IEICE TRANS. INF. & SYST., VOL.E89–D, NO.6 JUNE 2006

INVITED PAPER Special Section on Human Communication II

Toward Robots as Embodied Knowledge Media

Toyoaki NISHIDA[†], *Member*, **Kazunori TERADA**[††], **Takashi TAJIMA**[†††*], **Makoto HATAKEYAMA**[†††**], **Yoshiyasu OGASAWARA**[†††***], *Nonmembers*, **Yasuyuki SUMI**[†], *Member*, **Yong XU**[†], **Yasser F. O. MOHAMMAD**[†], **Kateryna TARASENKO**[†], **Taku OHYA**[†], *and* **Tatsuya HIRAMATSU**[†], *Nonmembers*

SUMMARY We describe attempts to have robots behave as embodied knowledge media that will permit knowledge to be communicated through embodied interactions in the real world. The key issue here is to give robots the ability to associate interactions with information content while interacting with a communication partner. Toward this end, we present two contributions in this paper. The first concerns the formation and maintenance of joint intention, which is needed to sustain the communication of knowledge between humans and robots. We describe an architecture consisting of multiple layers that enables interaction with people at different speeds. We propose the use of an affordance-based method for fast interactions. For medium-speed interactions, we propose basing control on an entrainment mechanism. For slow interactions, we propose employing defeasible interaction patterns based on probabilistic reasoning. The second contribution is concerned with the design and implementation of a robot that can listen to a human instructor to elicit knowledge, and present the content of this knowledge to a person who needs it in an appropriate situation. In addition, we discuss future research agenda toward achieving robots serving as embodied knowledge media, and fit the robots-as-embodied-knowledge-media view in a larger perspective of Conversational Informatics.

key words: *knowledge media, communicative artifacts, nonverbal communication, human-agent communication, intention, conversational informatics*

1. Introduction

Knowledge in the 21st century keeps evolving faster than ever. As a result, it is much harder for individuals to receive its full benefits. This hinders the penetration of knowledge for avoiding life's dangers and pitfalls which should serve as the basis of leading a peaceful existence, threatening the safety of society. We certainly need an effective means of helping people create and share knowledge.

The long term goal of this research was to develop robots that could serve as embodied knowledge media to help people communicate with one another. We focused on conversation because it is a natural means for people to communicate, and in fact knowledge is most effectively created, extended, consumed, and criticized in actual conversations. In this paper, we address issues with building a robot that can participate in conversations on the spot in the real world and can communicate knowledge through embodied interactions.

Conversation is a sophisticated intellectual process where meaning is associated with complex and dynamic interactions resulting from collaboration between the speaker and listener. Building artifacts that can be involved in the conversation process and produce useful interactions at the knowledge level is quite challenging.

Communication behaviors can generally be classified into those that are verbal and nonverbal. Roughly speaking, verbal communication means slow and logical communication. In contrast, nonverbal communication means fast and sophisticated coordination of conversational behaviors, playing an important role in forming and maintaining joint intentions in real time.

Nonverbal communication is considered to underlie and therefore precede verbal communication in forming and maintaining intentions, which is considered to be a fundamental capability of a real robot. Forming and maintaining intentions at the nonverbal communication level is used to interactively determine intentions at varying levels. For example, a listener may look away from an object in focus and look at the speaker's face when s/he does not follow the explanation [1]. This demonstrates how eye contact is used in ordinary conversations to signal the speaker to repeat the utterance so that the listener may be able to understand it. The process normally takes a short time and is carried out almost unconsciously in daily situations.

Nonverbal interaction makes up a significant portion of human-human interactions [2]. McNeill suggests that verbal and nonverbal expressions occur in parallel for some psychological entities called growth points [3].

Robots that can merely exchange verbal information with humans might fail to participate in human conversations in the real world, for people extensively use nonverbal information to coordinate their conversational behaviors. Robots should be able to detect various signs of nonverbal communications that participants produce, capture the meaning associated with interactions, and coordinate their behavior during the discourse. In other words, robots should be able to play the role of active and sensible participants in the conversation, rather than standing still listening to the speaker, or continuing to speak without regarding the listener.

Manuscript received October 25, 2005.

[†]The authors are with Kyoto University, Kyoto-shi, 606–8501 Japan.

[††]The author is with Gifu University, Gifu-shi, 501–1193 Japan.

[†††]The authors are with The University of Tokyo, Tokyo, 113–8656 Japan.

[*]Presently, with Matsushita Electric Industrial Co., Ltd.

[**]Presently, with NEC Corporation.

[***]Presently, with Sharp Co., Ltd.

DOI: 10.1093/ietisy/e89–d.6.1768

We took an ecological approach to overcome these difficulties. We attempted to give robots the ability of behaving according to the surface discourse of the conversation to capture or present information content, rather than exchanging meaning based on deep understanding. In other words, it appears feasible to aim at building robots that can mimic conversational behavior at least on the surface and act quickly to meet temporal requirements in nonverbal communication. For example, our robots will move eye gaze on the object when the partner has been recognized as paying attention to that object, successfully creating joint attention. The media equation theory [4] suggests that superficial similarities might allow people to coordinate behavior. In addition, it is also reasonable to expect that a robot will be able to infer that an object has a respective role in the conversation, which will enable it to add a proper discourse label to the record of the object.

2. Robots as Embodied Knowledge Media

This section overviews ecological approaches to implementing robots as embodied knowledge media. The key issue here is how to give robots the ability of associating interactions with information content while they are interacting with a communication partner. Conversation quantization gives the basis for associating interaction-oriented and content-oriented views with conversation. We point out that the ability of forming and maintaining joint intentions constitutes the foundation for making robots act as a communicative artifact. We introduce listener and presenter robots as prototypes of the idea of robots acting as embodied knowledge media.

2.1 Conversation Quantization as Implementation of Ecological Approach

Conversational quantization introduces conversation quanta to describe quantized segments of conversation. Each conversation quantum integrates conversation-as-interaction and conversation-as-content views.

The conversation-as-interaction view sheds light on how each conversation partner coordinates her/his behaviors to communicate with the other. Joint attention is a typical example. When the speaker looks at some object and starts talking about it, the listener should also look at it to demonstrate that the listener is paying attention to the explanation. When the listener loses the trail of discourse during the speaker's explanation of the object, s/he may look at the speaker's face and probably murmur to signal that s/he has lost the point (Fig. 1). The speaker should be able to recognize the flaw in communication, and take an appropriate action such as suspending the flow of explanation and supplementing it with more information. Thus, the conversation as a process is sustained and ceased as the result of collaboration between the speaker and listener.

The conversation-as-content view, on the other hand, focuses on how meaning emerges from interaction. Consider the situation where the speaker is telling the listener how to disassemble a device, saying "turn the lever this way to remove the component" (Fig. 2). Nonverbal behavior, such as the eye contact and gestures of the speaker, will associate the utterance with its information content as a sequence in the speaker's actions.

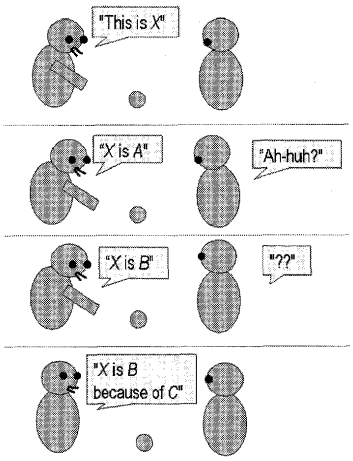


Fig. 1 Conversation as interaction.

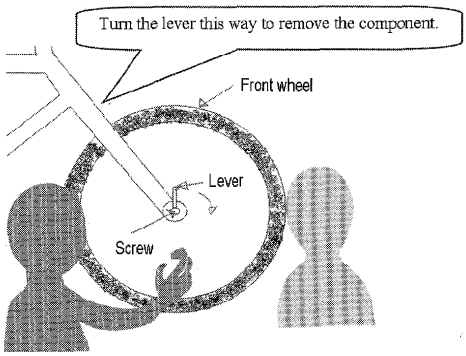


Fig. 2 Conversation as content.

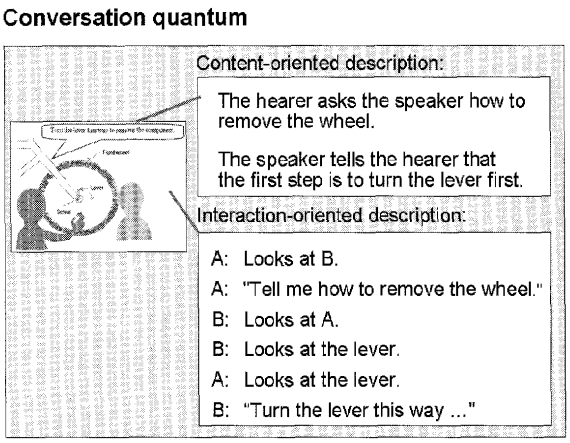


Fig. 3 Example of conversation quantum.

The role of the conversation quantum is to represent the association between information content and interaction within minimal units of conversation. For example, we can create the conversation quantum in Fig. 3 for the situation in Fig. 2. It contains a description of a visual scene where the

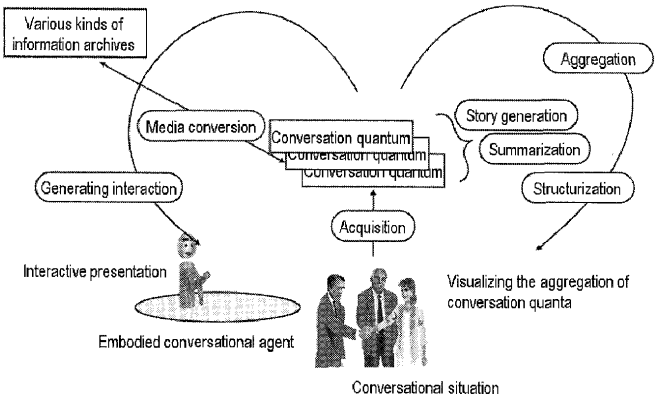
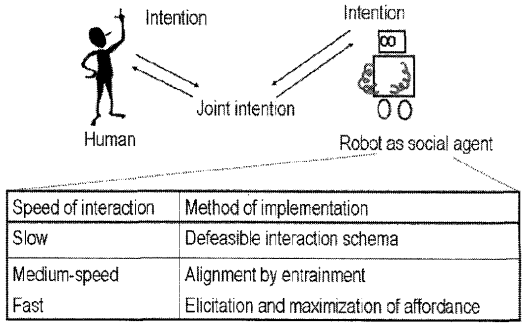


Fig. 4 Framework for conversation quantization.



Speed of interaction	Method of implementation
Slow	Defeasible interaction schema
Medium-speed	Alignment by entrainment
Fast	Elicitation and maximization of affordance

Fig. 5 Listener and presenter robots as embodied knowledge media.

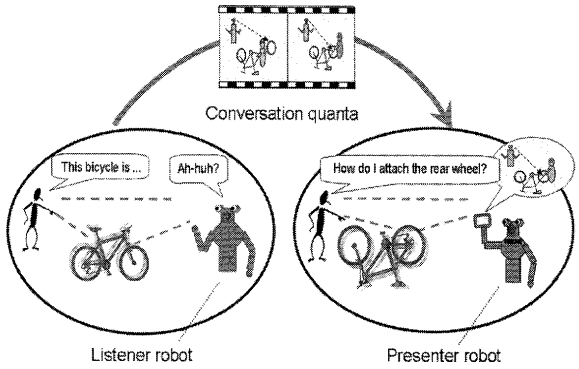


Fig. 6 Listener and presenter robots as embodied knowledge media.

speaker provides an answer to the hearer in response to the question raised by the hearer. It also contains a description of an interaction where the speaker is explaining by pointing out a component to the listener who is listening to her/him while paying attention to an object.

The approach to conversation quantization is to acquire, accumulate, and reuse conversation quanta (Fig. 4). In addition to the basic cycle of acquiring conversation quanta from a conversational situation and reusing them in the presentation through embodied conversational agents, conversation quanta may be aggregated and visually presented to the user, or some manipulation such as summarization may be added to transform one or more conversation quanta into another, or converted to and from various kinds of information archives [5].

The role of a robot is to acquire or present information and knowledge situated within the discourse, by engaging in appropriate behavior as a result of recognizing the other participants' conversational behaviors. In future, the aggregated collection of conversation quanta will serve as a knowledge base for driving robots.

Our current research focuses on establishing robust nonverbal communication that can serve as a basis for associating content with interaction.

2.2 Formation and Maintenance of Joint Intentions

We attempted to realize a communication schema that would allow two or more participants to repeat observations and reactions at varying speeds to form and maintain joint intentions to coordinate behavior, which may be called a "coordination search loop."

We propose an architecture consisting of layers to deal with interactions at different speeds to achieve this coordination search loop (Fig. 5) [6].

The lowest layer is responsible for fast interaction. We based the design of this level on affordance [7], which refers to the bundle of cues the environment provides the actor. We relied on people's capabilities of utilizing various kinds of affordances even though these are subtle. We designed the layer at this level so that a robot could suggest its capabilities to the human, coordinate its behavior with her/him, establish a joint intention, and provide the required service.

The intermediate layer is responsible for interactions at medium-speed. We introduced an entrainment-based alignment mechanism so that the robots could coordinate their behaviors with the interaction partner by varying rhythms of nonverbal behaviors.

The upper layer is responsible for slow and deliberate interactions such as those based on social conventions and knowledge to communicate more complex ideas based on the shared background. We introduced defeasible interaction patterns to describe typical sequences of behaviors actors are expected to undertake in conversational situations. A probabilistic description was used to cope with the vagueness of the communication protocol used in human society.

2.3 Listener Robot and Presenter Robot as Prototypes of Embodied Knowledge Media

We built listener and presenter robots on top of the three layer model, aiming at prototyping the idea of robots as embodied knowledge media. The pair of robots serves as a means of communicating embodied knowledge (Fig.6). The listener robot first interacts with the human with knowledge to acquire knowledge quanta. The presenter robot, equipped with a small display, will then interact with a human to show the appropriate video in appropriate situations where this knowledge is considered to be needed. Conversation quanta are used to encode knowledge transferred from the listener to the presenter robot.

We elaborate on these ideas and present some preliminary work in two sections that follow.

3. Architecture of Robots that Can Form and Maintain Joint Intentions with Humans

This section describes attempts to achieve robots that can form and maintain joint intentions to coordinate appropriate conversational behaviors with people. Some of the resulting techniques are incorporated in the listener and presenter robots described in the next section.

3.1 Elicitation and Maximization of Affordance

Affordance encompasses various kinds of information that the environment provides an actor attempting to achieve goals [7]. We considered that the robots could exhibit cooperative behaviors by maximizing the affordance they were expected to produce. The key issue was to define affordance as a simple collection of measurable physical features so that we can implement algorithms for recognizing/producing it.

To further investigate this idea, we developed an autonomous mobile chair that could dynamically produce a means of allowing a person to get a place to sit down [8]. The autonomous mobile chair perceives the relation between the surface of the actor's body and the surface of the environment, i.e., a measure called the affordance distance that is characterized as the minimal distance between the surface of the autonomous mobile chair and human body. The affordance distance decreases as the autonomous mobile chair approaches the human. The optimal action sequence depends on multiple factors such as the shape and locomotive ability of the autonomous mobile agent and the relative angle of the two surfaces.

We designed the autonomous mobile chair so that it could learn to move to a configuration where the affordance distance was minimal. The affordance distance is computed using a utility function†:

$$U(s) = R(s) + \max_a \sum_{s'} M^a_{ss'} U(s'),$$

where $R(i)$ is a reward function that will return the value of the reward in state i. $M^a_{ss'}$ is the transition probability of reaching state s' if action a is done in state s. $M^a_{ss'}$ is obtained by repeating the same action in the same state:

$$M^a_{ss'} = \frac{n_{s'}}{n^a_s},$$

where n^a_s is the number of times action a is undertaken in state s, and $n_{s'}$ is the number of states, s', reached then. We provide a reward when a certain point of the artifact's body touches a certain point of the human's body. The goal can be specified as:

$$(x, y, z)_p = (x, y, z)_q \quad \text{and} \quad (\theta, \phi, \varphi)_p = -(\theta, \phi, \varphi)_q,$$

where $(x, y, z)_p$ and $(x, y, z)_q$ are the coordinates for the points of the human's and artifact's bodies. $(\theta, \phi, \varphi)_p$ and $(\theta, \phi, \varphi)_q$ are the angles of the normal vector for points of the human's and artifact's bodies, respectively. To calculate the utility value of each state, we used a simple iterative algorithm

$$U_{t+1}(s) \leftarrow R(s) + \max_a \sum_{s'} M^a_{ss'} U_t(s'),$$

where $U_t(s)$ is the utility value of s after t iterations.

We implemented the autonomous mobile chair. Its shape, the utility function, and typical behaviors are shown in Fig. 7. We carried out several experiments. Interactions with several users are shown in Fig. 8. Although the users were all able to sit down on the chair as a result of coordinating behaviors, some users pointed out that the autonomous mobile chair should have communicated its intentions more explicitly.

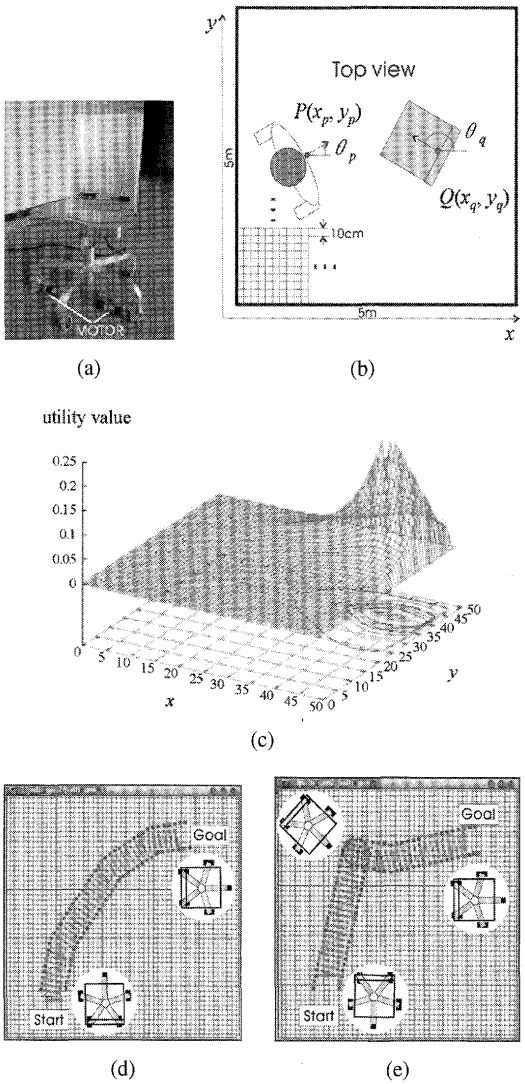
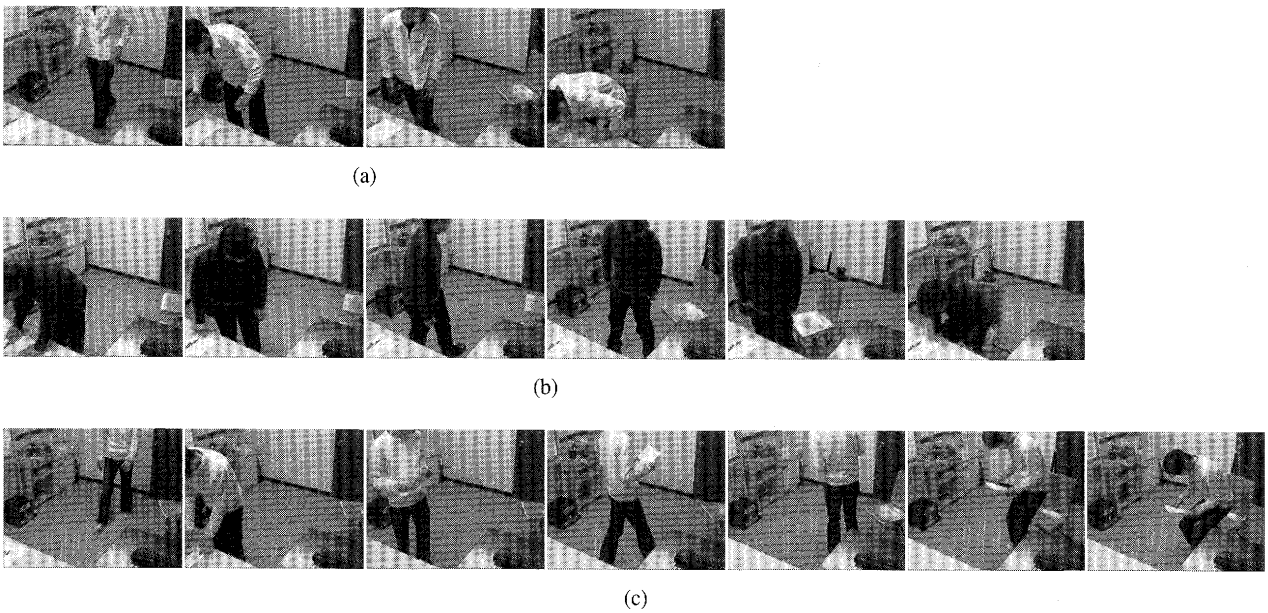


Fig. 7 Autonomous chair, its utility function, and typical behaviors. (a) Implemented autonomous chair, (b) Definition of parameters, (c) The shape of utility function, (d) One trajectory, (e) Another trajectory.

†The affordance distance is given as the inverse number of the value of the utility function.



(a)

(b)

(c)

Fig. 8 Autonomous mobile chair interacting with people. (a) Interaction with person A, (b) Interaction with person B, (c) Interaction with person C.

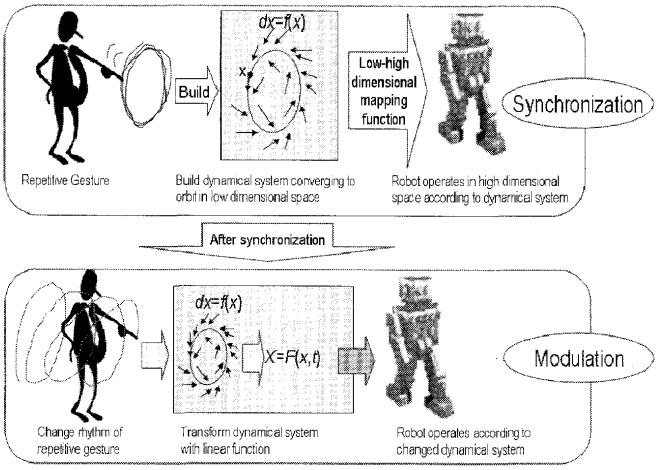


Fig. 9 Outline of entrainment-based interaction.

3.2 Entrainment-Based Interaction

Entrainment-based interaction allows a joint intention to be established in two steps (Fig. 9) [9]. The first step is called the synchronization phase. Assume one actor A wants to establish a joint intention with another actor B. First, A engages in rhythmic behavior, signaling an intention to establish a joint intention with B. When B recognizes this, B will change behavior so as to synchronize with the observed rhythm.

The second step is called the modulation phase. Once A observes that B is acting with the same rhythm, A may gradually change her/his rhythm so that B's behavior may become more desirable to A. This will cause B to modify her/his intention to converge to A's behavior.

From a theoretical point of view, synchronization and modulation are characterized as a means of bridging dynamical systems for A and B. Let us represent the state of A and B as vectors $\vec{x}$ and $\vec{y}$, and assume the intrinsic behaviors of A and B are governed by functions $f(\vec{x})$ and $g(\vec{y})$. Then, the behaviors of A and B are presented as:

$$\begin{cases} \dfrac{d\vec{x}}{dt} = f(\vec{x}) + \alpha(\vec{x}, \vec{y}) \\ \dfrac{d\vec{y}}{dt} = g(\vec{y}) + \beta(\vec{x}, \vec{y}) \end{cases}$$

where terms $\alpha(\vec{x}, \vec{y})$ and $\beta(\vec{x}, \vec{y})$ are connecting the dynamical systems for A and B.

Suppose the human and the robot play the roles of A and B, and the human's behavior is given as time series $\vec{i}(t)$ (which is assumed to be in the same vector space as $\vec{x}$ and $\vec{y}$). We pursue synchronization and modulation operations, when necessary to reflect the user's intention manifesting as $\vec{i}(t)$ on $\beta(\vec{x}, \vec{y})$. We first store the time series of $\vec{i}(t)$ for a certain period, W. When the amount of stored data reaches W, we calculate the autocorrelation function to compute cycle T after noise has been eliminated and power has been normalized. When the peak exceeds the threshold, we assume that the behavior is periodic and estimate the average pattern for the repeated behavior. High frequency segments are removed at this stage.

We then perform a method based on [10] to define $(dx/dt, dy/dt)$ in the phase space so that the repeated behavioral pattern becomes an attractor that asymptotically attracts nearby orbits.

We now proceed to the modulation phase when synchronization is detected. We calculate the deviation in input in this phase from the expected cyclic behavior. Once this is completed, we modify the basic repetitive behavior by adding the difference.

We implemented a simulated floor cleaning robot system. A human can interact with multiple robots with hand

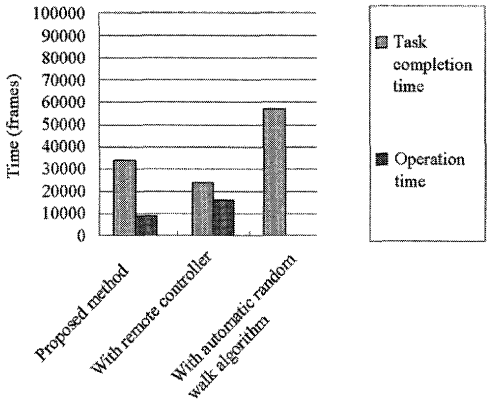


Fig. 10 Comparison of entrainment-based interaction with different algorithms.

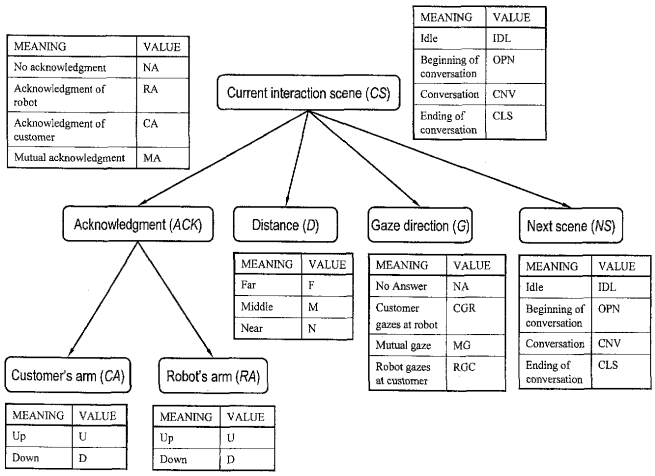


Fig. 11 Schema-based interaction using Bayesian network.

gestures captured by a motion capture device. We carried out some experiments to compare the amount of time it took to complete a cleaning task and the amount of the time required to manipulate the cleaning robot. We found that the method we propose falls between fully manual control by a remote controller and a fully automated cleaning algorithm based on a random walk, as plotted in Fig. 10.

3.3 Defeasible Interaction Patterns

We employed Bayesian networks to describe typical patterns of social conventions for conversational interactions. We used a probabilistic framework to cope with the fact that the social protocols are not rigid [11]. To investigate the issue further, we implemented a simplified version of a waiter robot that approached or stayed away depending on demands by a human. The waiter robot attempted to detect the human's intentions and the situation within the environment, by collecting and interpreting information about the inter-personal distance, direction of the gaze and the acknowledgment (ACK), in addition to the history record of human-robot interaction and the current action of the robot, based on knowledge encoded as defeasible interaction patterns.

In our experimental settings for evaluating the performance of a waiter robot interacting with a customer, we assumed that the customer had three tasks, i.e., asking for a drink, asking for an empty glass to be removed, and approaching the robot to get a drink. A typical sequence of interactions between the customer and the waiter robot was:

- the customer wanted to call the waiter robot;
- the customer raised or waved her/his hand at a distance;
- the robot approached the customer;
- the robot started to communicate with the customer.

Defeasible interaction patterns were implemented with a Bayesian network based on Bayes' principle:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(Y|X)P(X)}{P(Y)}.$$

A portion of the Bayesian network we employed is shown in Fig. 11. A random variable is represented by an elliptic node, and causality is represented by an edge connecting nodes. The table behind each node contains the random variable's possible values for the node.

We used conditional probabilities such as $P(CA|\,ACK)$ and $P(RA|\,ACK)$ to represent causal dependencies among variables, where $P(CA|\,ACK)$ stands for the probabilities of CA, i.e., the customer may lift her/his arm (CA = U) or put down her/his hand (CA = D) under various possible conditions for ACK including NA (no acknowledgment), RA (acknowledgment by robot), CA (acknowledge by customer) and MA (mutual acknowledgment). The probabilities are specified using matrices like:

$$\begin{array}{c} \\ ACK \end{array} \begin{array}{c} \begin{array}{cc} CA & \\ D & U \end{array} \\ \begin{bmatrix} 0.85 & 0.15 \\ 0.85 & 0.15 \\ 0.15 & 0.85 \\ 0.15 & 0.85 \end{bmatrix} \end{array} \begin{array}{l} \begin{array}{c} \\ NA \\ RA \\ CA \\ MA \end{array} \end{array} \qquad \begin{array}{c} \\ ACK \end{array} \begin{array}{c} \begin{array}{cc} RA & \\ D & U \end{array} \\ \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \\ 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \end{array} \begin{array}{l} \begin{array}{c} \\ NA \\ RA \\ CA \\ MA \end{array} \end{array},$$

which suggests that

$$P(CA = \mathrm{D}|\,ACK = \mathrm{NA}) = 0.85,$$
$$P(CA = \mathrm{U}|\,ACK = \mathrm{NA}) = 0.15,$$
$$P(CA = \mathrm{D}|\,ACK = \mathrm{RA}) = 0.85.$$

Similarly, we may also assume the values of other conditional probabilities:

$$\begin{array}{c} \\ CS \end{array} \begin{array}{c} \begin{array}{cccc} ACK & & & \\ NA & RA & CA & MA \end{array} \\ \begin{bmatrix} 0.65 & 0.15 & 0.15 & 0.05 \\ 0.1 & 0.4 & 0.4 & 0.1 \\ 0.05 & 0.15 & 0.1 & 0.7 \\ 0.3 & 0.2 & 0.1 & 0.4 \end{bmatrix} \end{array} \begin{array}{l} \begin{array}{c} \\ IDL \\ OPN \\ CNV \\ CLS \end{array} \end{array}$$

$$\begin{array}{c} \\ CS \end{array} \begin{array}{c} \begin{array}{ccc} D & & \\ F & M & N \end{array} \\ \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.5 & 0.4 & 0.1 \\ 0.05 & 0.15 & 0.8 \\ 0.05 & 0.15 & 0.8 \end{bmatrix} \end{array} \begin{array}{l} \begin{array}{c} \\ IDL \\ OPN \\ CNV \\ CLS \end{array} \end{array}$$

$$
\begin{array}{cc} & G \\ & \begin{array}{cccc} NA & CGR & RGC & MG \end{array} \\ CS \begin{array}{c} IDL \\ OPN \\ CNV \\ CLS \end{array} & \left[\begin{array}{cccc} 0.45 & 0.3 & 0.2 & 0.05 \\ 0.05 & 0.4 & 0.4 & 0.15 \\ 0.1 & 0.2 & 0.2 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 \end{array}\right] \end{array}
$$

$$
\begin{array}{cc} & NS \\ & \begin{array}{cccc} IDL & OPN & CNV & CLS \end{array} \\ CS \begin{array}{c} IDL \\ OPN \\ CNV \\ CLS \end{array} & \left[\begin{array}{cccc} 0.5 & 0.4 & 0.05 & 0.05 \\ 0.3 & 0.15 & 0.5 & 0.05 \\ 0.1 & 0.05 & 0.4 & 0.45 \\ 0.7 & 0.2 & 0.05 & 0.05 \end{array}\right]. \end{array}
$$

Given the values of a priori probability $P(CS)$,

$$P(CS = IDL) = 0.5, \quad P(CS = OPN) = 0.1,$$
$$P(CS = CNV) = 0.3, \quad P(CS = CLS) = 0.1$$

we can calculate the value of $P(\text{ACK} = \text{NA})$:

$$
\begin{aligned}
P(ACK{=}NA) &= P(ACK{=}NA|CS{=}IDL)\cdot P(CS{=}IDL) \\
&\quad + P(ACK{=}NA|CS{=}OPN)\cdot P(CS{=}OPN) \\
&\quad + P(ACK{=}NA|CS{=}CNV)\cdot P(CS{=}CNV) \\
&\quad + P(ACK{=}NA|CS{=}CLS)\cdot P(CS{=}CLS) \\
&= 0.65\times 0.5 + 0.1\times 0.1 + 0.05\times 0.3 + 0.3\times 0.1 \\
&= 0.38.
\end{aligned}
$$

Similarly, we can derive the values of other conditional probabilities as

$$P(ACK = RA) = 0.18, \quad P(ACK = CA) = 0.155,$$
$$P(ACK = MA) = 0.285, \quad P(CA = D) = 0.542,$$
$$P(CA = U) = 0.458.$$

Consider that CA = U (the customer has lifted her/his arm) is observed. Then, we obtain:

$$
\begin{aligned}
&P(ACK = NA|CA = U) \\
&= \frac{P(ACK = NA)\cdot P(CA = U|ACK = NA)}{P(CA = U)} \\
&= \frac{0.38\times 0.15}{0.458} \cong 0.124,
\end{aligned}
$$

which implies that the ACK = NA ("the customer does not acknowledge") becomes less probable. On the other hand, we also obtain:

$$
\begin{aligned}
&P(ACK = CA|CA = U) \\
&= \frac{P(ACK = CA)\cdot P(CA = U|ACK = CA)}{P(CA = U)} \\
&= \frac{0.155\times 0.85}{0.458} \cong 0.288,
\end{aligned}
$$

which implies that ACK = CA ("the customer acknowledges") becomes more probable.

The method was implemented using Robovie2 as a platform. Figure 12 shows how the implemented waiter robot could distinguish subtle differences resulting in a course of interaction with the customer. We carried out small-scale experiments and found that the waiter robot worked as designed.

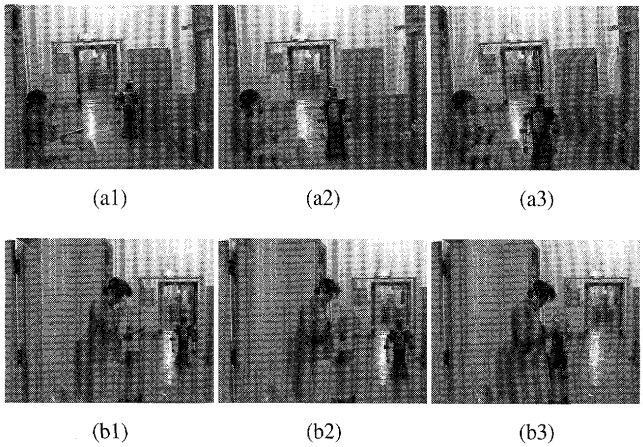


Fig. 12 Interaction with defeasible interaction patterns. (a1) C raises hand, (a2) R approaches, (a3) R approaches closer, (b1) C touches his hair, (b2) R approaches, (b3) R leaves. (C: customer, R: robot).

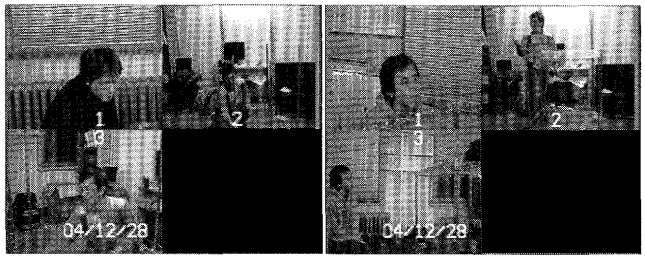


Fig. 13 Video data on explanatory scene.

4. Listener and Presenter Robots

We describe the listener and presenter robots as prototypes of the embodied knowledge medium paradigm in this section.

4.1 Listener Robot

The listener robot is designed to undertake appropriate nonverbal behavior while the presenter is explaining and producing a series of conversation quanta as records of the conversation. When the listener robot was first implemented, the subject domain was furniture assembly [12]. Later, we also applied the framework to bicycle assembly and disassembly.

We observed the conversation in detail to obtain a precise model of the behavior of the listener, where the instructor explained the procedure for assembling furniture to the listener.

We videotaped some scenes where a person as instructor was explaining how to assemble a piece of furniture (a metal rack, see Fig. 13) to another person as listener, and analyzed the video in detail using a video annotation tool, Anvil[†]. Table 1 lists the result of analysis. We found that when the instructor attended to an object, the listener attended to it more than 75% of the time. Thus, joint attention

[†]http://www.dfki.de/~kipp/anvil/

Table 1 Analysis of listener's behaviors.

	Listener's response to speaker's gaze		
	Joint attention	Gaze toward instructor	Nodding toward instructor
Listener 1	76.4	85.3	64.7
Listener 2	84.7	71.4	47.6

(%)

between the instructor and the listener was achieved very frequently. Moreover, when the instructor turned his gaze to the listener, he turned his gaze back to the instructor for more than 70% of occasions. In many cases, the listener nodded in concurrence with his gaze, but the frequency differed greatly depending on individuals. Of the exchanges occurring during communication, the following sequence of events was most frequently observed during short periods:

> the instructor turned his gaze toward the listener, the listener turned his gaze back to the instructor, the listener nodded (or did nothing), the instructor looked at the object to be explained, and the listener looked at it.

We suspect that the above sequence took place when the instructor wanted to confirm the listener was paying attention. We also obtained additional observations:

- Showing the object with his hand and turning his gaze frequently attracted the listener's attention,
- the listener usually attended to the object after the instructor exhibited multiple behaviors (e.g. showing and gazing), and
- when the instructor moved or changed his posture, the listener paid attention to the instructor himself instead of the object.

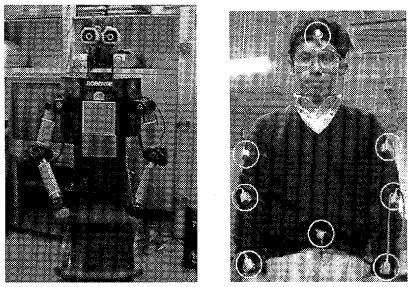
Based on the observation that attention behaviors were frequently represented with more than one modality or in repeating fashion, we propose that redundancy of attention behaviors should be applied to the design of a natural communication environment using a listener robot. The instructor's redundant behaviors strongly suggest the intention of his behaviors. Recognizing redundancy enables the listener robot to easily understand the instructor's intention. The effectiveness of applying informative redundancies to an agent's sensors and actuators is also suggested by [13].

We assumed that there were two types of redundancy in instructor-listener communication:

- Redundancy of modality: Simultaneously using multiple modalities of behaviors.
- Redundancy of time: Using repetitive or persistent behaviors.

We designed the behavior of the listener robot so that it could distinguish four modes of communication, i.e., the talking-to, talking-about, confirming, and busy modes.

In the talking-to mode, the instructor (the human) is mainly watching the listener (the robot) and is involved in the cognitive space based on the relation between them. As



(a) (b)

Fig. 14 Robobvie2 and attachment of 3D position sensors to the human (Two more sensors are attached to back of head and waist).

the instructor in the talking-to mode expects the listener to be involved in the same conversation, s/he should pay attention to her/him.

In the talking-about mode, the instructor is mainly watching the target to be explained, expecting the listener to cognitively share the target. Whenever the listener recognizes that the instructor is in the talking-about mode, the listener will attend to the target in turn.

In the confirmation mode, the instructor alternately looks at the listener and the target, suggesting that s/he is interested in whether or not the listener is paying attention to the target. The listener should sensitively react to the behaviors of the instructor.

In the busy mode, the instructor is devoting himself to his work without talking to the listener. The instructor is attending more to the target than in the talking-about mode, and tends to ignore the listener. This situation is not favorable for any explanatory task, but it frequently occurs when the instructor is not skillful. The listener robot can keep attending to the target during the busy mode.

We used a humanoid robot, called Robovie2 (Fig. 14 (a)) as the robot platform. Robovie2 is almost as tall as a human and can move its hands, head, and eyes. The motions of the instructor and the location of significant objects are captured by a motion capture system called MotionStar. The motion capture system's ten 3D position sensors are attached to the human's body (Fig. 14 (b)). In addition, several more 3D position sensors are attached to the salient objects in the environment, which will be referred to during the course of explanation. Although we sensed the speech of the human, only its power was utilized and no speech recognition facilities were employed.

The attention behaviors of the instructor were recognized in two steps.

First, the basic communicative behaviors of the human, such as eye gaze (head direction), pointing, holding, repetitive hand gestures (e.g., tapping), physical relationship to objects (i.e., distance and body direction), were recognized from the sensory data and attention toward each target was inferred. The confidence of each behavior was calculated using simple algorithms, except for the repetitive behaviors. For example, the direction of the gaze was estimated from the directions of two markers attached to the instructor's

head. Holding was recognized by calculating the distance between the hand and object. However, we did not infer the precise direction of the eyes, for this was difficult to recognize from motion capture information. Repetitive gestures of the hands were recognized by employing the methods described in Sect. 3.2.

The confidence values of respective behaviors were regarded as scores, and the redundancy of attention behaviors toward the target (redundancy of modality) was inferred by summing up the confidence values with respective weights.

Each weight depended on the current communication mode and the duration of behaviors (redundancy of time). For example, in the talking-to mode, the weights were less than in any other mode, for gazing attention at the target did not frequently occur in the talking-to mode.

The communication mode was inferred based on the recognition of target attention (with the largest redundancies), gaze at the listener robot, body direction toward the listener robot, attention behavior (e.g., posture change), speech (power, duration), and preceding communication mode. For example, the talking-to mode was inferred when frequent speech, and the instructor's body and face toward the listener were recognized, while the confirming mode was inferred if momentary gaze toward the listener was observed during target attention.

The listener robot's behaviors were determined based on the redundancy of attention behavior and the communication mode, as described in Fig. 15. The object with the largest redundancy was selected as the target for the listener robot's attention.

We implemented a listener robot. Figure 16 shows how the listener robot interacts with the human instructor. Figure 17 shows critical scenes captured during interaction between the human instructor and listener robot.

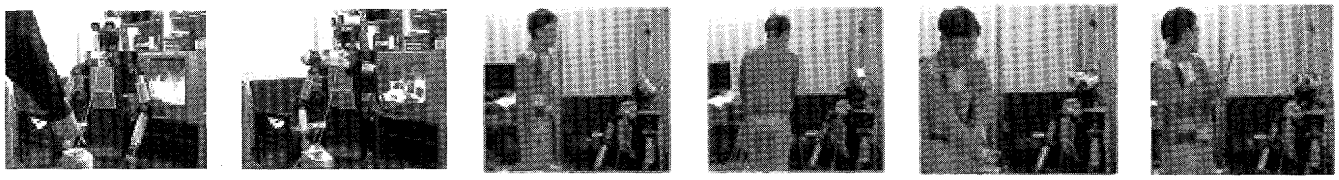
4.2 Presenter Robot

The presenter robot is designed to approach the task area when it detects the human listener's need for help [14]. It will then adjust the position and angle of the display according to the listener's position and posture, and show the video to the listener.

To approach the listener's work area and play suitable explanatory videos as shown in Fig. 18, the robot will determine distance D and direction F to the listener, based on location information obtained from motion capture. Let the

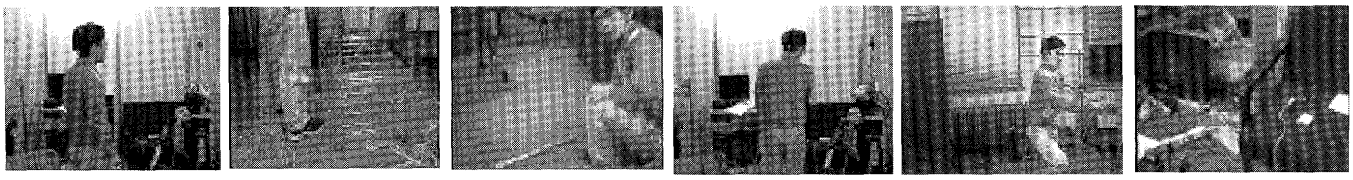
- **Talking-*to*:** Turns head to speaker and randomly nods when speech is aborted.
  - **Low redundancy:** *No attention*
  - **Medium redundancy:** *Gaze attention within short period*
  - **High redundancy:** *Slight head movement and gaze attention within short period*
- **Talking-*about*:** Uses various methods.
  - **Very low redundancy:** *No attention*
  - **Low redundancy:** *Gaze attention*
  - **Medium redundancy:** *Head attention after gaze attention*
  - **High redundancy:** *Prompt head attention*
  - **Very high redundancy:** *Head movement and verbal responses*
- **Confirming:** Takes glance at speaker, nods with verbal responses, and then returns to original state. Nodding and verbal responses are randomly produced.
- **Busy:** In same fashion as talking-*about* mode, but with no utterances.

Fig. 15 Listener robot's behavior in four communication modes.



(a) (b) (c) (d) (e) (f)

Fig. 16 Listener robot interacting with instructor. (a) Attention by gaze and head orientation, (b) Using arm to confirm object of attention, (c) Attention to instructor, (d) Joint attention by instructor pointing, (e) Joint attention by instructor lifting object, (f) Confirmation of instructor's intention of drawing attention.



(a) (b) (c) (d) (e) (f)

Fig. 17 Content captured by listener robot and cameras on floor. (a) Presentation to listener, (b) Attention to task being explained, (c) Confirmation by instructor, (d) Overview of task, (e) Instructor pointing, (f) Attention to pointed object of reference.

Fig. 18 Adjusting position of display with touch sensor.

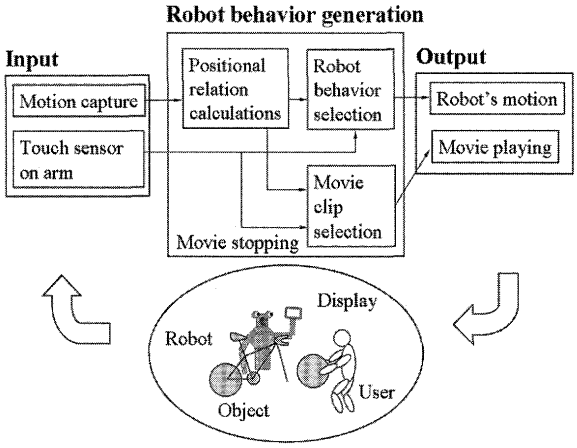


Fig. 19 Architecture of presenter robot.

center position of the robot be $O = (O_x, O_y, O_z)$, the direction of the robot's estimated gaze be $A = (A_x, A_y, A_z)$, the center position of the listener be $U = (U_x, U_y, U_z)$. D and F are calculated as:

$$D = |U - O|$$

$$F = \arccos\left(\frac{A \cdot (U - O)}{|A| \cdot |U - O|}\right).$$

The robot will move to the listener and face to her/him if D is larger than a predetermined threshold and F deviates from zero (the robot is not appropriately faced to the listener).

To select an appropriate explanatory video for the listener, the robot will search for one with features that match the current situation. To determine the zoom ratio, the presenter robot will measure how much time s/he spends on attending to the target object. The longer s/he looks at the target object, the more zoomed in the video will be.

We allow the listener to adjust the display's position so that s/he can clearly see the display. The robot will calculate the height of the listener's head and waist based on data from sensors attached to her/his body. Parameters H (that denotes the default height of an average user), α and β (the user's feedback parameters with an initial value of 0) will be used to determine the display's height, $H + \alpha$, the display's direction, β, from which the robot arm's axis, roll and pitch will be derived to adjust to the listener. The display will keep still if it is likely that the listener is working on the object. The display's optimal position depends on her/him. It will be adjusted according to her/his height and posture. The robot will also permit her/him to adjust the display's position directly by touching the tactile sensors on the robot's arm.

The architecture of the presenter robot is outlined in Fig. 19. The implementation of the presenter robot is currently in progress with the bicycle assembly as the subject domain. According to the change in task, we also reimplemented the listener robot in this domain.

5. Future Work

To complete the work on listener and presenter robots, we need to solve a handful of remaining problems. The major problem is to automate the acquisition of conversation quanta. We are currently implementing an algorithm based on analysis of nonverbal communication means and intelligent annotation. In addition, we are addressing the following problems.

(1) Intention detection using signal processing

The robot should be able to behave efficiently, to establish robust communication between itself and the humans even in noisy environments. We consider low-level intention detection using signal processing is promising and propose interactive and active perception that can serve as an amplifier of intended behavior [15].

(2) Learning

Interaction is a complex behavior, and the construction of a knowledge base for creating complex interactive behaviors is a bottleneck. Preliminary work toward modeling and acquiring knowledge for alignment is in progress. We are employing dynamic Bayesian networks as the basis and are working on implementing a learning algorithm [16].

(3) Mutual adaptation

To achieve a robust communication robot, we need to implement mutual adaptation not only by giving the robot the ability of adapting to its partner's behaviors, but also by making the learning capabilities comprehensible by the communication partner.

To collect detailed observations about humans' mutual adaptation, we devised an experiment environment [17]. We are currently carrying out extensive studies on analyzing and modeling human behavior.

6. Conversational Informatics — Communicative Robots in a Larger Perspective

The robots-as-embodied-knowledge-media view discussed in this paper fits in a larger perspective of Conversational Informatics [18], which is a field of research aiming at investigating human conversational behaviors as well as designing conversational artifacts that can interact with people in a conversational fashion. Conversational Informatics attempts to establish a new technology consisting of environmental media, embodied conversational agents, and management of conversational contents, based on the foundation of Artificial Intelligence, Pattern Recognition, and Cogni-

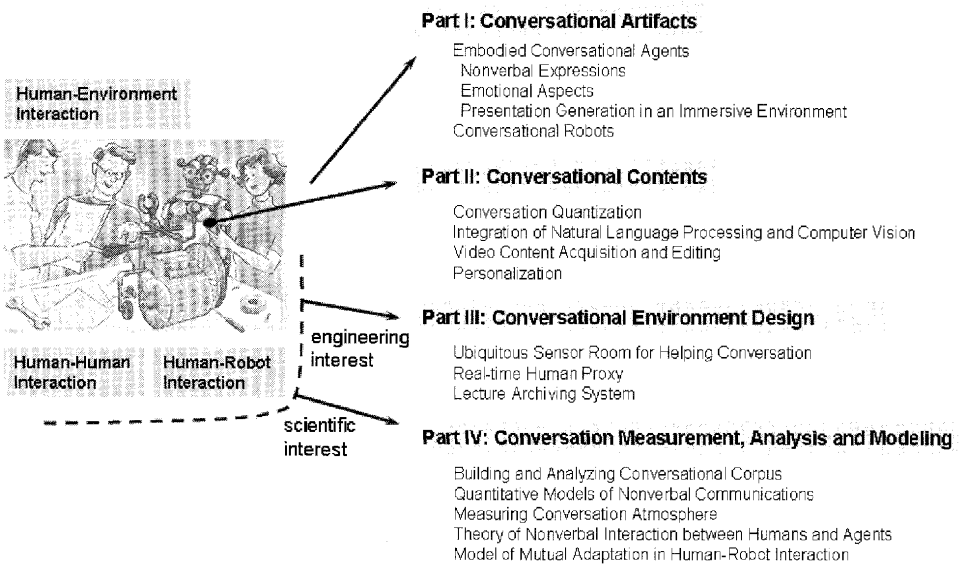


Fig. 20 Conversational Informatics.

tive Science. The major application areas of Conversational Informatics involve knowledge management and e-learning. Although Conversational Informatics covers a broad field of research encompassing linguistics, psychology and human-computer interaction and interdisciplinary approaches are truly important, the engineering aspects are emphasized that are more prominent in recent novel technical developments such as conversational content acquisition, conversation environment design and quantitative conversational modeling.

The current technical development of Conversational Informatics centers around four subjects (Fig. 20).

The first subject is conversational artifacts. We address how to build artifacts, synthetic characters on the computer screen or intelligent robots that can help the user by making a conversation not only with natural language but also with eye gaze, facial expressions, gestures, or other nonverbal communication means. Example of the major technological contributions is knowledgeable embodied conversation agents that can automatically produce emotional and socially proper communication behaviors in human-computer interaction [19], [20].

The second subject is conversational contents. We address building a suite of techniques for acquiring, editing, distributing, and utilizing the contents that can be produced and applied in the conversation. Example of the major contributions to this subject is Sustainable Knowledge Globe (SKG), which supports people to manage conversational content by using geographical arrangement, topological connection, contextual relation, and a zooming interface. By using SKG, a user can construct his content in the virtual landscape, and then explore the landscape along a conversational context. [21].

The third subject is conversation environment design. We address designing an intelligent environment that can sense the conversational behaviors to either help participants be involved in collaboration even though they may be acting at distant places or record conversation accompanying the atmosphere of the conversational behavior for later use or review. Example of contributions to this subject is a ubiquitous sensor room that can be used to measure and capture conversational behaviors [22].

The last subject is conversation measurement, analysis and modeling. Motivated by scientific interest, we take a data-driven quantitative approach to understanding conversational behaviors by measuring conversational behaviors with advanced technologies and building detailed quantitative models about various aspects of conversation. Example of contributions to this subject include is analysis of synchrony tendency, which is nonverbal behaviors such as body movement and speech interval that tend to synchronize and become mutually similar [23].

The robots-as-embodied-knowledge-media view fits in multiple issues in the perspective of Conversational Informatics. Obviously, the presenter and listener robots can be discussed in the context of conversational artifacts. An important future challenge is to build a full-fledged conversational robot that can communicate humans with verbal and nonverbal communication means. In this paper, we placed much emphasis on integration of the conversation-as-interaction and conversation-as-content views. Conversation quantization is proposed as a major framework for it, and is discussed in the context of conversational contents where conversational robots may be characterized as a device that can produce association between interaction and content. The viewpoints of Conversational Environment Design should be taken into account when we actually design an artificially augmented environment where humans and robots co-habit (Fig. 21). As the perceptual and communicative abilities of robots might be limited, the entire environment should be carefully designed so that humans and artifacts may produce useful interactions. An ecological approach is considered to be promising in the current state-of-the-art. Conversation Measurement, Analysis and Modeling constitute the basis of designing the artificial environment, especially when an ecological approach is taken.

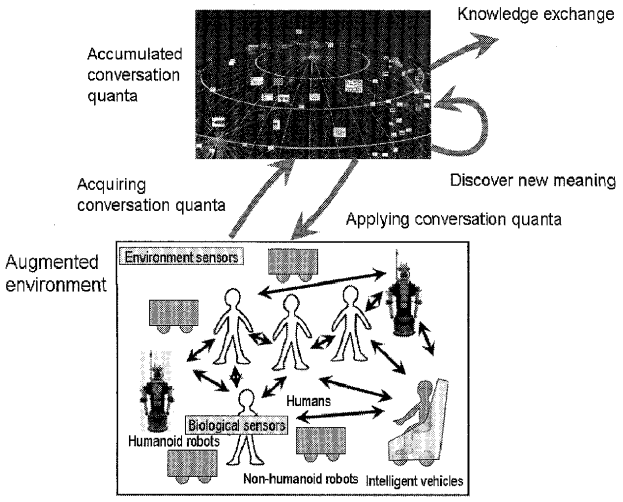


Fig. 21 Robots-as-embodied-knowledge-media in a larger perspective.

7. Conclusion

Building a system of conceptualization grounded on real world situations has been a long-term goal of Artificial Intelligence and Human Computer Interaction. In this paper, we have presented the robots-as-embodied-knowledge-media view, and described work with a layered approach to allow robots to form and maintain joint intention with humans, and a listener and presenter robot pair to create associations of information content and interaction, as a step toward achieving this goal.

References

[1] Y.I. Nakano, G. Reinstein, T. Stocky, and J. Cassell, "Towards a model of face-to-face grounding," Proc. 41st Annual Meeting of the Association for Computational Linguistics (ACL 03), pp.553–561, 2003.

[2] A. Kendon, Gesture: Visible Action As Utterance, Cambridge University Press, 2004.

[3] D. McNeill, Gesture and Thought, The University of Chicago Press, 2005.

[4] B. Reeves and C. Nass, The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places, Cambridge University Press, 1996.

[5] T. Nishida, "Conversation quantization for conversational knowledge process," Special Invited Talk, ed. S. Bhalla, DNIS 2005, LNCS 3433, pp.15–33, Springer, 2005.

[6] T. Nishida, "Real agents," in Robot Informatics, Chapter 4, ed. Y. Anzai, H. Tokuda, T. Nishida, H. Hagita, and M. Hirose, Iwanami Lecture Series, Robotics, vol.5, Iwanamishoten, 2005 (in Japanese).

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

[8] K. Terada and T. Nishida, "Active artifacts: For new embodiment relation between human and artifacts," Proc. 7th International Conference on Intelligent Autonomous Systems (IAS-7), Marina del Rey, CA, March 2002.

[9] T. Tajima, Y. Xu, and T. Nishida, "Entrainment based human-agent interaction," 2004 IEEE Conference on Robotics, Automation and Mechatronics, Singapore, Dec. 2004.

[10] M. Okada, K. Tatani, and Y. Nakamura, "Polynomial design of the nonlinear dynamics for the brain-like information processing of whole body motion," Proc. of IEEE International Conference on Robotics and Automation (ICRA2002), vol.2, pp.1410–1415, 2002.

[11] M. Hatakeyama, Human-Robot Interaction Based on Interaction Schema, Master's Thesis, Graduate School of Information Science and Technology, the University of Tokyo, 2004.

[12] Y. Ogasawara, M. Okamoto, Y.I. Nakano, and T. Nishida, "Establishing natural communication environment between a human and a listener robot," Proc. Symposium on Conversational Informatics for Supporting Social Intelligence and Interaction, AISB, pp.42–51, 2005.

[13] R. Pfeifer and C. Scheier, Understanding Intelligence, Bradford Books, 1999.

[14] T. Ohya, T. Hiramatsu, Y. Xu, Y. Sumi, and T. Nishida, "Towards robot as an embodied knowledge medium — Having a robot talk to humans using nonverbal communication means," SID-2006, accepted for presentation, 2006.

[15] Y.F.O. Mohammad and T. Nishida, "Interactive perception for amplification of intended behavior in complex noisy environment," SID-2006, accepted for presentation, 2006.

[16] K. Tarasenko and T. Nishida, "Dynamic Bayesian networks for modeling of mutual adaptation with communicative robots," SID-2006, accepted for presentation, 2006.

[17] Y. Xu, K. Ueda, T. Komatsu, T. Okadame, T. Hattori, Y. Sumi, and T. Nishida, "WOZ experiments for understanding mutual adaptation," SID-2006, accepted for presentation, 2006.

[18] T. Nishida, ed., Special Issue on Conversational Informatics, J. Jpn. Soc. Artif. Intell., vol.21, no.2, 2006.

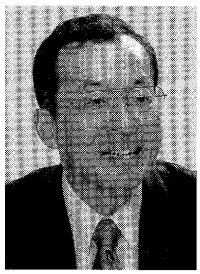
[19] Y.I. Nakano, T. Murayama, and T. Nishida, "Multimodal story-based communication: Integrating a movie and a conversational agent," IEICE Trans. Inf. & Syst., vol.E87-D, no.6, pp.1338–1346, June 2004.

[20] M. Okamoto, Y.I. Nakano, K. Okamoto, K. Matsumura, and T. Nishida, "Producing effective shot transitions in CG contents based on a cognitive model of user involvement," Special Issue of Life-like Agent and Its Communication, IEICE Trans. Inf. & Syst., vol.E88-D, no.11, pp.2623–2532, Dec. 2005.

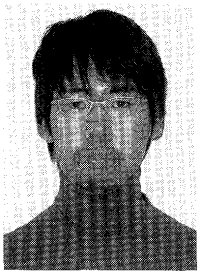
[21] H. Kubota, Y. Sumi, and T. Nishida, "Sustainable knowledge globe: A system for supporting content-oriented conversation," presented at AISB 2005 Symposium Conversational Informatics for Supporting Social Intelligence & Interaction — Situational and environmental information enforcing involvement, 2005.

[22] Y. Sumi, S. Ito, T. Matsuguchi, S. Fels, and K. Mase, "Collaborative capturing and interpretation of interactions," Pervasive 2004 Workshop on Memory and Sharing of Experiences, pp.1–7, April 2004.

[23] C. Nagaoka and S. Yoshikawa, "Synchrony tendency of body movement in therapist-client counseling situation," Paper presented at Annual Meeting of Japanese Psychological Society, 2005.

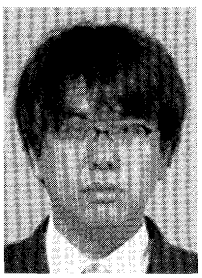


Toyoaki Nishida received his Ph.D. in Information Science from Kyoto University in 1984. He is currently a Professor at the Department of Intelligence Science and Technology, Kyoto University, Kyoto, Japan. His research centers on artificial intelligence and human computer interaction. His current research focuses on social intelligence design and communicative intelligence. He led several projects, including a JSPS (Japan Society for the Promotion of Science) project on intelligent media technology for supporting natural communication between people, and a RISTEX (The Research Institute of Science and Technology for Society) project on the conversational knowledge process for risk communication.



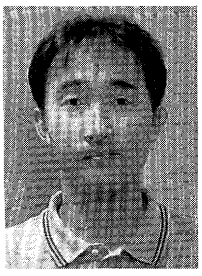
Kazunori Terada received his Ph.D. in Information Science from Nara Institute of Science and Technology in 2001. He is currently a Research Associate at the Department of Information Science, Gifu University, Gifu, Japan. His research interests include artificial intelligence and cognitive science, focusing on human-artifact interaction, affordance, and visual-haptic integration.

Takashi Tajima received his B.S. in Electronics Engineering from the University of Tokyo in 2002, and received his M.S. in Information and Communication Engineering from the University of Tokyo in 2004. He joined Matsushita Electric Industrial Co., Ltd. in 2004. He is currently working at the Advanced Technologies Research Laboratories of Matsushita Electric Industrial Co., Ltd. His research interests include human agent interaction. He is a member of the Japanese Society for Artificial Intelligence.

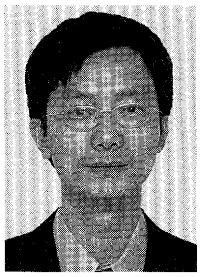


Makoto Hatakeyama received his B.E. and M.E. in Information and Communication Engineering from the University of Tokyo in 2002 and 2004. He joined NEC Corporation in 2004 and is now a researcher at Internet Systems Research Laboratories. He is engaged in research on identity management.

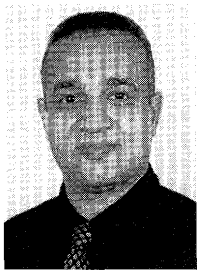
Yoshiyasu Ogasawara received his B.E. and Master of Science and Technology in Information and Communication Engineering from the University of Tokyo in 2003 and 2005. He joined Sharp Corporation in 2005. He is currently working with the Audio Visual System Group at Sharp Corporation.



Yasuyuki Sumi is an associate professor in Graduate School of Informatics at Kyoto University since 2003. He received his B.Eng. from Waseda University in 1990, and his M.Eng. and D.Eng. degrees in Information Engineering from the University of Tokyo in 1992 and 1995. He was a researcher at ATR from 1995 to 2003. His research interests include knowledge-based systems, creativity supporting systems, interface/social agents, ubiquitous/wearable computing, Web intelligence, multimedia processing, and their applications for facilitating human interaction and collaboration.

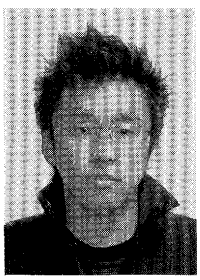


Yong Xu graduated with B.E. in Computer Software and Applications from Suzhou Vocational University in 1993, and received his Master's in Circuits and Systems from the Department of Electronic Science and Technology at the University of Science and Technology of China in 2001. He is currently a Ph.D. student at the Department of Intelligence Science and Technology of Kyoto University in Japan. His research centers on human-robot interaction. His current research focuses on a mutually adaptive human-robot interface.



Yasser F. O. Mohammad received his B.S. and M.S. in Electrical Engineering from the Faculty of Engineering of Assiut University in 1998 and 2004. During 1999–2005 he worked as a TA, and RA in the Faculty of Engineering of Aswan University and Assiut University studying Network Security, Computer Vision, and Robotics. He is now a Ph.D. student at the Nishida-Sumi Laboratory of Kyoto University.

Kateryna Tarasenko received her B.S. and M.S. in Intelligence Systems for Information Processing and Decision Making from the National Technical University of Ukraine, Kiev Polytechnic Institute in 2001 and 2003. She entered the Graduate School of Informatics of Kyoto University, Japan as a research student in 2005. Her research interests include embodied conversational agents, and the simulation of non-verbal communication behaviors.



Taku Ohya received his B.S. from Kyoto University in 2005. He studied learning algorithms for human computer interaction. He is now working in the laboratory of Professor Toyoaki Nishida at the Department of Intelligence Science and Technology of Kyoto University, and is studying social intelligence design and communicative intelligence.

Tatsuya Hiramatsu is an MS student at the Information and Mathematical Science Department of Kyoto University, studying Information Science. His present research focuses on interactions between humans and robots.