Cognitive Development in Partner Robots for Information Support to Elderly People

Akihiro Yorita, *Student Member, IEEE*, and Naoyuki Kubota, *Member, IEEE*

***Abstract*—This paper discusses an utterance system based on the associative memory of partner robots developed through interaction with people. Human interaction based on gestures is quite important to the expression of natural communication, and the meaning of gestures can be understood through intentional interactions with a human. We therefore propose a method for associative learning based on intentional interaction and conversation that can realize such natural communication. Steady-state genetic algorithms (SSGA) are applied in order to detect the human face and objects via image processing. Spiking neural networks are applied in order to memorize the spatio–temporal patterns of human hand motions and various relationships among the perceptual information that is conveyed. The experimental results show that the proposed method can refine the relationships among this varied perceptual information that can then inform an updated relationship to natural communication with a human. We also present methods of assisting memory and assessing a human's state.**

***Index Terms*—Associative memories, cognitive science, intelligent systems, robots, speech communication.**

I. Introduction

ONE OF THE great problems of the aging of society is the increased number of elderly people who live alone and separate from their children. Such elderly people often do not even experience daily conversation, a lack that can lead to cognitive decline and a high risk of dementia. In particular, the decline of their capacity for memory, attention, and planning can have a terrible impact on the safety of their daily lives [1], [2]. Various types of useful tools have been developed to support the memory of the elderly [3]. For example, there is a system that reminds us them of things they did in the past by taking snapshots of their kitchen [4]. Most of these kinds of methods, however, are not interactive, which is why communication robots are also used to offer support to the elderly [5]. Paro, for example, is a robotic baby seal designed for use in healthcare environments [6], [7] that has a healing effect equal to that of real pets. The conversational robot "ifbot" is also used in nursing homes [8]. It is based on the principle that nursing care can foster the health of the elderly by providing conversations with robots [9], [10]. Robotic conversation can activate the brain of the elderly and improve both their concentration and memory. These kinds of robots can gather information about a person and actually learn from the contents of their dialog. In this way, the conversational capability of a robot is applied to the prevention of dementia in the elderly. It is difficult, however, for a robot to converse appropriately with a person even if various contents of the conversation are designed in advance. For this reason, personal information that is natural to the flow of conversation is required for more natural conversation with robots. That is, the robot should be able to perceive personal information about the elderly in the course of their actual communication and interaction with them.

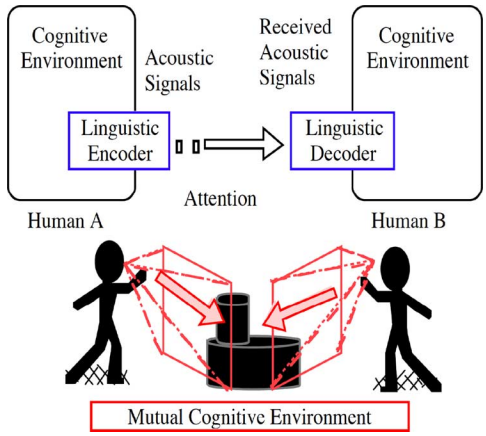

Fig. 1. Mutual cognitive environment in communication between humans.

The issue of social communication has been discussed in sociology, developmental psychology, relevance theory, and embodied cognitive science [11]–[16]. Cognitive psychology has tried to construct a mind with a computer [17]. In the society of mind theory proposed by Minsky, intelligence is explained as a being a combination of multiple simpler things. He said that although our agent is intelligence itself, it is not enough to simply explain what each separate agent does. Rather, it is a group of agents that can accomplish things [14]. The relevance theory also offers insight to a discussion of human communication [15]. According to this theory, human thought is not just transmitted, but is in fact a shared event between two people. Each person has his/her own cognitive environment, as shown in Fig. 1. One person can understand the meaning of an unknown word spoken by another because the person makes the symbol correspond to the percept, even though they speak different languages. Therefore, an important role of utterance and gesture is their stimulation of attention, and utterances and gestures can enlarge the cognitive environment of other people. Such a shared cognitive environment is called a mutual cognitive environment. To be effective, then, a robot should also have a cognitive environment. To this end, the relationships among

Manuscript received March 09, 2010; revised August 29, 2010; accepted December 15, 2010. Date of publication January 17, 2011; date of current version March 16, 2011.

The authors are with the Department of Systems Design, Tokyo Metropolitan University, Hino, Tokyo, Japan (e-mail: yorita-akihiro@sd.tmu.ac.jp; kubota@tmu.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAMD.2011.2105868

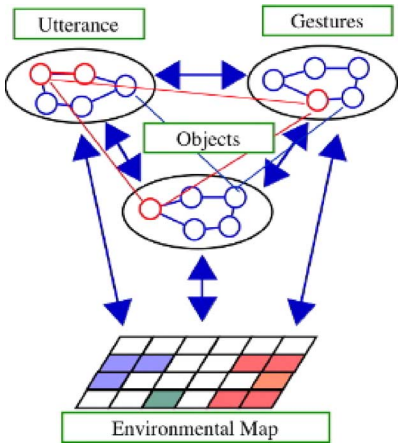

Fig. 2. Environmental state.

linguistic terms, gestures, and objects are built upon the environmental state, as shown in Fig. 2.

To develop a cognitive environment in robots, we focus on the refinement of associative memory by using the symbolic information used for utterances and patterns, which is based on visual information obtained through interaction with people.

With regard to associative memory, Nakano, Kohonen, and Anderson proposed it in 1970s [18]–[20]. After that, the Hopfield network that was proposed, which applied associative memory in 1982 [21]. In general, the Hopfield network is an autoassociative fully connected network that consists of a single layer of nodes. On the other hand, the architecture of bidirectional associative memory (BAM) is not a matrix, but rather, is a two-layer neural network. BAM is a heteroassociative, nearest-neighbor, pattern-matching network that encodes binary or bipolar pattern pairs using Hebbian learning [22]. There are two types of recall, autoassociative means recalling the whole part from a piece, heteroassociative means recalling one thing to another. Heteroassociative memory is used in daily conversation.

We have proposed the concept of associative learning and discussed the importance of the total architecture of the learning mechanism [23]. Association is defined that the relation has already been learned and new information is learned over the relation. Recall is performed using this relation. A role of associative learning is to associate new information with information he/she has already possessed. We use BAM to associate many elements.

In the case of using temporal information like voice recognition, a time-delay neural network was used. We use spiking neural network because it can learn spatio–temporal patterns and avoid sequential spike output.

Our system enables a robot to communicate with a human and to exchange appropriate content. Beyond this, the robot actually learns from the content of a conversation, and from that point forward, it understands certain individual characteristics of the person it has spoken with.

This paper is organized as follows. Section II introduces the idea of partner robots. Section III explains computational intelligence technologies and the total architecture of associative learning. Section IV presents the experimental results obtained from partner robots based on the proposed method.

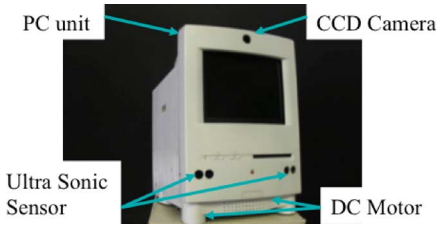

Fig. 3. Partner robot, MOBiMac.

II. Cognitive Development of Partner Robots

A. Partner Robots

We have developed a partner robot, which is a mobile PC called MOBiMac [24], in order to realize social communication with a human (see Fig. 3). The robot has two CPUs and many sensors, such as a CCD camera, microphone, and ultrasonic sensors, which enable the robot to perform image processing, voice recognition, target tracing, collision avoidance, map building, and imitative learning.

In this paper, we focus on the cognitive development of partner robots through their interaction with people. In this application, then, no movement is required of the robot. As a basic policy of this study, we employed flexible and adaptive methods for search and learning. Various types of methods have been proposed that can accomplish this; we selected steady-state genetic algorithms (SSGA) for the search, and spiking neural networks (SNN) for the memorization of spatio–temporal information [25], [26].

Fig. 4 shows a total architecture of the perception, decision making, learning, and action. First, the voice recognition and image processing are performed to extract visual and verbal information through the interaction with a person. In this paper, the robots use perceptual modules for various modes of image processing, such as differential extraction, human detection, object detection, and human hand-motion recognition. We used Voice Elements DTalker 3.0, which was developed by EIG Co., Ltd., Japan, for voice recognition and synthesis in the robot [27]. It was able to perform voice recognition using a sound segment network that made speaker-independent recognition possible. In addition, with the number of words that are recognized dependent on the memory, it achieved a recognition rate of 96.5% (for 200 words).

After that, the robot selects the conversation mode from: 1) scenario-based conversation; 2) daily conversation; and 3) learning conversation. In the scenario-based conversation mode, the robot makes utterances sequentially according to the order of utterances in a scenario. In the daily conversation, the robot uses a long-term memory based on SNN. The robot selects an utterance according to the long-term memory corresponding to the internal states of spiking neurons. In the learning conversation, the robot updates the relationship between spiking neurons used in long-term memory by the associative learning. Finally, the robot makes utterance. In the following sections, we

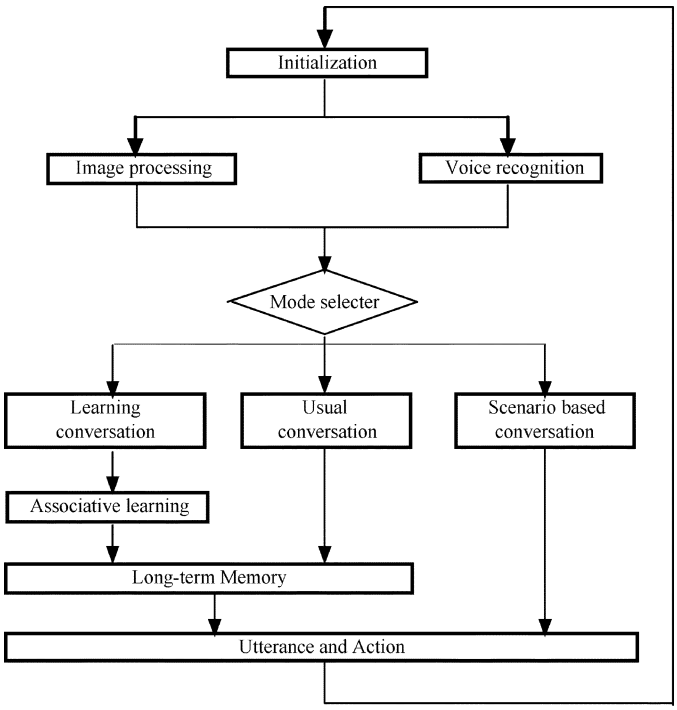

Fig. 4. Flow of learning.

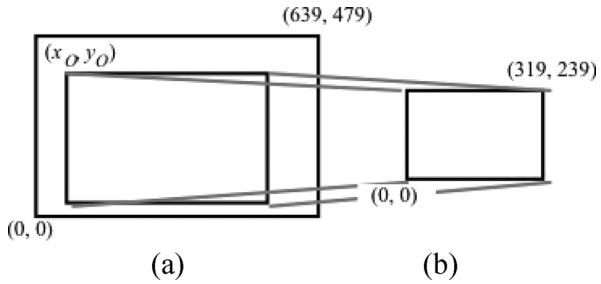

Fig. 5. Human face detection for joint attention. (a) Original image (b) Attention range.

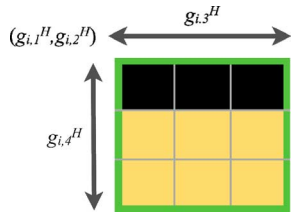

Fig. 6. Template used for human detection in SSGA-H.

explain the image processing based on SSGA, and associative learning between perceptual information and verbal words.

B. Human Detection and Tracking

Various types of pattern matching methods such as template matching, a cellular neural network [28], neocognitron [29], and dynamic programming (DP) matching have been applied to human detection in image processing. In general, pattern matching is composed of two steps: target detection and target recognition. The aim of target detection is to extract a target candidate from an image, and the aim of target recognition is to identify the target from among the classification candidates.

Since image processing consumes much costly computational time, full-size image processing of each image is not practical. We therefore used a reduced size image to detect a moving object to achieve rapid human candidate detection. First, an image of RGB color space is taken by a CCD camera installed on the partner robot. Next, the robot calculates the center of gravity (COG) of the pixels that are different from those in the previous image as a differential extraction. The size of the image used in the differential extraction is updated according to the previous result of human detection. Here, the area generated by the differential extraction is called an attention range. If the robot does not move, the COG of the difference represents the location of the moving object. To achieve rapid human detection, then, the main search area for detection of a human is formed according to the COG in the attention range. In this paper, the original size of an image is 640×480, and the size of this image is reduced to 320×240 as an attention range according to the reduction level ($1.0 \leq RL \leq 2.0$) and the origin (x_O, y_O) of the attention range (see Fig. 5). If the reduction level is 1, the same resolution of the image is cut off from the original image. Otherwise, each pixel in the attention range is interpolated according to the four surrounding pixels based on the reduction level.

The robot must swiftly recognize a human face against a complex background. To achieve this, as one of the search methods we used a SSGA for human detection. Using template matching, the SSGA extracts the human face candidate positions based on human skin and hair colors (see Fig. 6).

SSGA is used as one of the stochastic search methods because it can easily obtain feasible solutions through environmental changes with low computational cost. SSGA simulates a continuous model of the generation, which eliminates and generates a few individuals in a generation (iteration) [30], [31]. The genotype is represented by $g_{i,j}$ $(i = 1, 2 \ldots G, j = 1, 2 \ldots M)$ and the fitness value is represented by f_i. One iteration is composed of selection, crossover, and mutation. The worst candidate solution is eliminated (using a "delete least fitness" selection strategy), and is replaced by the candidate solution generated by the crossover and the mutation.

We used elitist crossover and adaptive mutation [24]. Elitist crossover randomly selects one individual and generates an individual by combining genetic information from the selected individual and the individual with the best crossover probability. If the crossover probability is satisfied, the elitist crossover is performed. Otherwise, a simple crossover is performed between two randomly selected individuals. Next, the following adaptive mutation is performed upon the generated individual:

$$g_{i,j} \leftarrow g_{i,j} + \left(\alpha_j \cdot \frac{f_{\max} - f_i}{f_{\max} - f_{\min}} + \beta_j \right) \cdot N(0,1) \quad (1)$$

where f_i is the fitness value of the ith individual, $f_{\max}$ and $f_{\min}$ are the maximum and minimum fitness values in the population, $N(0,1)$ denotes a normal random variable with a mean of zero and a variance of one, and α_j and β_j are the coefficients ($0 < \alpha_j < 1.0$) and offset ($\beta_j > 0$), respectively. In adaptive mutation, the variance of the normal random number is changed

relatively according to the fitness values of the population in the case of maximization problems.

Fig. 6 shows a candidate solution of a template used for detecting a human face. A template is composed of the numerical parameters $g^H_{i,1}$, $g^H_{i,2}$, $g^H_{i,3}$, and $g^H_{i,4}$. The number of individuals is G^H. A superscript H stands for the parameter for human detection. The fitness value of the ith individual is calculated by the following equation:

$$f^H_i = C^H_{\rm Skin} + C^H_{\rm Hair} + \eta^H_1 \cdot C^H_{\rm Skin} \cdot C^H_{\rm Hair} - \eta^H_2 \cdot C^H_{\rm Other} \quad (2)$$

where $C^H_{\rm Skin}$, $C^H_{\rm Hair}$ and $C^H_{\rm Other}$ indicate the numbers of pixels of the colors corresponding to human skin, human hair, and other colors, respectively; and η^H_1 and η^H_2 are the coefficients ($\eta^H_1, \eta^H_2 > 0$). Because this results in the problem of maximization, the iteration of SSGA is repeated until the termination condition is satisfied. Here, the SSGA for human detection is called SSGA-H.

Since SSGA extracts the area of skin colors and hair colors for human detection, various objects other than humans might also be detected. For this reason, human tracking is performed according to the time series position of the ith human candidate $(g^H_{i,1}, g^H_{i,2})$ obtained by SSGA-H. The position of the jth human candidate in the human tracking $(X_{k,1}, X_{k,2})$ is updated by the nearest human candidate position within the tracking range. In addition, the width and height of the human candidate for human tracking $(X_{k,3}, X_{k,4})$ are updated by the size of the detected human $(g^H_{i,3}, g^H_{i,4})$. This update is performed as follows ($j = 1$, 2, 3, 4):

$$X_{k,j} \leftarrow (1-\lambda)X_{k,j} + \lambda \cdot g^H_{i,j}. \quad (3)$$

Furthermore, a time counter is used to provide reliability in the human tracking. If the position of the human candidate in human tracking is determined, the time counter is incremented; if not, it is decremented. If the time counter exceeds the threshold (*HT*), a human count is started. Sometimes, several human candidates are close to one another, because human detection is able to generate several human candidates in a single human. When human candidates coexist within the tracking range in this way, removal processing is performed.

The direction the face is pointing can be approximately extracted using the relative positions of the hair and face. We apply spiking neurons to extract the direction of the detected human face, and we use the relative position of the COG of areas corresponding to the hair and face. The relative positions of the COG against the central position of the detected face region are used as inputs to the spiking neurons in order to extract the direction that the face is pointing. Fig. 7 shows experimental results of human detection. The snapshots show the system can detect face regardless of the person has hair or not [see Fig. 7(a), (b)].

C. Object Recognition

We will now explain a method for object recognition. We focus on color-based object and shape recognition using SSGA based on template matching. Here, the SSGA for object recognition is called SSGA-O. The shape of a candidate template is generated by the SSGA-O. We used an octagonal template. Fig. 8 shows a candidate template used for detecting a target in which the jth point $g^O_{i,j}$ of the ith template is represented by $(g^O_{i,1} + g^O_{i,j}\cos(g^O_{i,j+m}), g^O_{i,2} + g^O_{i,j}\sin(g^O_{i,j+m}))$, $i = 1, 2 \ldots G^O$, $j = 3, 4 \ldots 2m+2$; $O_i (= (g^O_{i,1}, g^O_{i,2}))$ is the center of a candidate template on the image; and G^O and m are the number of candidate templates and the searching points used in a template, respectively. A superscript O stands for the parameter for object recognition. Therefore, a candidate template is composed of the numerical parameters of $(g^O_{i,1}, g^O_{i,2} \ldots g^O_{i,2m+2})$. The fitness value is calculated as follows:

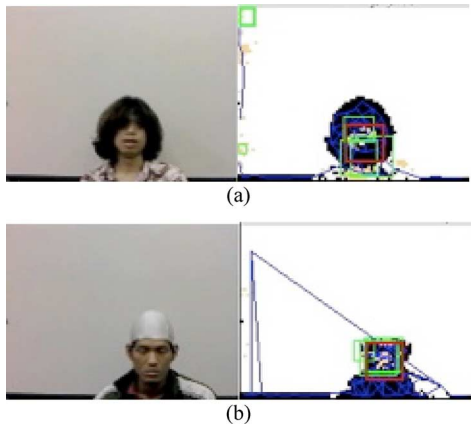

Fig. 7. Results of human face detection in SSGA-H. (a) The person who has a thick head of hair. (b) The person who wears a swim cap to hide his hair.

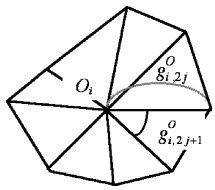

Fig. 8. Template used for object detection in SSGA-O.

$$f^O_i = C^O_{\rm Target} - \eta^O \cdot C^O_{\rm Other} \quad (4)$$

where η^O is a coefficient for penalty ($\eta^O > 0$), and $C^O_{\rm Target}$ and $C^O_{\rm Other}$ denote the number of pixels of the colors corresponding to a target and to other colors included in the template, respectively. The target color is selected according to the pixel color that occupies most of the template candidate, so that the largest area of a single color is then extracted on the reduced color space of the image.

Furthermore, we apply a k-means algorithm for the clustering of candidate templates in order to find several objects simultaneously. The inputs to the k-means algorithm are the central positions of the template candidates: $\mathbf{u}_j$ $(= (g^O_{i,1}, g^O_{i,2}))$, $j = 1, 2 \ldots K$). The number of clusters is K. When the reference vector of the kth cluster is represented by $\mathbf{r}_k = (r_{k,1}, r_{k,2} \ldots r_{k,m})$, the Euclidian distance between the ith input vector $\mathbf{u}_i = (g_{i,1}, g_{i,2} \ldots g_{i,k})$ and the kth reference vector is defined as

$$d_{i,k} = \|\mathbf{u}_i - \mathbf{r}_k\|. \quad (5)$$

Next, the reference vector minimizing the distance $d_{i,k}$ is selected by

$$\mathbf{c}_i = \arg\min_k \{\|\mathbf{u}_t - \mathbf{r}_k\|\} \tag{6}$$

where $\mathbf{c}_i$ is the cluster number that the ith input belongs to. After selecting the nearest reference vector to each input, the kth reference vector is updated by the average of the inputs belonging to the kth cluster. If the update is not performed during the clustering process, the updating process is complete. The crossover and selection are performed with the template candidates from each cluster. Therefore, SSGA-O tries to find different objects within each cluster according to the spatial distribution of objects in the image.

D. Human Hand Motion Extraction and Learning

To extract and classify human hand gesture, we use a SNN and self-organizing map.

Cluster analysis is used for grouping or segmenting observations into subsets or clusters based on similarity. A self-organizing map (SOM), a K-means algorithm, growing neural gases, and a Gaussian mixture model are often applied as clustering algorithms [32]. An SOM can be used for incremental learning, while a K-means algorithm and Gaussian mixture model use all the data observed in the learning phase (batch learning). In this paper, we apply SOM to the clustering of the spatio–temporal patterns of pulse outputs from the SNN. Furthermore, the neighboring structure of units can be used in a further discussion of the similarity of clusters.

Various types of artificial neural networks have been proposed to realize clustering, classification, nonlinear mapping, and control [33]–[35]. Basically, artificial neural networks are classified into pulse-coded neural networks and rate-coded neural networks, from the viewpoint of their level of abstraction [33]. A pulse-coded neural network approximates the dynamics of the ignition phenomenon of a neuron and the propagation mechanism of the pulse between neurons. The Hodgkin–Huxley model, one of the classic neuronal spiking models, has four differential equations. An integrate-and-fire model with a first-order linear differential equation is known as a neuron model of a higher abstraction level. A spike response model is slightly more general than the integrate-and-fire model, because the spike response model can choose kernels arbitrarily. Rate-coded neural networks, on the other hand, neglect the pulse structure, and are therefore considered to be neuronal models of a higher level of abstraction. McCulloch-Pitts and the Perceptron are also well known as famous rate-coding models [34], [35]. One important feature of pulse-coded neural networks is their temporal coding capability. In fact, various types of SNNs have been applied to the memorization of spatial and temporal context.

We use a simple spike response model to reduce the computational cost. First of all, the internal state $h_i(t)$ is calculated as follows:

$$h_i(t) = \tanh\left(h_i^{\mathrm{syn}}(t) + h_i^{\mathrm{ext}}(t) + h_i^{\mathrm{ref}}(t)\right). \tag{7}$$

Here, a hyperbolic tangent is used to avoid the bursting of neuronal fires, $h_i^{\mathrm{ext}}(t)$ is the input to the ith neuron from the external environment, and $h_i^{\mathrm{syn}}(t)$, which includes the output pulses from other neurons, is calculated by

$$h_i^{\mathrm{syn}}(t) = \gamma^{\mathrm{syn}} \cdot h_i(t-1) + \sum_{j=1, j\neq i}^{N} w_{j,i} \cdot h_j^{EPSP}(t). \tag{8}$$

Furthermore, $h_i^{\mathrm{ref}}(t)$ indicates the refractoriness factor of the neuron, $w_{j,i}$ is a weight coefficient from the jth to ith neuron, $h_j^{\mathrm{EPSP}}(t)$ is the excitatory postsynaptic potential (EPSP) that is approximately transmitted from the jth neuron at the discrete time t, N is the number of neurons, and γ^{syn} is the temporal discount rate. The presynaptic spike output is transmitted to the connected neuron according to the EPSP, which is calculated as follows:

$$h_i^{\mathrm{EPSP}}(t) = \sum_{n=0}^{T} \kappa^n p_i(t-n) \tag{9}$$

where κ is the discount rate ($0 < \kappa < 1.0$), $p_i(t)$ is the output of the ith neuron at the discrete time t, and T is the time sequence to be considered. If the neuron is fired, R is subtracted from the refractoriness value in the following:

$$h_i^{\mathrm{ref}}(t) = \begin{cases} \gamma^{\mathrm{ref}} \cdot h_i^{\mathrm{ref}}(t-1) - R, & \text{if } p_i(t-1) = 1 \\ \gamma^{\mathrm{ref}} \cdot h_i^{\mathrm{ref}}(t-1), & \text{otherwise} \end{cases} \tag{10}$$

where γ^{ref} is the discount rate. When the internal potential of the ith neuron is larger than the predefined threshold, a pulse is outputted as follows:

$$P_i(t) = \begin{cases} 1, & \text{if } h_i(t) \geq q_i \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where q_i is the threshold for firing. The weight parameters are trained based on the temporal Hebbian learning rule as follows:

$$w_{j,i} \leftarrow \tanh\left(\gamma^{wgt} \cdot w_{j,i} + \xi^{wgt} \cdot h_j^{\mathrm{EPSP}}(t-1) \cdot h_i^{\mathrm{EPSP}}(t)\right) \tag{12}$$

where γ^{wgt} is the discount rate and ξ^{wgt} is the learning rate.

SOM is often applied to extract a relationship among observed data, since it can ascertain the hidden topological structure from the data. The inputs to SOM are given as the weighted sum of pulse outputs from the neurons

$$\mathbf{v} = (v_1, v_2 \ldots v_N) \tag{13}$$

where v_i is the state of the ith neuron. In order to consider the temporal pattern, we use $h_i^{\mathrm{EPSP}}(t)$ as v_i, although the EPSP is used when the presynaptic spike output is transmitted. When the ith reference vector of SOM is represented by $\mathbf{r}_i$, the Euclidian distance between an input vector and the ith reference vector is defined as

$$d_i = \|\mathbf{v} - \mathbf{r}_i\| \tag{14}$$

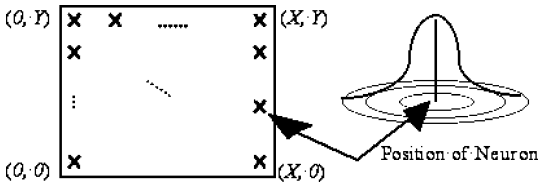

Fig. 9. Spiking neurons for gesture recognition.

where $\mathbf{r}_i = (r_{1,i}, r_{2,i} \ldots r_{N,i})$ and the number of reference vectors (output units) is M. Next, the kth output unit that minimizes the distance d_i is selected by

$$k = \arg\min_i \{\|\mathbf{v} - \mathbf{r}_i\|\}. \tag{15}$$

Furthermore, the reference vector of the ith output unit is trained by

$$\mathbf{r}_i \leftarrow \mathbf{r}_i + \xi^{\mathrm{SOM}} \cdot \zeta^{\mathrm{SOM}}_{k,i} \cdot (\mathbf{v} - \mathbf{r}_i) \tag{16}$$

where ξ^{SOM} is a learning rate ($0 < \xi^{\mathrm{SOM}} < 1.0$), and $\zeta^{\mathrm{SOM}}_{k,i}$ is a neighborhood function ($0 < \zeta^{\mathrm{SOM}}_{k,i} < 1.0$).

The robot extracts human hand motion from a series of images using SSGA-O, in which the maximal number of images is TG. The sequence of hand positions is represented by $\mathbf{G}(t) = (Gx(t), Gy(t))$ where $t = 1, 2 \ldots TG$. Here, the spiking neurons are arranged on a planar grid (see Fig. 9) and $N = 25$. By using the value of a human hand position, the input to the ith neuron is calculated by the Gaussian membership function as follows:

$$h_i^{\mathrm{ext}}(t) = \exp\left(-\frac{\|\mathbf{c}_i - \mathbf{G}(t)\|^2}{2\sigma^2}\right) \tag{17}$$

where $\mathbf{c}_i = (c_{x,i}, c_{y,i})$ is the position of the ith spiking neuron on the image, and σ is the standard deviation. The sequence of pulse outputs $p_i(t)$ is obtained using the human hand positions $\mathbf{G}(t)$. Because the adjacent neurons along the trajectory of the human hand position are easily fired as a result of the temporal Hebbian learning, the SNN can memorize the temporal firing patterns of various gestures.

Accordingly, the output unit that is selected is the pattern that is most similar to the previously learned human hand motion patterns.

For example, we show the person moving the ball. The robot recognizes the ball and gesture (see Fig. 10).

E. Associative Learning for Cognitive Development

This subsection explains a method for associative learning in the perceptual system for cognitive development. Symbolic information is quite useful and helpful for learning the relationships among patterns. In this paper, we focus on refining the association of the perceptual information with other information (see Fig. 11). To do this, we use SNNs.

Various types of utterance systems and language processing systems have been proposed [36]–[39]. Expert systems and guide robots have only to answer questions, therefore they

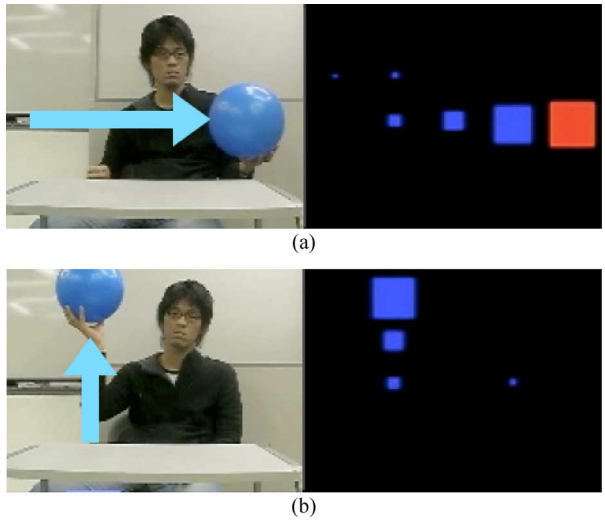

Fig. 10. Gesture recognition of the person. (a) The gesture of horizontal direction. (b) The gesture of vertical direction.

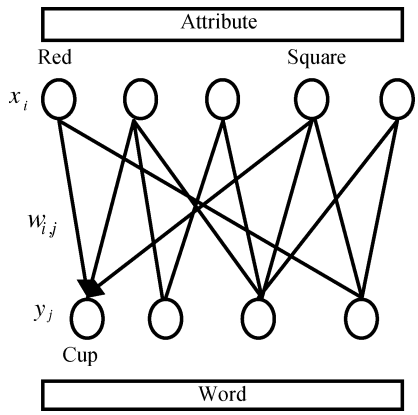

Fig. 11. Learning relationship with SNN.

do not need to learn and it is desirable to have knowledge in advance. They have only to do scenario conversation. On the other hand, the robot living with human needs to be able to support effectively by exchanging information each other. To know him/her, we developed learning conversation mode for the robot. And talking from the robot, we also developed usual conversation mode. It is thought that the robot adapts actual environment not only having knowledge but also getting it by learning. In [39], bayesian network is used but we used SNN to learn temporal patterns. In this paper, we propose an utterance system in which there are three modes. In the *scenario-based conversation mode*, a human speaks to the robot using words of greeting. In the *usual conversation mode*, the robot speaks words it has learned in the *learning conversation mode* according to a human's state. Basically, the robot speaks from input images. The learning conversation mode begins when the human says, "look this."

In the scenario-based conversation mode, the robot utters words that are prepared in advance. In the usual conversation mode, the robot utters words it has learned. In the learning conversation mode, a human teaches the robot regarding the name, color, shape, and use of an object. The robot obtains

Fig. 12. Gesture pattern of respective objects and as an example of book gesture.

human information from this and utters words that are appropriate to the individual and/or the context. The robot relates the input from the voice and image by associative learning. Using this relationship in the usual conversation mode, the robot can then utter the appropriate words. The selection probability (s_i^P) of the ith utterance group is calculated using a Boltzmann selection scheme as follows:

$$S_i^P = \frac{\exp\left(\frac{w_{j,i}}{\tau^U}\right)}{\sum_{J=1}^{J} \exp\left(\frac{w_{j,i}}{\tau^U}\right)} \tag{18}$$

where τ^U is a positive parameter called the temperature. When the temperature is high, the robot randomly selects an utterance group. As the temperature decreases, the robot deterministically selects the utterance group with the highest selection strength. This system enables the robot to integrate perceptual information and symbolic information and to produce utterances based on the external environment.

III. Experimental Results

This section presents the experimental results of a conversation with a partner robot. The number of utterance words is 50. The population size of SSGA-H and SSGA-O is 100. The number of spiking neurons in the gesture recognition is 25. The number of gestures in SOM is 50. The gesture recognition for object handling begins if the position of the hand and object is near and their velocity is also similar. Gestures used in the experiment is shown in Fig. 12. We use three types of gestures.

Fig. 13 presents different stages of image processing: (a) the original image; a photograph; (b) differential extraction; (c) the reference vectors of SOM corresponding to gestures; (d) object recognition results by SSGA-O; (e) human detection results by SSGA-H; and (f) EPSP of the spiking neurons. The subject was reading a yellow book, which he held in front of himself throughout this experiment. The proposed method for human detection and tracking extracted his face and his hands. In Fig. 13(e), the green box indicates the candidates for human face position produced by SSGA-H, the red box indicates the face position produced by human tracking, and the pink box indicates the hand position. SSGA-O was able to detect a red cup, a yellow book, and a blue ball, as shown in Fig. 13(d). The robot noticed the yellow book because there is a white rectangle on it. Fig. 13(f) shows the degree of EPSP from a spiking neuron, which indicates the spatio–temporal pattern captured from the subject's hand motion. The red rectangle is EPSP, and it gradually diminishes, turns blue, and becomes smaller. Fig. 13(c) shows the reference vectors of SOM that are learned through the interaction with the subject.

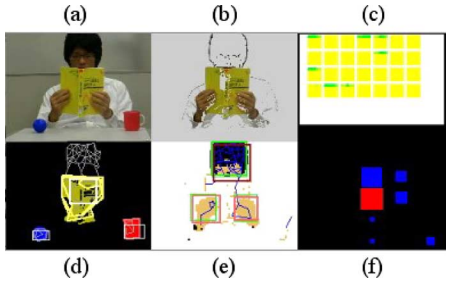

Fig. 13. Robot performs associative learning interacting with the person. (a) The original image, a photograph, (b) differential extraction, (c) the reference vectors of SOM corresponding to gestures, (d) object recognition results by SSGA-O, (e) human detection results by SSGA-H, the green box indicates the candidates for human face position produced by SSGA-H, the red box indicates the face position produced by human tracking, and the pink box indicates the hand position and (f) EPSP of the spiking neurons. which indicates the spatio–temporal pattern captured from the subject's hand motion. The red rectangle is EPSP, and it gradually diminishes, turns blue, and becomes smaller.

TABLE I
List of Learning Relationships

Name	Cup, Mug, Glass	Tea, Green tea	Ball	Book
Color	Red	Green	Blue	Yellow
Shape	Square	Square	Round	Square
Do	Drink	Drink	Play	Read

A. Learning Process Between Symbolic Information and Perceptual Information

We list the attributes of the objects used in this experiment in Table I. For each object, we teach the robot its use, color, and shape. The robot learns the relationship between words and attributes. In a concrete manner, it is then able to share a cognitive environment with a human.

First, we taught the robot concerning a cup, a ball, and a book, one by one, and confirmed the state of learning.

We show the results of learning in Figs. 14 and 15. Fig. 13 shows the relationship among words, colors, shapes, and gestures. The nodes represent each element, and the edge between nodes represents whether or not a relationship exists. The robot learned relationships from the initial state. In initial state, the robot learned nothing and then did not respond when the person showed objects. But after learning, the robot shared cognitive environment with the person, the robot can talk about the object that the person showed. As advantage point, because the robot talked to the person, the robot takes the initiative in the conversation. This is the effect of usual conversation mode, after usual conversation mode, the robot shifts scenario-based conversation mode.

The content that the robot utters are different from the words for learning. When the person make the robot learn the relationship, the person only utters words. On the other hand the robot utters sentence concerning the word the robot learned. For example, when the subject had a cup, the robot uttered "where is

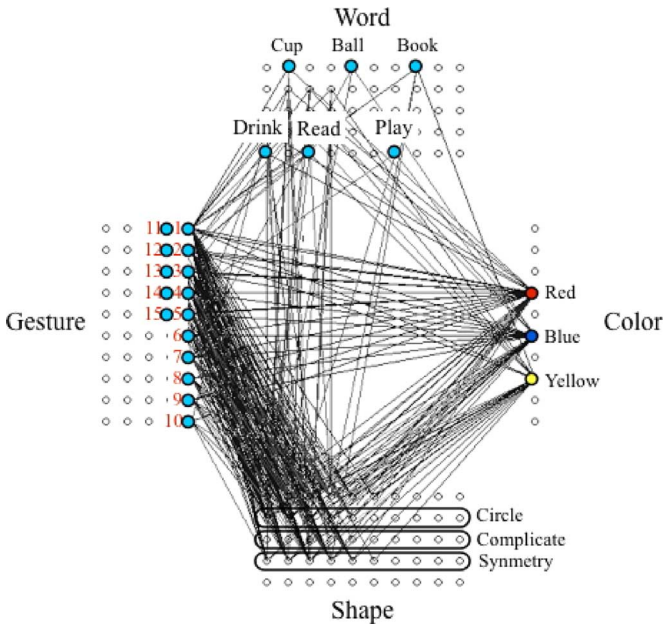

Fig. 14. Relationships of several words.

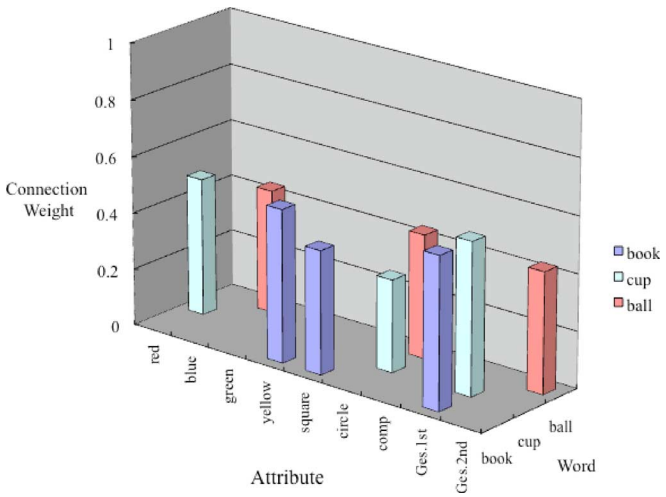

Fig. 15. Relationship among words.

the cup?" We made such the content in advance. When the robot perceives red, the words such as "cup" or "drink" are recalled, (18) determines which word is selected. The number of gesture is different in each word, these are distinguished by gesture. Therefore gesture is important in communication. In the case of showing cup, "drink" is selected, the robot utters "what do you drink?" Fig. 14 quantitatively shows the relationships only for words in a noun form (cup, book, and ball). As regards color, the robot can learn other characteristics through words, but as regards shape and gesture, the robot learned the same characteristics per words. As regards object recognition, because the shape changes how objects are shown or the lighting condition, it was a little difficult for the robot to learn what we wanted it to. Following this, we created the same circumstance and the robot uttered related words.

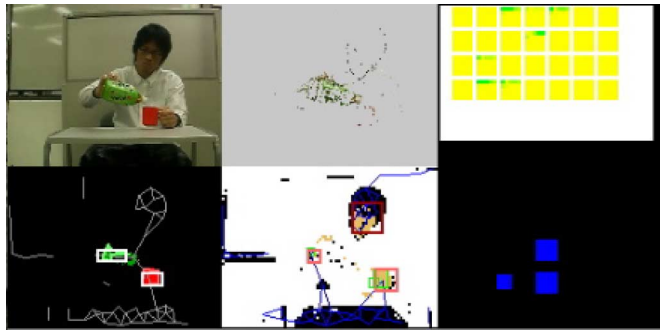

Fig. 16. Experimental results of human tracking, object recognition, and gesture recognition.

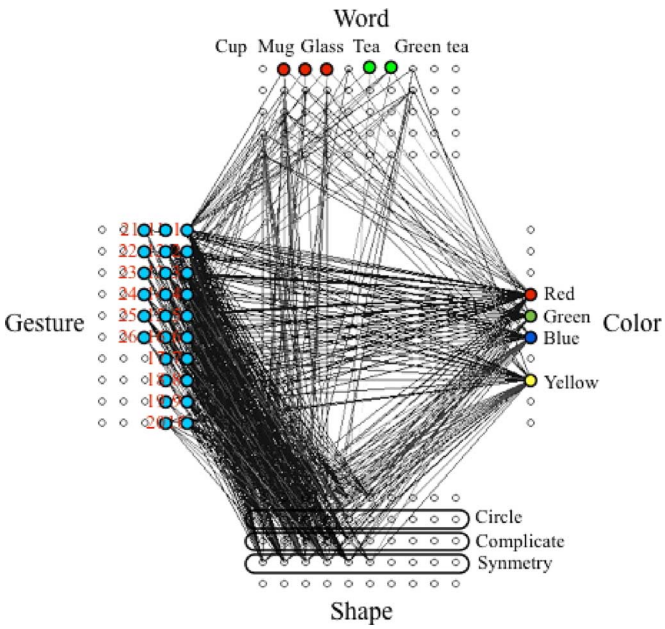

Fig. 17. Relationships of a number of words.

B. Learning Among Object Names

Next, we taught the robot the names of several objects. In this case, it was the names of three kinds of objects: cup, mug, and glass. The object of this is to hold a conversation with the robot using objects whose names everybody normally uses.

In this experiment, the robot could utter three types of words and could engage in conversation with a human using words that are easy to use. The robot can recognize only words registered in advance. And it is probably that the objects are called different names by different person, to associate multiple names with same perceptual information becomes the appropriate way that the person communicates with the robot.

C. Learning Process Between Action and Symbolic Information

Next, we put two objects in front of the robot to see whether it could explain them.

Here, we added "tea" to the learning content that the robot would learn in Fig. 16. In Japanese, there are two ways to refer to tea (*ocha* and *ryokucha*).

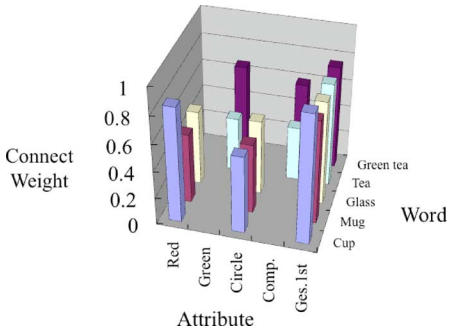

Fig. 18. Relationships among words.

After the learning took place, we interacted with the robot. We wished to project an image of the situation of drinking tea, and thus showed the robot a cup and tea. The robot was able to utter words related to the attention that it had paid to these objects. The results are shown in Figs. 17 and 18. The robot was able to learn that the same attribute could pertain to different things. It is possible to talk about both cup and tea by same gesture and to talk about various contents by making multiple contents with different words.

IV. Conclusion

This paper has discussed the capability of associative learning for partner robots that is produced through interaction with a human, based on the relevance theory. We proposed methods of associative learning that can lead a robot to produce natural utterances and to assist with memory while evaluating a human state. The experimental results show the effectiveness of these methods for human–robot interaction and show that a robot can learn the relationships among a variety of symbolic information used for utterances and can also make determinations based upon visual information. As a result, the associative capability of the robot is able to be refined through its actual interaction with a human. In this way, the proposed method is able realize more natural communication with people that can be applied to support the health and well-being of the elderly.

As a future work, we will also include the factor of emotional intelligence. It has been shown that humans easily remember events that they experienced in a certain affective state [40]. We will then extract facial expressions to be used in combination with associative memory. We will conduct experiments on associative learning based on the actions of partner robots in homes for the elderly. Furthermore, we will discuss the learnability of the proposed method in detail.

References

[1] D. M. Rentz and S. Weintraub, "Neuropsychological detection of early probable Alzheimer's disease," in *Early Diagnosis and Treatment of Alzheimer's Disease*, L. F. M. Scinto and K. R. Daffner, Eds. Totowa, NJ: Humana Press, 2000, pp. 169–189.

[2] P. Barberger-Gateau *et al.*, "Neuropsychological correlates of self-reported performance in instrumental activities of daily living and prediction of dementia," *J. Gerontol. Series B: Psychol. Sci. Social Sci.*, vol. 54, no. 5, pp. 293–303, 1999.

[3] M. Pollack, "Intelligent Technology for an aging population: The use of AI to assist elders with cognitive impairment," *AI Mag.*, vol. 26(2), pp. 9–24, 2005.

[4] Q. T. Tran, G. Calcaterra1, and E. D. Mynatt, *COOK'S COLLAGE*. Boston, MA: Springer-Verlag, 2005.

[5] N. Roy, G. Baltus, D. Fox, F. Gemperle, J. Goetz, T. Hirsch, D. Margaritis, M. Montemerlo, J. Pineau, J. Schulte, and S. Thrun, "Towards personal service robots for the elderly," in *Proc. Workshop Interact. Robot. Entertainment*, Pittsburgh, PA, 2000.

[6] K. Wada, T. Shibata, T. Saito, and K. Tanie, "Effects of robot assisted activity to elderly people who stay at a health service facility for the aged," in *Proc. 2003 IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Las Vegas, NV, 2003, pp. 2847–2852.

[7] W. Taggart, S. Turkle, and C. D. Kidd, "An interactive robot in a nursing home: Preliminary remark," *Toward Social Mech. Android Sci., Cogn. Sci. Soc.*, pp. 56–61, 2005.

[8] M. Kanoh, S. Kato, and H. Itoh, "Facial expressions using emotional space in sensitivity communication robot "ifbot"," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sendai, Japan, 2004, pp. 1586–1591.

[9] M. Heerink, Kröse, B. Wielinga, and V. Evers, "Enjoyment intention to use and actual use of a conversational robot by elderly people," in *Proc. 3rd ACM/IEEE Int. Conf. Human Robot Interact. (HRI 2008)*, Amsterdam, The Netherlands, 2008, pp. 113–120.

[10] J. Cassell, "Embodied conversational agents: Representation and intelligence in user interface," *AI Mag.*, vol. 22, no. 3, pp. 67–83, 2001.

[11] R. Pfeifer and C. Scheier, *Understanding Intelligence*. Cambridge, MA: The MIT Press, 1999.

[12] M. W. Eysenck, *Psychology: An Integrated Approach*. Harlow, Essex, U.K.: Longman, 1998.

[13] R. L. Gregory, *The Mind*. London, U.K.: Oxford Univ. Press, 1998.

[14] M. Minsky, *The Society of Mind*. New York: Simon and Schuster, 1986.

[15] D. Sperber and D. Wilson, *Relevance – Communication and Cognition*. Oxford, U.K.: Blackwell, 1995.

[16] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: A survey," *IEEE Trans. Autonom. Mental Develop.*, vol. 1, no. 1, pp. 12–34, May 2009.

[17] U. Neisser, *Cognitive Psychology*. New York: Appleton, 1967.

[18] K. Nakano, "Associatron – A model of associative memory," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, no. 3, pp. 380–388, Jul. 1972.

[19] T. Kohonen, "Correlation matrix memories," *IEEE Trans. Comput.*, vol. C-21, no. 4, pp. 353–359, Apr. 1972.

[20] J. A. Anderson, "A simple neural network generating interactive memory," *Math. Biosc.*, vol. 14, pp. 197–220, 1972.

[21] J. J. Hopfield, "Neurons with graded response have collective com- putational properties like those of two-state neurons," in *Proc. Nat.Acad. Sci. USA*, 1984, vol. 81, pp. 3088–3092.

[22] B. Kosko, "Bi-directional associative memories," *IEEE Trans. Syst. Man., Cybern.*, vol. 18, pp. 49–60, Jan. 1988.

[23] N. Kubota, "Computational intelligence for structured learning of a partner robot based on imitation," *Inform. Sci.*, no. 171, pp. 403–429, 2005.

[24] N. Kubota and K. Nishida, "Development of internal models for communication of a partner robot based on computational intelligence," in *Proc. 6th Int. Symp. Adv. Intell. Syst.*, Nagoya, Japan, 2005, pp. 577–582.

[25] N. Kubota, "Computational intelligence for human detection of a partner robot," in *Proc. (CD-ROM) of World Automation Congress*, Seville, Spain, 2004.

[26] N. Kubota, Y. Tomioka, and M. Abe, "Temporal coding in spiking neural network for gesture recognition of a partner robot," in *Proc. Joint 3rd Int. Conf. Soft Comput. Intell. Syst. 7th Int. Symp. Adv. Intell. Syst.*, 2006, pp. 737–742.

[27] DTalker for Mac OSX Ver3.0 [Online]. Available: http://www.createsystem.co.jp/dtalkerMacOSX.html

[28] L. O. Chua and L. Yang, "Cellular neural networks: Theory," *IEEE Trans. Circuits Syst. I*, vol. CAS1-35, no. 10, pp. 1257–1272, Oct. 1988.

[29] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 193, p. 202, 1980.

[30] G. Syswerda, "A study of reproduction in generational and steady-state genetic algorithms," in *Foundations of Genetic Algorithms*. San Mateo, CA: Morgan Kaufmann, 1991.

[31] D. B. Fogel, *Evolutionary Computation*. Piscataway, NJ: IEEE Press, 1995.

[32] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Berlin, Heidelberg: Springer-Verlag, 2001.

[33] W. Gerstner, *Pulsed Neural Networks*, W. Maass and C. M. Bishop, Eds. Cambridge, MA: MIT Press, 1999, pp. 3–53.

[34] J.-S.R. Jang, C.-T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing*. Englewood Cliffs, NJ: Prentice-Hall Inc., 1997.

[35] J. A. Anderson and E. Rosenfeld, *Neurocomputing*. Cambridge, MA: MIT Press, 1988.
[36] J. Weizenbaum, "Eliza–A computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, 1966.
[37] F. Hayes-Roth, D. A. Waterman, and D. B. Lenat, *Building Expert Systems*. Reading, MA: Addison-Wesley, 1983.
[38] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, N. Hagita, and Y. Anzai, "Humanlike conversation with gestures and verbal cues base on a three-layer attention-drawing model," *Connect. Sci.*, vol. 18, no. 4, pp. 379–402, 2006.
[39] T. Inamura, M. Inaba, and H. Inoue, "A dialogue control model based on ambiguity evaluation of users'instructions and stochastic representation of experiences," *J. Robot. Mechatron.*, vol. 17, no. 6, pp. 697–704, 2005.
[40] G. H. Bower, "Mood and memory," *Amer. Psychol.*, vol. 36, no. 2, pp. 129–148, 1981.

Akihiro Yorita (S'08) graduated from Saitama University, Saitama, Japan, in 2007. He received the M.E. degree from Tokyo Metropolitan University, Tokyo, Japan, in 2009. He is currently working towards the Ph.D degree at Tokyo Metropolitan University.

His research interests include a dialog system of partner robots and human–robot interaction.

Naoyuki Kubota (S'95–A'97–M'01) received the B.Sc. degree from Osaka Kyoiku University, Kashiwara, Japan, in 1992, He received the M.Eng. degree from Hokkaido University, Hokkaido, Japan, in 1994, and the D.E. degree from Nagoya University, Nagoya, Japan, in 1997.

He joined the Osaka Institute of Technology, in 1997. In 2000, he joined the Department of Human and Artificial Intelligence Systems, Fukui University as an Associate Professor. He joined the Department of Mechanical Engineering, Tokyo Metropolitan University in 2004, and has been an Associate Professor of the Department of System Design, Tokyo Metropolitan University, Tokyo, Japan, since 2005.