervous because this was my first time to talk with a robot.")
5. No comment

Table 4 Classification of types of interview results

	About explanation		About interaction		No comment
	Praise	Criticism	Praise	Criticism	
Deictic	23 (50%)	7 (15%)	9 (20%)	3 (7%)	4 (8%)
Speech-only	19 (42%)	7 (16%)	10 (22%)	4 (9%)	5 (11%)

Table 4 shows the analysis results. Many either praised its explanation capability (23/46=50% in the deictic condition and 19/45=42% in the speech-only condition) or criticized it (7/46=15% and 7/45=16%). Some commented on the interaction experience: 20% (9/46) and 22% (10/45) praised it and 7% (3/46) and 9% (4/45) criticized it. 8% (4/46) and 11% (5/45) did not comment. We conducted a chi-squared test, which did not show any significant difference ($x^2(4) = .677$, $V = .086$, $p = .954$).

6 TOWARD AUTONOMOUS SYSTEMS

We investigated whether the robot's capability to perceive pointing gestures encouraged the children to ask questions. Since our experiment was conducted with a Wizard-of-Oz approach, in this section we next discuss the possibility of implementing autonomous robots based on our findings. First we discuss the recognition of pointing gestures using the gathered data and then the speech recognition and content preparations.

6.1 Can We Recognize Pointing Gestures?

6.1.1 System that recognizes children's pointing

Although a number of studies have addressed gesture recognition, they typically assume that one sensor is located in front of many users. For example, Schauerte et al. developed a saliency-based model that identifies and recognizes the objects at which people are pointing [36]. They also extended their model to guide visual attention in human-robot interaction by integrating the recognitions of pointing gestures and spoken references [37]. Nagi et al. developed a vision-based system to recognize the spatial gestures of human operators to interact with multi-mobile robots by integrating

Fig. 12 Pointing-gesture recognition from depth images

Fig. 13 Boy drawing a circle while pointing at target

color-based segmentation and optical flow from camera images [38]. Cosgun et al. developed a pointing gesture of an interaction partner with a depth sensor and modeled the uncertainty of people's pointing gestures in spherical coordinates [39].

However, when a user moves freely, as in typical situations where users are browsing around an exhibit, sometimes an arm is occluded by a body in the view of a single sensor. To cope with this problem, we developed a system that recognizes pointing gestures by multiple depth sensors and designed it to be computationally light for real-time processing. Due to space limitations, we briefly explain our purpose and discuss what we learned from our Wizard-of-Oz approach.

Our algorithm extends a previous people-tracking algorithm [34]. For each depth image from each sensor, it runs a clustering algorithm for the point clouds so that each individual forms one cluster. Then from the top, it creates layers to analyze the location of the head and the shoulders. Next it seeks the location of the hands, chooses the farthest point from the shoulders, creates a cluster with the nearby points (Fig. 12), and segments the neighbor points until it reaches one of the shoulders. If it reaches either of them, it checks the shape of the segments to distinguish an arm from other body parts like legs. If it finds an arm, it computes the pointing direction as a vector that connects the shoulder and the hand. Finally, it combines the information from all of the sensors and assumes that in a pointing gesture, the arm will be extended and remain still for 500 milliseconds.

6.1.2 Result

We investigated our system's performance with the collected data. Our system detected 176 pointing gestures, among which

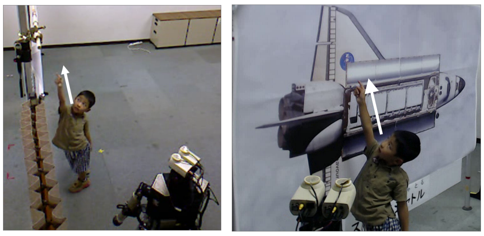

Fig. 14 Boy pointing at ceiling near poster

116 were correct (116/176=65.9%), two outputs were false positives, and 58 were wrongly identified as pointed targets. We further analyzed the incorrect cases and identified three types of failures: *deviation of pointing style*, *deviation of pointing direction*, and *system settings*. For the pointing style, although our algorithm assumed that a child would extend his arm without moving it while he is pointing, in 19 cases the children's pointing deviated from that assumption. They often drew a circle (Fig. 13) or a line while pointing. Based on this observation, we identified the need to perceive such arm motion patterns while pointing.

In the pointing direction, we identified four cases where the pointing vector was inadequately aimed at the target. For instance, when a boy pointed at the poster's top while standing too close to it, his gesture seemed to indicate the ceiling (Fig. 14). In this case, our recognition system needs to be tolerant of such inaccurate pointing. The other 35 incorrect cases were in part due to the location of sensors (e.g., occlusion near the poster) and the program's parameters (e.g., duration of pointing gestures is shorter than the pre-defined threshold) whose adjustment can be improved using the collected data.

Overall, due to the Wizard-of-Oz approach, we collected the naturalistic behaviors of children and learned how to improve our recognition system. In addition, we believe that we can probably apply a learning technique to automatically learn from such collected data.

6.2 Can We Recognize Children's Utterances?

Speech recognition remains difficult. One study reported that it is only 21.3% successful in the real world, although it was designed to be robust for noise [40]. In addition, a child's voice is more difficult to recognize than an adult's due to differences of vowel formant frequencies and the speech's bandwidth [41]. For such problems, if we use the pointing-gesture recognition algorithm from Section 6.1, 53.5% (116/217, since 116 pointing gestures were correctly recognized) of the questions in the deictic condition were correctly recognized even without any speech recognition. Like Siri and Google speech recognition, training with a large amount of data will improve performance. Gathering speech data in a realistic situation is another possible contribution.

6.3 Can We Prepare Robot's Utterances in Advance?

Since children's interests are diverse, one might prepare necessary robot responses in advance. In our study, the children in the deictic condition asked 217 questions, 25 of which were not covered by the prepared utterances (88.5% were covered). The operator typed responses for them. Children in the speech-only condition asked 115 questions, 37 of which were not covered (67.8% were covered). The covered ratio was higher in the deictic condition because children often pointed at parts of the exhibits, which we easily anticipated.

The main reason that some situations were not covered is that sometimes children asked rather easy questions but differently than we expected (7 in the deictic condition, 18 in the speech-only condition). For instance, although we had prepared explanations about the space station's solar cells, some children asked, "Does it have any solar cells?" We did not expect such basic questions. Perhaps the children asked questions whose answers they already knew. The children also asked about topics the robot had just mentioned in a previous response (12 cases in the deictic condition, 9 in the speech-only condition). For instance, after the robot answered a question by saying, "The module at the space station's lower right is a laboratory built in Japan. It's the biggest laboratory in the space station," a child asked, "Where's the second largest laboratory?" In the remaining cases (6 in the deictic condition, 10 in the speech-only condition), children asked difficult, unexpected questions: "Except for humans, who else has ridden on the space shuttle?" and "How does the space shuttle land?" Even though we prepared utterances to cover various situations (e.g., we conducted preliminary trials with 20 children), it was not easy to cover the diversity of their questions. We believe that the Wizard-of-Oz approach is very useful for such initial use to increase the robot's preparation.

7 DISCUSSION

7.1 Design Implications

First, our study revealed the importance of the robots' capability to perceive user pointing. Children asked a robot more questions with such a capability than a robot without it. This seems particularly critical for young children since the 1st graders always asked questions by pointing in the deictic condition. Younger children might be interested in the individual items depicted in the posters, while older children might be more interested in the poster as a single entity. Also, their linguistic abilities to explain their questions without pointing as well as their basic knowledge are quite different based on their grade in school. These discrepancies result in a different number of pointing gestures while they ask questions. If we design such a guide robot for children, in particular, younger children, deictic interaction capability is critical. Concerning the contents, some observations influenced and led to design implications.

In this study we only focused on an educational situation with children, but deictic interaction capability is useful in a situation where people do not know the actual words and/or pronunciations. For example, even if people get lost and cannot ask for directions from a guide robot in a large shopping mall or a tourist resort, they could use pointing gestures and a map to get appropriate answers if the robot has a deictic interaction capability.

Sometimes the children tested the robot and asked simplistic questions to which they seemed to already know the answers. On the other hand, because some children appeared very familiar with the exhibited target, they asked rather advanced questions. These observations suggest the need to prepare answers for a wide range of questions, even when the robot is used with children. In addition, we observed a pattern where children remained silent and only started to speak after being encouraged by the robot. We believe that such robot behavior will benefit children.

7.2 Limitations

Since our study was conducted in a specific country with a particular robot and users, its generalizability is limited. We did not investigate any long-term influence, e.g., whether the children became more interested in exhibits, understood them better, or wanted to study more, all of which are important for use as a guide. Moreover, it is difficult to control children's pre-knowledge and interest in the contents of the posters and the robot. In addition, the degree of understanding about explanations from the robot in the experiment would vary due to differences of their ages. Such differences would create variability between children and influence question-asking behaviors about the robot in the experiments.

In this study, we told the children that the robot's eyes do not function in the speech-only condition, because in early preliminary trials we found that it was too difficult for children to ask valid questions if they received no instructions; they seemed unable to understand why the robot could not recognize their pointing gestures. Thus, we expect that without such instruction, children would perhaps ask more invalid questions or ask them quicker (the time until the first questions would decrease) without necessarily expecting more valid questions. Children may or may not understand the robot's instructions requesting them to use only verbal information when asking questions. Children who understand what kinds of information they can use to ask questions would probably be less influenced, and those who fail to understand would continue to point or get confused and stop asking questions.

Some children in the speech-only condition asked questions by pointing, although we instructed them to avoid gestures. This was possible because of the robot's embodiment, including its human-like appearance and the fact that it used pointing gestures. Because we did not have a condition without an embodied robot, it is an open question how such embodiment helps or discourages children from asking questions. Even though some did not ask any questions, we did not determine the reason. Perhaps they did not have enough knowledge to verbalize about the exhibit as hypothesized, or perhaps they were too shy. Pragmatically, it is useful to know whether a robot's capability to perceive pointing helps children ask questions; room remains for further investigation.

We did not conduct any tests to identify domain knowledge about the exhibit before the experiment. We expected that most of the children would not have domain knowledge because most of the questions were factual (e.g., ”what is this?” by pointing at the posters), but we had no evidence about domain knowledge beforehand. Therefore, any detailed effects of the children's domain knowledge about asking questions remain unknown.

Our experiment was conducted with a poster, although in the real world exhibits are often three-dimensional. We do not expect a substantial difference (e.g., frequency of questions) to be caused by such differences in a setting, although a robot should probably more actively move around the exhibit in such cases. Perhaps the required interaction will be more complex.

8 CONCLUSION

We investigated whether a robot's capability to understand pointing encourages children to ask questions. We conducted an experiment with 92 elementary-school children and compared the existence of such a capability in a robot. The children asked significantly more questions of a robot with a deictic interaction capability than one without it. We also identified the relationship between their school grades/ages and the number of asked questions.

Since our study used a Wizard-of-Oz approach, we also discussed the possibility of implementing autonomous robots based on our finding. For this purpose, we developed a system to recognize the pointing gestures of children and evaluated the system's performance. We also discussed speech recognition and content preparations toward an autonomous system.

9 ACKNOWLEDGMENTS

We thank the assistants for their support during our experiments. This work was in part supported by JSPS Grants-in-aid for Scientific Research Numbers 25240042 and 25280095 and JSPS KAKENHI Grant Numbers JP15H05322 and JP16K12505.

10 Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. S. Thrun et al., “MINERVA: a second-generation museum tour-guide robot,” *IEEE Int. Conf. on Robotics and Automation (ICRA1999)*, pp. 1999-2005, 1999.
2. M. Ghosh and H. Kuzuoka, “An ethnomethodological study of a museum guide robot's attempt at engagement and disengagement,” *Journal of Robotics*, 2014.
3. M. Saerbeck, T. Schut, C. Bartneck, and M. D. Janse, “Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor,” *ACM Conf. on Human Factors in Computing Systems (CHI2010)*, pp. 1613-1622, 2010.
4. J. Han, M. Jo, S. Park, and S. Kim, “The educational use of home robots for children,” *IEEE Int. Workshop on Robot and Human Interactive Communication (RO-MAN2005)*, pp. 378-383, 2005.
5. B. L. McCombs and J. S. Whisler, “The learner-centered classroom and school: strategies for increasing student motivation and achievement,” The Jossey-Bass Education Series., Jossey-Bass Inc., Publishers, 350 Sansome St., San Francisco, CA 94104, 1997.
6. I. Howley, T. Kanda, K. Hayashi, and C. Rosé, “Effects of social presence and social role on help-seeking and learning,” *ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI2014)*, pp. 415-422, 2014.
7. Goldin-Meadow, Susan, “Hearing gesture: How our hands help us think,” *Harvard University Press*, 2005.
8. McNeill, D, “Hand and mind: What gestures reveal about thought,” *University of Chicago Press*, 1992.
9. Dittmann, Allen T. and Llewellyn, Lynn G, “Body movement and speech rhythm in social conversation,” *Journal of Personality and Social Psychology*, Vol. 11, Issues 2, pp. 98-106, 1969
10. Gordon W. Hewes, “Primate Communication and the Gestural Origin of Language,” *Current Anthropology*, Vol. 32, pp. 65-84, 1992.
11. F. H. Rauscher, R. M. Krauss, and Y. Chen, “Gesture, Speech, and Lexical Access: The Role of Lexical Movements in Speech Production,” *Psychological Science*, vol. 7, no. 4, pp. 226-231, 1996.
12. Alibali, Martha Wagner, and Alyssa A. DiRusso, “The function of gesture in learning to count: More than keeping track,” *Cognitive development*, Vol. 14, Issue 1, pp. 37-56, 1999.
13. Goldin-Meadow, Susan, Alibali, Martha W, and Church, R Breckinridge, “Transitions in concept acquisition: Using the hand to read the mind,” *Psychological review*, Vol. 100, Issues 2, pp. 279-297, 1993.
14. K. J. Pine, H. Bird, and E. Kirk, “The effects of prohibiting gestures on children's lexical retrieval ability,” *Developmental Science*, vol. 10, no. 6, pp. 747-754, 2007.
15. M. G. Sauter, D. H. Uttal, A. Schaal, S. C. Levine, and S. Goldin-Meadow, “Learning what children know about space from looking at their hands: The added value of gesture in spatial communication,” *Journal of Experimental Child Psychology*, pp. 587-606, vol. 111, no. 4, 2012 (Epub 2011).
16. B. J. Matlen, K. Atit, T. Göksun, M. A. Rau, and M. Ptouchkina, “Representing space: Exploring the relationship between gesturing and geoscience understanding in children,” *Int. Conf. on Spatial Cognition VIII*, pp. 405-415, 2012.

17. R. M. Krauss, “Why do we gesture when we speak?,” *Current Directions in Psychological Science*, pp. 54-60, 1998.
18. M. Alibali, “Gesture in spatial cognition: expressing, communicating, and thinking about spatial information,” *Spatial Cognition and Computation*, pp. 307-331, 2005.
19. H. Kuzuoka, S. Oyama, K. Yamazaki, K. Suzuki, and M. Mitsuishi, “Gestureman: a mobile robot that embodies a remote instructor’s actions,” *ACM Conf. on Computer-supported Cooperative Work (CSCW2000)*, pp. 155-162, 2000.
20. B. Scassellati, “Investigating models of social development using a humanoid robot,” *Biorobotics*, B. Webb and T. Consi eds., MIT Press, 2000.
21. Y. Okuno, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, “Providing route directions: design of robot’s utterance, gesture, and timing,” *ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI2009)*, pp. 53-60, 2009.
22. M. Lohse, R. Rothuis, J. Gallego-Pérez, D. E. Karreman, and V. Evers, “Robot gestures make difficult tasks easier: the impact of gestures on perceived workload and task performance,” *ACM Conf. on Human Factors in Computing Systems*, pp. 1459-1466, 2014.
23. A. Sauppé and B. Mutlu, “Robot deictics: how gesture and context shape referential communication,” *ACM/IEEE Int. Conf. on Human-robot interaction (HRI2014)*, pp. 342-349, 2014.
24. Ng-Thow-Hing, Victor, Pengcheng Luo, and Sandra Okita, “Synchronized gesture and speech production for humanoid robots.” *Intelligent IEEE/RSJ International Conference on Robots and Systems (IROS 2010)* pp. 4617-4624, 2010.
25. Huang Chien-Ming and Bilge Mutlu., “Modeling and Evaluating Narrative Gestures for Humanlike Robots.” Robotics: Science and Systems,” *Robotics: Science and Systems*, 2013.
26. Yasuhiko Hato, Satoru Satake, Takayuki Kanda, Michita Imai, and Norihiro Hagita, “Pointing to space: modeling of deictic interaction referring to regions,” *ACM/IEEE Int. Conf. on Human-robot interaction (HRI2010)*, pp. 301-308, 2010.
27. Paul Bremner and Ute Leonards, “Iconic gestures for robot avatars, recognition and integration with speech,” *Frontiers in Psychology*, Vol. 7, pp. 1-14, 2016.
28. K. Dautenhahn, “Methodology and themes of human-robot interaction: a growing research field,” *Int. Journal of Advanced Robotic Systems*, vol. 4, no. 1, pp. 103-108, 2007.
29. M. Van den Bergh et al., “Real-time 3D hand gesture interaction with a robot for understanding directions from humans,” *IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN2011)*, pp. 357-362, 2011.
30. D. Droeschel, J. Stückler, D. Holz, and S. Behnke, “Towards joint attention for a domestic service robot-person awareness and gesture recognition using time-of-flight cameras,” *IEEE Int. Conf. on Robotics and Automation (ICRA2011)*, pp. 1205-1210, 2011.
31. 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.
32. O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, “Natural deictic communication with humanoid robots,” *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2007)*, pp. 1441-1448, 2007.
33. N. Dahlbäck, A. Jönsson, and L. Ahrenberg, “Wizard of Oz studies: why and how,” *Int. Conf. on Intelligent User Interfaces (IUI1993)*, pp. 193-200, 1993.
34. D. Brscic, T. Kanda, T. Ikeda, and T. Miyashita, “Person tracking in large public spaces using 3d range sensors,” *IEEE Trans. on Human-Machine Systems*, pp. 522-534, 2013.
35. L. W. Anderson, D. R. Krathwohl, and B. S. Bloom, “A taxonomy for learning, teaching, and assessing: a revision of Bloom’s taxonomy of educational objectives,” Allyn & Bacon, 2001.
36. Schauerte, Boris, Jan Richarz, and Gernot A. Fink., “A Saliency-based identification and recognition of pointed-at objects,” *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2010)*, pp. 4638-4643, 2010.
37. Schauerte, Boris and Rainer Stiefelhagen, gLook at this! learning to guide visual saliency in human-robot interaction,h *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2014)*, pp. 995-1002, 2014.
38. Nagi, Jawad, Giusti, Alessandro, Gambardella, Luca M, and Di Caro, Gianni A, “Human-swarm interaction using spatial gestures,” *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2014)*, pp. 3834-3841, 2014.
39. Cosgun, Akansel, Alexander JB Trevor, and Henrik I. Christensen, “Did you Mean this Object?: Detecting Ambiguity in Pointing Gesture Targets,” *Int. Workshop. on Towards a Framework for Joint Action Workshop in HRI2015*, 2015.
40. M. Shiomi, D. Sakamoto, T. Kanda, C. T. Ishi, H. Ishiguro, and N. Hagita, “Field trial of a networked robot at a train station,” *Int. Journal of Social Robotics*, pp. 27-40, 2011.
41. Q. Li and M. J. Russell, “Why is automatic recognition of children’s speech difficult?” *European Conf. on Speech Communication and Technology*, pp. 2671-2674, 2001.

Tsuyoshi Komatsubara received M.Eng. degree from Osaka University, Osaka, Japan, in 2016. He was an internship student at the Intelligent Robotics and Communication Laboratories (IRC), Advanced Telecommunications Research Institute International (ATR). His research interests included human-robot interaction and educational robots.

Masahiro Shiomi received Ph.D. degrees in engineering from Osaka University, Osaka, Japan, in 2007. He is currently a Group Leader with the ATR Intelligent Robotics and Communication Laboratories. His research interests include human-robot and human-agent interaction.

Takayuki Kanda received the B.Eng., M. Eng., and Ph.D. degrees in computer science from Kyoto University, Kyoto, Japan, in 1998, 2000, and 2003, respectively. He is currently a Group Leader with the ATR Intelligent Robotics and Communication Laboratories. His research interests include intelligent robotics and, human-robot interaction.

Hiroshi Ishiguro received a D. Eng. in systems engineering from the Osaka University, Japan in 1991. He is currently Professor of Department of Systems Innovation in the Graduate School of Engineering Science at Osaka University (2009-) and Distinguished Professor of Osaka University (2017-). He is also visiting Director (2014-) (group leader: 2002-2013) of Hiroshi Ishiguro Laboratories at the Advanced Telecommunications Research Institute and an ATR fellow. His research interests include sensor networks, interactive robotics, and android science. He received the Osaka Cultural Award in 2011. In 2015, he received the Prize for Science and Technology (Research Category) by the Minister of Education, Culture, Sports, Science and Technology (MEXT).