Active lexicon acquisition based on curiosity

Masaki Ogino, Masaaki Kikuchi and †Minoru Asada
Dept. of Adaptive Machine Systems, †Handai FRC,
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*— It is observed that human infants 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. It is thought that the infants utilize various kinds of cognitive biases for efficient learning. This paper proposes a lexical acquisition model which makes use of curiosity to associate visual features of observed objects with the labels that is uttered by a caregiver. This model is applied to a virtual robot. The degree of curiosity that the robot feels to the objects is determined by the two kinds of saliency; habituation saliency and knowledge one. The former saliency is related to the habituation and the latter one is related to the strength of the association between the visual features and the labels. A robot changes its attention and learning rate based on curiosity. Simulation experiments show that the learning model with curiosity effectively associate the labels with the observed visual features.**

***Index Terms*— lexicon acquisition, curiosity, saliency, SOM, Hebbian learning**

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 [7]. 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 [1].

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 [2] [3] [10] [4]. 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 set 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 [6] proposed the whole object constraint and the mutual exclusivity constraint. Landau et al. [5] 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] [11].

Moreover, infants are not passive creatures. They actively and intentionally interact with the environment around them [8]. 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 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.

In the following, the next section introduces the proposed lexical acquisition model based on curiosity. Then the simulation experiment to show the efficiency of the proposed learning model is described. Finally, discussion and conclusion is 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

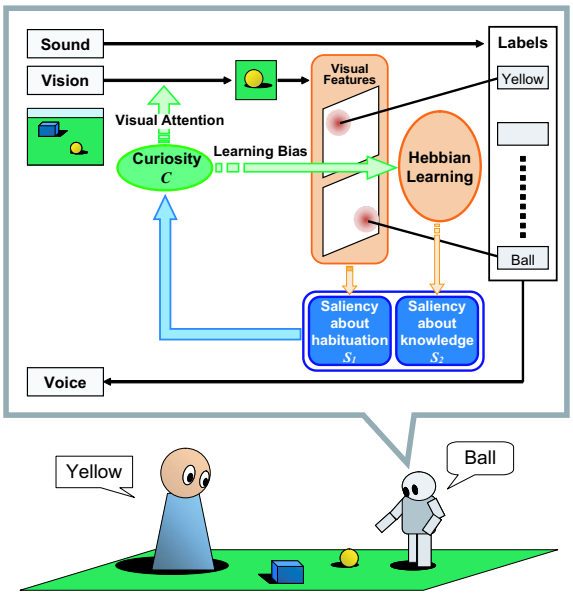


Fig. 1. An overview of the proposed system for acquiring lexicon utilizing curiosity

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 saliency: 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 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 length, 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 the 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,

$$\epsilon = c S_2^n, \tag{10}$$

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

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

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

exceeds a certain threshold.

III. Simulation Experiment

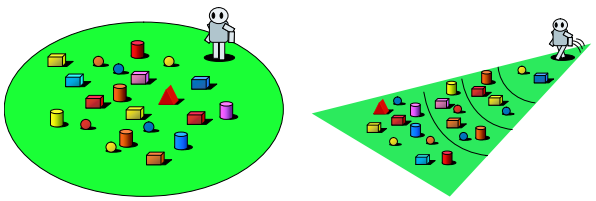
A. Experimental conditions

The effectiveness of the proposed system is examined in simulation experiments.

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 he 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 do nothing.

To examine the effectiveness of the proposed system, the learning performance under various conditions are compared; with and without the learning bias in Hebbian learning, and with and without the selection of an object to be attended. Another condition is the environmental setting; uniform environment in which all objects are uniformly distributed (Fig. 2 (a)) and incremental environment in which observable objects gradually increases (Fig. 2 (b)). Here, the gradual environment means that the robot encounters the new visual variety of the visual features gradually. In the first step, the presented objects in the environment have 3 variations in each visual feature. When the robot does not feel salient in the presented objects, it searches new objects with new variations of visual features. In this simulation experiment, the robot encounters n varieties of visual features in the n-th searching step.



(a) Uniform environment (b) Incremental environment

Fig. 2. Experimental environments

The examined conditions in the simulations are as follows,

1. **with no bias, uniform env.**
2. **with learning bias, uniform env.**
3. **with attention bias, uniform env.**
4. **with attention and learning bias, uniform env.**
5. **with no bias, incremental env.**
6. **with learning bias, incremental env.**
7. **with attention bias, incremental env.**

8. **with attention and learning bias, incremental env.**

In the conditions without learning bias (conditions 1, 3, 5, 7), the connection weights are updated by the following equation, instead of eq. (9),

$$\Delta w_{l \to i} = \epsilon \, a_l \left(I^i(t) - \text{threshold} \right). \tag{14}$$

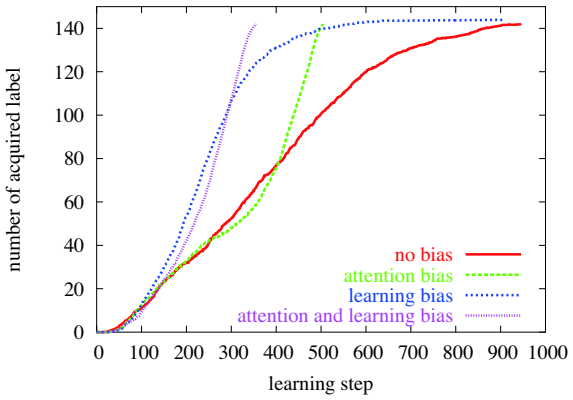
In the conditions with attention bias (conditions 3, 4, 7, 8), four objects are presented to the robot at each learning step. The robot attends to the most salient object among the presented ones. If the curiosity levels of all presented objects are lower than certain threshold, the robot changes its attention to other objects with new variations of the visual features in the environment.

B. results

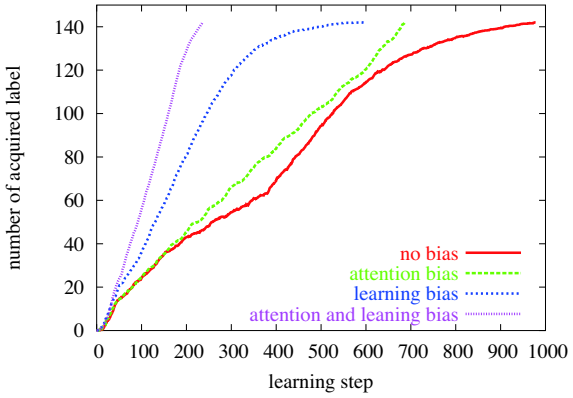
Fig. 3 shows the resultant learning curves averaged in 10 trials. The horizontal axis indicates the learning steps and the vertical axis indicates the number of acquired vocabulary. One learning step is defined as the process in which the caregiver teaches the robot one label and the robot updates its Hebbian learning network. The 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: 35 % less steps than that of simple Hebbian learning (red curve) in the uniformly distribution environment, 25 % less steps than that of simple Hebbian learning (red curve) in the incremental environment.

Figs. 4 and 5 show the time courses of the number of uttered words (red) and the errors (green) in each condition, respectively. The horizontal axis indicates the learning step and the vertical axis is the number of uttered words and the errors. Figs. 6 and 7 show the distribution of the number of learning steps that the robot takes to acquire one word in the respective conditions. The horizontal axis indicates the number of teaching by the caregiver and the vertical axis indicates the number of acquired words.

1) effectiveness of the attention bias: The active selection of an object to be attended affects the learning speed in the late stage of the learning process. The learning curves without attention bias (red and blue curves in Fig. 3) show slow convergence in the late stage of the learning process. This is because objects are randomly presented to the robot regardless whether the visual features included in presented objects are learned or not. It is also indicated in Fig. 4 (a) and Fig. 5 (c) which show that in almost steps the robot utters the right labels. On the other hand, the attention bias effectively accelerates the learning speed in the last half of the learning steps (red and green curves in Figs. 3 (a, b)), nevertheless the the number of learning process that the robot needs to acquire one association is almost same between the learning processes with and without attention bias (Figs. 6 (a, b) and Figs. 7 (a, b)).



(a) Result of uniform environment

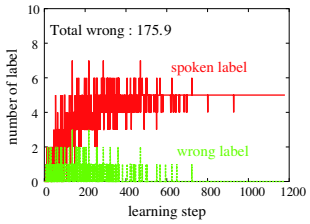


(b) Result of incremental environment

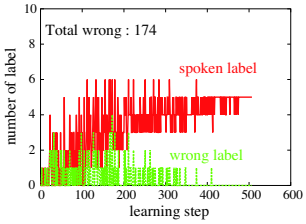
Fig. 3. Learning curves of respective learning conditions in each environment

2) effectiveness of the learning bias: Figs. 4 and 5 show that the frequency of errors in the late stage of the learning process is much less in the conditions with learning bias (Figs. 4 (c, d) and Figs. 5 (c, d)) than without learning bias (Figs. 4 (a, b) and Figs. 5. (a, b)). It is also shown in Figs. 6 and 7 that the teaching number per one association are less in the conditions with learning bias. These results indicate that the learning bias helps the robot to acquire the more proper associations with less number of teaching.

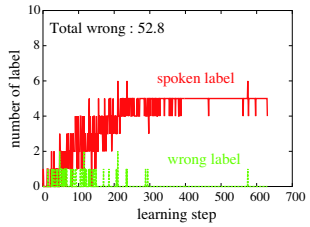
3) effectiveness of the learning environment: Fig. 3 shows that the learning rate in the early stage of the learning process is faster in the gradually increased environment than in the uniformly distributed environment. Fig. 5 show that In the gradually increased environment the robot utters many labels from the early stage of learning. This is because the probability that the robot encounters an object that has the limited number of new features is higher, so that the robot



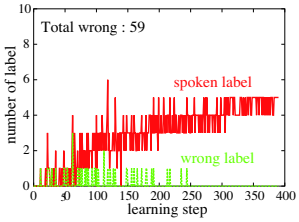
(a) no bias



(b) attention bias

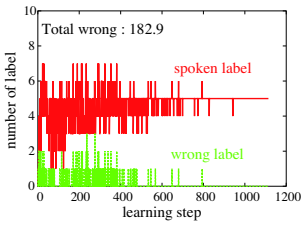


(c) learning bias

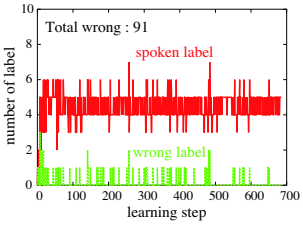


(d) attention and learning bias

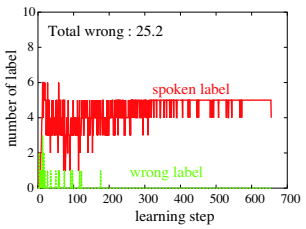
Fig. 4. Number of spoken labels and wrong labels in uniform environment



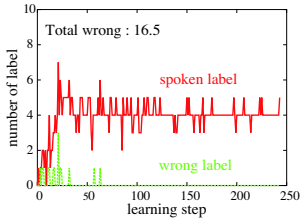
(a) no bias



(b) attention bias



(c) learning bias



(d) attention and learning bias

Fig. 5. Number of spoken labels and wrong labels in incremental environment

can effectively learn the association with the learning bias by attending to the new feature. Especially with the learning bias the robot can learn about 70 % of words (100 words) by only one teaching (Figs. 7 (c) and (d)). On the other hand, Figs. 7 (a) and (b) show that the robot without the learning bias fails to attend to the new feature and takes much more time to learn the associations (nevertheless, it is more effective than in the uniformly distributed environment (Fig. 6) thanks to the gradual learning.).

IV. Discussion and Conclusion

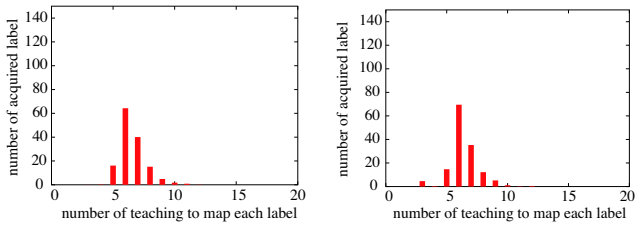
This paper proposed the lexicon acquisition model in which the robot can effectively associate the observed visual feature with the spoken labels based on curiosity. The curiosity level calculated based on the saliency levels based on habituation and the acquired knowledge have effects on the visual 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 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 mutual exclusively constraint proposed by Markman [6]. The mutual exclusively 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.

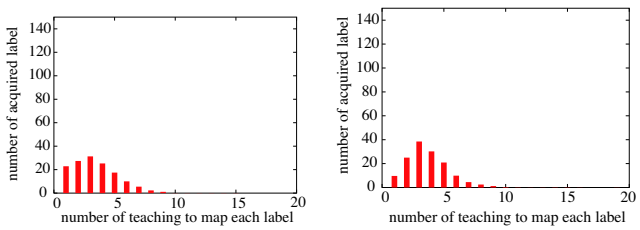
The implicit point in the proposed method is the joint attention between the robot and the caregiver. The coincidence of the curiosity, and so the coincidence of the attention, between the robot and the caregiver is thought to be very important in language acquisition. In this paper, the curiosity model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

Exploring the learning model in which the learner and the caregiver share the same saliency model is next challenge. 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.

The application of this method to a real robot is also the next challenging problem. As well as the speech recognition, we are now trying to develop a visual recognition system with which a robot can represent various features using self organizing maps. Especially, the representation of the shape information independent of viewpoints is important in the real environments. Moreover, it is necessary to consider to

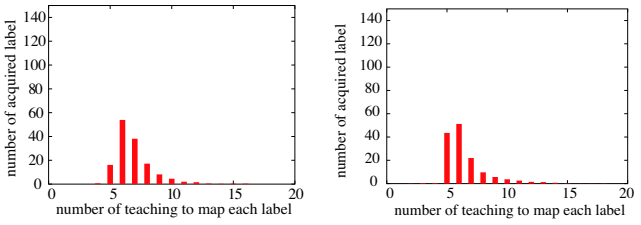


(a) no bias (b) attention bias

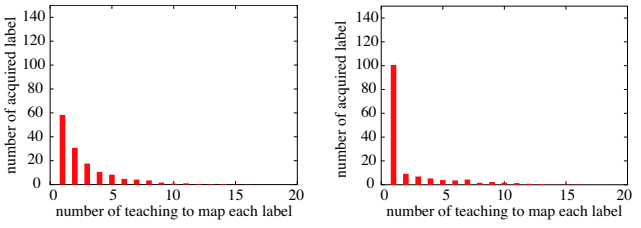


(c) learning bias (d) attention and learning bias

Fig. 6. Number of teaching to map each label in uniform environment



(a) no bias (b) attention bias



(c) learning bias (d) attention and learning bias

Fig. 7. Number of teaching to map each label in incremental environment

combine other constraints such as grammatical information with the proposed learning model.

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] Asada, M., MacDorman, K. F., Ishiguro, H. and Kuniyoshi, Y.: 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.

[2] Asoh, H., Akaho, S., Hasegawa, O., Yoshimura, T. and Hayamizu, S.: Intermodal Learning of Multimodal Interaction Systems, in *In Proceedings of the International Workshop on Human Interface Technology*, 1997.

[3] Ishiguro, K., Otsu, N. and Kuniyoshi, Y.: Inter-modal Learnig and Object Concept Acquisition, in *In Proceedings of IAPR Conference on Machine Vision Applications (MVA2005)*, 2005.

[4] Iwahashi, N.: Language acquisition through a human-robot interface by combining speech, visual, and behavioral information, *Information Sciences*, Vol. 156 (2003), pp. 109–121.

[5] Landau, B., Smith, L. B. and Jones, S.: The importance of shape in early lexical learning, *Cognitive Development*, Vol. 3 (1988), 299–321.

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

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

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

[9] Roy, D. K.: Learning Visually-Grounded Words and Syntax for a Scene Description Task, *Computr Speech and Language*, Vol. 16 (2002), 353–385.

[10] Steels, L. and Kaplan, F.: AIBO's first words. The social learning of language and meaning, *Evolution of Communication*, (2001).

[11] Toyomura, A. and Omori, T.: A Computational Model for Taxonomy-Based Word Learning Inspired by Infant Developmental Word Acquisition, *IEICE Information and Systems*, Vol. 88 (2005), 2389–2398.

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