How can humanoid acquire lexicon? —active approach by attention and learning biases based on curiosity—

Masaki Ogino, Masaaki Kikuchi and †Minoru Asada
Dept. of Adaptive Machine Systems, †*JST ERATO Asada Synergistic Intelligence Project,*
Graduate School of Engineering, Osaka University
2-1, Yamadaoka, Suita, Osaka 565-0871, Japan
{*ogino, kikuchi*}*@er.ams.eng.osaka-u.ac.jp, asada@ams.eng.osaka-u.ac.jp*

***Abstract*— Observation study of human infants tells us that they can successfully acquire lexicon; understanding the relationship between the meaning and the uttered word from only one teaching by caregiver, even though there are many other possible mappings. This paper proposes a lexical acquisition model which makes use of curiosity to associate visual features of observed objects with the labels that are uttered by a caregiver. A robot changes its attention and learning rate based on curiosity. In the experiment with a real humanoid robot, the visual features are represented with self organizing maps which adaptively represents the shape of observed objects independent of the viewpoints.**

I. Introduction

Human infants learn new words at an incredible rate from around 18 months, and they acquire a vocabulary of 1,000 to 2,000 words by the time they are two [1]. This is called ”language explosion” or ”lexical explosion”, and one of the biggest mysteries of human cognitive developmental process. A constructive approach to this mystery by building a robot that can reproduce this function seems promising to reveal the underlying mechanism of this process [2].

The existing bottom-up approach in machine learning to lexicon acquisition has focused on symbol grounding problem in which the problem treated is how to connect sound information from caregiver and sensor information that a robot captures from the environment [3] [4] [5] [6]. A typical method proposed in these studies is based on the estimation of the co-occurrence probabilities between the words uttered by a caregiver and the visual features that a robot observes. In these experiments, training data sets are given by the caregiver, and the robot passively learns them.

However, such a statistical method does not seem sufficient to explain the lexical explosion. It is observed that human infants can acquire the lexical relationship between the meaning and the uttered word only from one teaching, even though there are many other possibilities. Cognitive psychologists have proposed that infants utilize some rules or constraints to acquire lexicon efficiently. Markman [7] proposed the whole object constraint and the mutual exclusivity constraint. Landau et al. [8] proposed the geometrical constraint. The word order can be used for constraining the meaning of the words, and some methods are proposed that use grammatical information to acquire the lexical relationship and to categorize the acquired words [9] [10].

Moreover, infants are not passive creatures. They actively and intentionally interact with the environment around them [11]. The period when an infant starts to learn is overlapped with that when he/she starts to walk. The existing methods proposed in machine learning have neglected this active attitude of infants, and training data are passively received by the infants. It is well known that infants have selectivity for novel things and events. It is shown from many observations that they look longer at novel things than at known ones. This selectivity is thought to take an effective role in acquiring information for new events and so in language acquisition.

The active selection of motions including visual attention might take an important role in lexicon acquisition. It is important to make a curiosity model with which an agent decides how to react to the environment depending on its current knowledge so that it can acquire necessary information. Saliency is one of the fundamental factors for making this conscious and subconscious motivational process. Saliency is supposed to be evaluated by comparing with something in novelty and frequency. Walther et al. [12] proposed a visual attention model in which saliency level is calculated based on the spatial comparison with surrounding features.

In this paper, we focuses on the temporal aspect of saliency, which is evaluated based on temporal comparison in short-term and long-term memory of an agent, and propose a lexical acquisition model in which saliency evaluated based on a robot’s experience affects to the visual attention and learning rate of a robot. A robot evaluates saliency for each visual feature of observed objects depending on habituation and learning experience. The curiosity based on the evaluated saliency affects to the selection of objects to be attended and changes the learning rate for lexical acquisition.

The rest of the paper is organized as follows: the next section introduces the proposed lexical acquisition model based on curiosity, and the previous simulation experiment to show the efficiency of the proposed learning model is summarized. Then, the experiment with a real humanoid is presented with

the adaptively representation model of visual features. Finally, discussion and conclusion are given.

II. Lexicon acquisition learning based on curiosity

The proposed system learns lexicons on shapes and colors of an observed object through communication with a caregiver. When a robot attends to an object, it acquires the visual features on shapes and colors through visual sensors. At the same time, a caregiver teaches a label: a word that corresponds to the visual feature of the observed object. Here, it is supposed that labels given to the robot are independent to each other (exclusive relation among them).

A. learning system

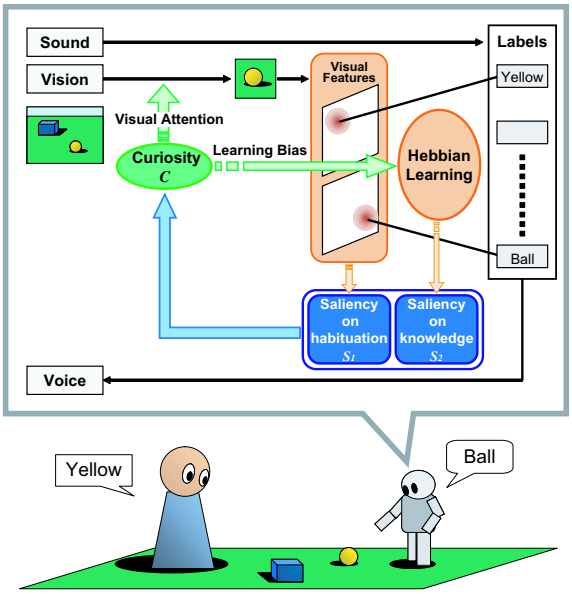


Fig. 1. An overview of the proposed system for acquiring lexicon utilizing curiosity

Fig. 1 shows the proposed system. The robot learns the lexicon in the following way.

1) The robot selects one salient object among many in its field of view.
2) The robot points the selected object and utters the labels from its own knowledge corresponding to the visual features of the selected object, so that the caregiver can be informed what is known label or unknown one.
3) The robot associates the visual features of the object with the label uttered by the caregiver.

In this learning process, curiosity that the robot feels has effects on the selection of the object to be attended and on the learning rate of the association between the label and the visual features. In the following, curiosity based on saliency is formulated, and the effects of curiosity on attention and learning are given.

B. curiosity based on saliency

The curiosity that a robot feels on the visual features consists of two kinds of salinecy: the habituation saliency, S_1, and the knowledge saliency, S_2.

1) the habituation saliency, S_1: The first saliency, S_1, is characterized by habituation. The robot feels low saliency for the visual feature that is always observed. On the other hand, it feels high for the feature that is observed for the first time or that is not observed for a long time. To realize this feature, the habituation saliency is updated as follows,

$$S_1^i(t) = S_1^i(t-1) + \Delta S_1^i(t-1) \tag{1}$$

$$\Delta S_1^i(t) = \frac{\alpha\left(1 - S_1^i(t)\right) - \beta\, S_1^i(t)\, I^i(t)}{\tau} \tag{2}$$

where α is a constant that characterizes spontaneous recovery, β is a constant that determines the rate of habituation, τ is the time constant, and I^i is the activation level of the i-th neuron in the visual feature map.

2) the knowledge-driven saliency, S_2: The second saliency, S_2, is characterized by the acquired knowledge. The robot feels more salient for the visual feature that is not associated with any other label than learned one. This saliency is expected to accelerate the lexicon learning by suppressing the association between a label and the visual feature that has already been associated with another label. The acquired lexical knowledge is represented as the connection strength, w, between a visual feature and a label. Let the connection strength from the l-th label to the i-th visual feature neuron $w_{l\to i}$, The label that connects to the i-th visual feature neuron with the maximum stlength, L, is

$$L = \arg\max_l(w_{l\to i}). \tag{3}$$

The connection, $w_{L\to i}$, can be used as the index of the familiarity of the i-th visual feature neuron.

$$S_2^i = 1 - sigmoid(w_{L\to i}), \tag{4}$$

$$sigmoid(w) = \frac{1}{1+e^{-a(w-\theta)}}, \tag{5}$$

where a is the parameter that determines the rate of rise of sigmoid function, and θ is a threshold.

3) curiosity: The curiosity level that the robot feels for the i-th visual feature can be calculated by the product of the two saliency as follows,

$$C^i(t) = \left(S_1^i(t) + c_1\right) \times \left(S_2^i(t) + c_2\right), \tag{6}$$

where c_1 and c_2 are constants.

C. Attention bias

The robot selects one object to be attended among the observed ones based on the curiosity level. The curiosity level for the n-th object is evaluated by the maximum value of the product of the activated level I and the curiosity level C of each observed visual feature,

$$M_n = \max_i(\, I_n^i \times C^i\,). \tag{7}$$

The robot attends to the object that has the maximum M value,

$$N = \arg\max_n M_n. \tag{8}$$

However, it is supposed that when M_n does not exceed the minimum threshold, the robot does not show any interest to the observed objects and searches for another one.

D. Learning bias

A visual feature is associated with a label based on Hebbian learning. When the caregiver teaches the robot the label l the activated neuron corresponding to the visual feature that the robot observes at that time is associated with this label l. The learning is biased by the curiosity defined previously. The more salient the visual feature is, the more strongly the connection with the label is bound. Let the activation level of the l-th label a_l, then the update equation is given by

$$\Delta w_{l\to i} = \epsilon\, a_l \left(I^i - \text{threshold} \right) C^i, \tag{9}$$

where ϵ denotes the learning rate. When the k-th label is taught by the caregiver, $a_{l=k} = 1$ and $a_{l\neq k} = 0$. The learning rate ϵ is biased by the second saliency S_2 as follows,

$$\epsilon = cS_2^n, \tag{10}$$

$$n = \arg\max_i (w_{l\to i}), \tag{11}$$

where c is a constant. When the l-th label is already connected to the n-th visual feature, the second saliency S_2^n becomes small, and it is expected that the connection with another visual feature is suppressed.

When the l-th label uttered by the robot is wrong and corrected by the caregiver, the corresponding connection is weakened by the following update equation,

$$\Delta w_{l\to i} = -\epsilon'\, I^i, \tag{12}$$

where ϵ' is a constant learning rate.

When the visual feature $\boldsymbol{I}$ is observed, the robot utters the l-th label, if the utterance value a_l which is defined as

$$a_l = \sum_i I^i\, w'_{l\to i}, \tag{13}$$

exceeds a certain threshold.

E. Simulation Experiment

Before applying to a real robot, we examined the proposed system in the simulation environment. The detail setting and the results of the simulation are described in the previous paper [13]. Here, only the points of the results are summarized.

In the simulation experiment, the task of a robot is to learn the name (labels) for the corresponding visual features. The assumed visual features that the robot can detect are 5 types: color, shape, size, weight, and hardness. The variations of objects are 40 for color, 80 for shape, and 8 for size, weight and hardness. The robot has the 144 neurons, each of which is activated when the corresponding visual feature is detected. 144 labels are taught to the robot by the caregiver. When the object to be attended is determined, the robot utters all labels that it has learned. Depending on the robot's utterance, the caregiver teaches labels in the following way.

- When the robot does not utters or utters without any error, the caregiver teaches only one label that is not uttered by the robot.
- When the robot's utterance is wrong, the caregiver points out that the utterance is wrong and teaches the right label.
- When the labels the robot utters are all right, the caregiver does nothing.

Fig. 2 shows the resultant learning curves averaged in 10 trials. The horizontal and vertical axes indicate the learning steps and the number of acquired vocabulary, respectively. One learning step is defined as the process in which the caregiver teaches the robot one label and the robot updates its Hebbian learning network. The number of acquired vocabulary is calculated as the sum of the connection weights that exceeds 0.9. One trial is defined as the process that the robot learns all labels. This graph shows the proposed method (purple curve) effectively learns the labels: 25 % less steps than that of simple Hebbian learning (red curve) in the incremental environment.

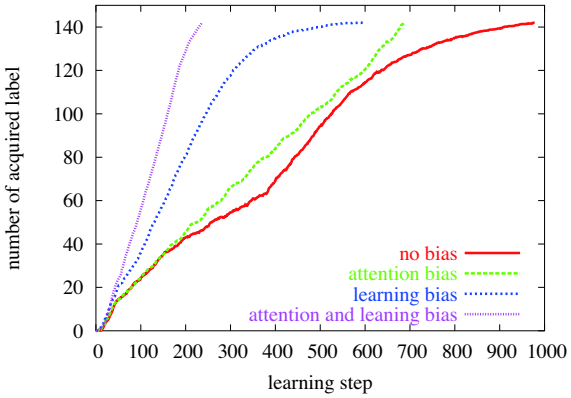


Fig. 2. Learning curves

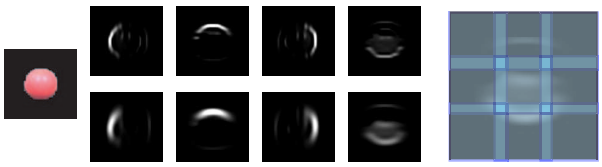
The active selection of an object to be attended seems to affect the learning speed in the late stage of the learning process. On the other hand, the learning bias helps the robot to acquire the more proper associations with less number of teaching.

III. Experiment with a Real Humanoid

A. visual features

In this experiment, the visual features of color and shape are extracted as visual inputs. The visual feature for color is the averaged value of the color of an observed object. The visual feature for shape is the output of Gabor filters [14]. Fig. 6 (a) shows an example of output of the Gabor filters, in which the parameters of the Gabor filters are $r = (4, 8)$, and $\theta = (0, 60, 120, 180, 240, 300)$. 3×3 receptive fields are arranged in the image as shown in Fig. 6 (b). The values of the Gabor filters are summed in each receptive field with the weight of the distance between the each pixel position and

the center of each receptive field. Thus the shape feature is represented in $9 \times 2 \times 6 = 108$ dimensions.



(a) Examples of Gabor filter (r=10,20 θ=0,90,180,270)

(b) Receptive Field

Fig. 3. Primal features about shape

The visual features such as shape and color are represented with self organizing map (SOM) [15]. In this SOM, the representational vectors are self organized based on the input visual data, and after learning the activated levels of neurons are used as a feature vector I for the system.

The activation level of each neuron in SOM is calculated depending on the distance between the input vector and the representational vector of the neuron.

$$a = 1 - \frac{1}{1 + \exp^{-\gamma(d - \text{threshold})}}, \quad (14)$$

where d denotes the distance between the input vector and the representational vector (Fig. 4). Let the activated level of the i-th neuron be a^i $(0 \leq a^i \leq 1)$, then the visual feature vector I is described as

$$\boldsymbol{I} = \begin{pmatrix} a_c^0 \\ \vdots \\ a_c^{n_c} \\ a_s^0 \\ \vdots \\ a_s^{n_s} \end{pmatrix} \quad (15)$$

where n_c and n_s are the numbers of neurons of the color and shape SOM, respectively.

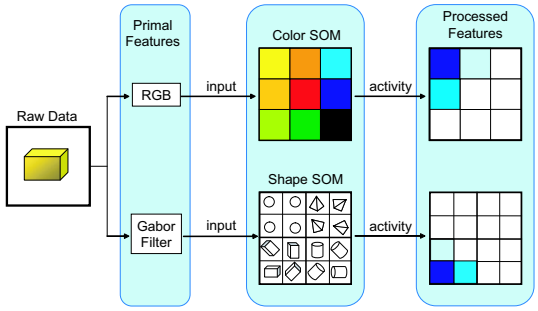


Fig. 4. The flow of feature extraction

B. representation of shapes independent of view points

Unlike in the simulation environment, the shape appearence in the real environment changes depending on the robot's viewpoint. The simple mapping between neurons of the visual feature SOM and labels cannot be used. We assume that the the visual features observed continuously belong to the same class of the shape. The proposed shape map consists of small SOM segments, to each of which given labels are mapped. The activation level of each SOM segment is the maximum activation level of the neuron belonging to it. A new SOM segment is added adaptively when the existing SOMs cannot represent an observed shape. Moreover, when multiple SOMs are activated during the interaction with one object, those SOMs are merged. In this case, the connection weights of two SOMs $w_{l \to i}(t)$, $w_{l \to j}(t)$, are merged to new one $w_{l \to i}(t+1)$ as follows,

$$w_{l \to i}(t+1) = w_{l \to i}(t) + w_{l \to j}(t), \quad (16)$$
$$w_{l \to j}(t+1) = 0. \quad (17)$$

The reconstruction process of shape map is described as follows,

1) When an object to be attended is decided, the robot starts to interact with the object so that the data of the various appearances of the object can be collected.
2) After the interaction with the object, the collected data are used for incremental learning of the SOM that is activated during the interaction.
3) If no existing SOMs are activated, a new SOM is added to the shape map.
4) If multiple SOMs are activated, these SOMs are merged because it is highly likely that they represent the different appearances of the same shape.

Here, in the reconstruction such as incremental learning or merging, the new SOM is constructed using the training data which consist of the representational vectors of the old SOMs and new input data.

C. acquiring lexicon by a humanoid robot

To evaluate the validity of the proposed method, the lexicon acquisition experiment is done in the real environment using a humanoid robot. Fig. 5 shows the environmental setting. The humanoid robot, Hoap2 [16], enhanced with the USB camera, has motions like moving its neck for searching objects, walking for approaching to them, and kicking for interacting with objects, and pointing the objects by the arms to mention the caregiver which object the robot attends to. Instead of uttering real sound, the robot communicates with the caregiver via displayed word and keyboard inputs.

The process of the interaction with the objects and communication with the caregiver is as follows. Firstly, the robot selects the object to be attended based on its curiosity level, and approaches to the attended object. The robot interacts with the object by kicking. During approaching and kicking, the appearance data are collected and used for the reconstruction of the shape map. After one kicking, the robot points to the object, and indicates the known labels about the pointed object to the caregiver in the computer display. If the indicated labels are incorrect, the caregiver correct them and teaches the

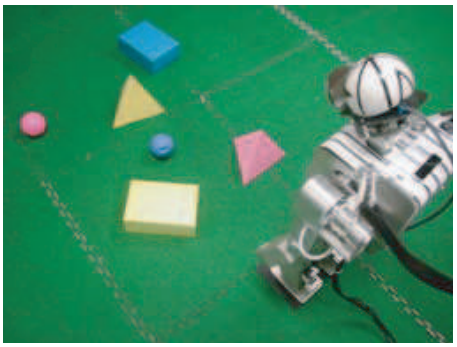


Fig. 5. Environmental setting

correct ones. If the indicated labels are correct or no labels are indicated, the caregiver teaches the new one of the pointed object. After labels are taught, the saliency level S_1 is updated according to the eqs. (1) and (2). If the saliency levels S_1 on all the features of the attended object go down less than a certain threshold, the robot changes its attention to a new object.

Fig. 6 shows the saliency changes and the activity levels of color and shape SOMs in the first several learning steps. The upper table shows the observed images of the objects (the last images in the interaction), the labels uttered by the robot, and the labels taught by the caregiver. The graph in the middle shows the time course of the curiosity and saliency for the shape SOM1, as well as the connection weight w between the label "Box" and the corresponding shape SOM (Shape SOM1). The tables in the bottom shows the activation level of the color SOM and the shape SOMs (the activation level at the end of the interaction is indicated). Followings are the details of this process.

1) In the first step of the learning (from 0 to 25 [sec]), the first shape SOM is composed based on the data that are collected during the interaction. After the label "Box" is taught by the caregiver, the robot gives its attention to the current object until the salient level S_1 goes down the threshold (=0.3), and changes its attention to a new object. At this stage, the label "Box" is connected both to the neuron in color SOM corresponding to yellow and to the shape SOM1.
2) In the second step, the robot attends to a new object. The shape SOM1 is not activated and the new shape SOM2 is added. At the same time, the salient level S_1 for the shape SOM1 is recovered. At this stage, the label "Ball" is connected to both to the neuron in color SOM corresponding to blue and to the shape SOM2.
3) In the third step, the robot attends to the object that has the same shape as the first step. However, the robot cannot recognize as the same because the viewpoint is different. The neuron in color SOM corresponding to blue is activated and the label connected to the neuron "Ball" is uttered. The caregiver corrects its mistake and the connection between the blue neuron and the label "Ball" is weakened. At this stage, the label "Ball" is connected to the correct shape SOM, and the label "Box" is connected to the neurons in color SOM corresponding to yellow and blue, and the shape SOMs 1 and 3.
4) In the forth step, the robot attends to the object that is the same as the first step. During the interaction with the object, the shape SOMs 1 and 3 are activated and so they are merged, so that the connection weight goes up and the saliency level about knowledge S_2 decreases.
5) In the fifth step, the robot attends to the object, that is, a yellow ball. The label "Box" is uttered because it is still connected to the yellow neuron. The caregiver corrects its mistake, then the label "Box" is correctly connected to the corresponding shape SOM1. This causes the decrease of the saliency level about knowledge S_2". Afterwards, the shape SOM1 is not easily connected to other labels.

Fig. 7 shows the learning curve in the learning process.

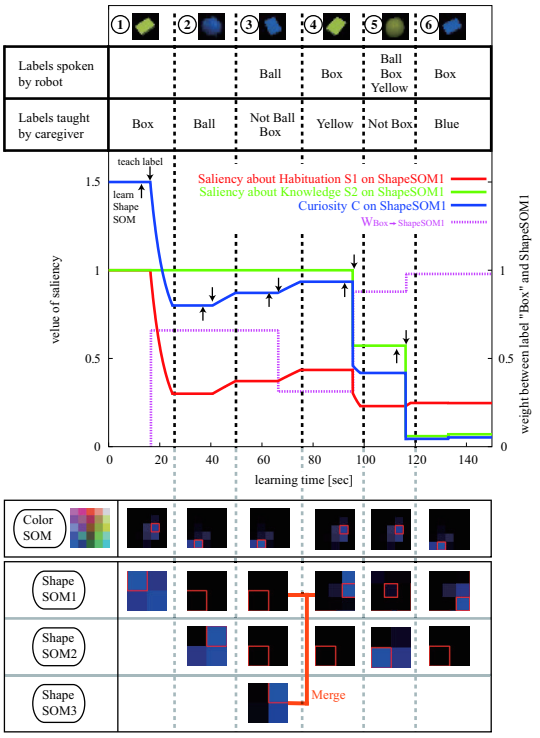
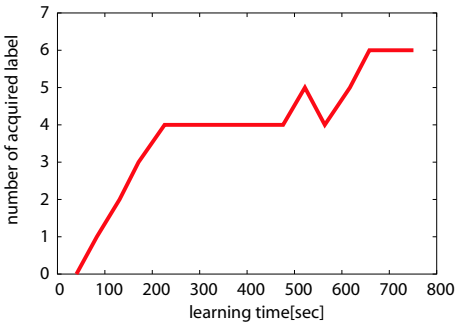


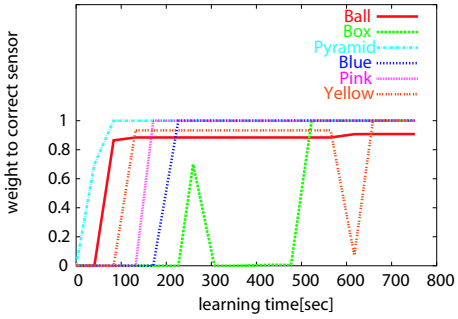
Fig. 6. Saliency changes and activated labels in shape SOMs at the first several steps

IV. Discussion and Conclusion

This paper proposed the lexicon acquisition model in which the robot can effectively associate the observed visual feature with the spoken labels based on curiosity. The curiosity level calculated based on the saliency levels from habituation and the acquired knowledge has effects on the visual attention and the learning rate of the robot. The simulation result shows that the learning model with curiosity acquires the given labels much faster than the simple Hebbian learning model. Moreover, the proposed learning model shows better



(a) Number of acqured label



(b) Weight changes of each label

Fig. 7. Experimental result using real robot

performance in the environment in which the number of objects exposed to robots is gradually increased.

Even if an agent and a caregiver shares joint attention to one object, the agent cannot associate the visual feature with the word uttered by the caregiver without understanding which feature the uttered word is intended to. The proposed learning model solved this problem by associating the uttered label with the unlearned feature more effectively based on the curiosity. This is thought to be one formulation of the mutually exclusive constraint proposed by Markman [7]. The mutually exclusive constraint will be more effective if the robot preferentially selects the object whose features are only partially known. This preference is not included in this paper. The robot feels equal curiosity to the objects which has any new feature. However, the effectiveness of the preference to partially known objects is shown in the simulation results in the environment in which the number of objects exposed to robots is gradually increased.

The proposed method implicitly supposes the joint attention mechanism between the robot and the caregiver. The coincidence of the curiosity, and so the coincidence of the attention, between the robot and the caregiver is thought to be very important in language acquisition. In this paper, the curiosity model is adopted only in the learner.

Exploring the learning model which the learner and the caregiver share the same saliency model is our future work. Sharing the same saliency will not be difficult when the learner and the caregiver shares the same saliency model. This may be possible by learning saliency model each other as if human infants and caregivers do.

ACKNOWLEDGMENT

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (A), 16200012, 2006.

REFERENCES

[1] K. D. Pruett, *Me, Myself and I: How Children Build Their Sense of Self —18 to 36 Months*. Goddard Press, Inc., 1999.

[2] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," in *Proceeding of the 1st IEEE/RSJ International Conference on Humanoid Robots*, 2000, p. CDROM.

[3] H. Asoh, S. Akaho, O. Hasegawa, T. Yoshimura, and S. Hayamizu, "Intermodal learning of multimodal interaction systems," in *In Proceedings of the International Workshop on Human Interface Technology*, 1997.

[4] K. Ishiguro, N. Otsu, and Y. Kuniyoshi, "Inter-modal learnig and object concept acquisition," in *In Proceedings of IAPR Conference on Machine Vision Applications (MVA2005)*, 2005, p. CDROM.

[5] L. Steels and F. Kaplan, "Aibo's first words. the social learning of language and meaning," *Evolution of Communication*, 2001.

[6] N. Iwahashi, "Language acquisition through a human-robot interface by combining speech, visual, and behavioral information," *Information Sciences*, vol. 156, pp. pp. 109–121, 2003.

[7] E. M. Markman, "Categorization in children: Problems of induction," *Cambridge, MA: MIT Press, Bradford Books*, 1989.

[8] B. Landau, L. B. Smith, and S. Jones, "The importance of shape in early lexical learning," *Cognitive Development*, vol. 3, pp. 299–321, 1988.

[9] D. K. Roy, "Learning visually-grounded words and syntax for a scene description task," *Computr Speech and Language*, vol. 16, pp. 353–385, 2002.

[10] A. Toyomura and T. Omori, "A computational model for taxonomy-based word learning inspired by infant developmental word acquisition," *IEICE Information and Systems*, vol. 88, no. 10, pp. 2389–2398, 2005.

[11] P. Rochat, *The Infant's World*. Harvard Univ Press, 2004.

[12] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Computer Vision and Image Understanding*, vol. 100, pp. 41–63, 2005.

[13] M. Ogino, M. Kikuchi, and M. Asada, "Active lexicon acquisition based on curiosity," in *Proceedings of the Fifth International Conference on Development and Learning*, 2006, p. CDROM.

[14] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, pp. 1233–1258, 1987.

[15] T. Kohonen, "Self-organizing maps," *Springer-Verlag Verlin Heidelverg*, 1995.

[16] Y. Murase, Y. Yasukawa, K. Sakai, and etc., "Design of a compact humanoid robot as a platform," in *In Proceedings of the 19th Conference of Robotics Society of Japan*, 2001, pp. 789–790.