2005 IEEE/RSJ International Conference on Intelligent Robots and Systems

Extracting Multi-Modal Dynamics of Objects using RNNPB

Tetsuya Ogata[†], Hayato Ohba[†], Jun Tani[‡], Kazunori Komatani[†], and Hiroshi G. Okuno[†]

[†] *Graduate School of Informatics, Kyoto University, Kyoto, Japan*
{ogata, hayato, komatani, okuno}@kuis.kyoto-u.ac.jp
[‡] *Brain Science Institute, RIKEN, Saitama, Japan*
tani@brain.riken.jp

Abstract **- Dynamic features play an important role in recognizing objects that have similar static features in colors and or shapes. This paper focuses on active sensing that exploits dynamic feature of an object. An extended version of the robot, Robovie-IIs, moves an object by its arm to obtain its dynamic features. Its issue is how to extract symbols from various kinds of temporal states of the object. We use the *recurrent neural network with parametric bias* (RNNPB) that generates self-organized nodes in the parametric bias space. The RNNPB with 42 neurons was trained with the data of sounds, trajectories, and tactile sensors generated while the robot was moving/hitting an object with its own arm. The clusters of 20 kinds of objects were successfully self-organized. The experiments with unknown (not trained) objects demonstrated that our method configured them in the PB space appropriately, which proves its *generalization* capability.**

Index Terms - Active Sensing, Humanoid Robot, Recurrent Neural Network

I. Introduction

Our final goal is to develop techniques to enable robots to manipulate tools designed for humans. Conventional robots only manipulate specific tools designed for robot hands. It is still quite difficult for mechanical systems to handle the dynamics of objects and generate adaptive behaviors through the learning of a dynamic environment.

A crucial problem for such tool manipulation is object recognition and there have been some studies concerning "active sensing" [1] to solve this problem [2][3][4]. Noda et al. reported a study using the humanoid robot, Wamoeba-2Ri, which grasps objects with its hand and recognizes them by integrating multiple sensory data: size, weight, and color images [2]. Since that study used a three-layered SOM (Self-Organizing Map [5]) which can only deal with static features and it required over a thousand neurons for processing multi-modal sensory data, it was quite difficult for the robot to apply the recognition results to its motion planning. Arsenio et al. focused on rhythmic motion as the dynamics of objects and merged the visual-audio sensory data to recognize them using the humanoid robot Cog [3]. Though that study showed cross-modal dynamics was essential for object recognition and manipulation, the target was only "rhythmic motion" generated not by the robot but by the human operators. Therefore, it was not enough for the robot to plan more general tool manipulations. The common problem of these studies is that their target was recognition of fewer than 10 objects designed and/or selected for the robot system.

In this paper, we propose a novel active-sensing method using the dynamics of objects. This method uses a recurrent neural net (RNN) trained using the multi-modal sensory data generated while a robot is moving/hitting objects. The RNN enables robots to use the dynamic features for various object-recognition and motion-prediction methods. Furthermore, the proposed method has generalization capability that can configure unknown (not trained) objects appropriately.

Section II introduces the recurrent neural network model as the learning method. Section III described the actual design of active sensing, such as, motion design, target objects, sensors, and configuration of the neural network. Section IV shows some experiments and the results of our proposed methods. Section V discusses the characteristics of our method and compares them with those of conventional recognition methods. Section VI concludes this paper and describes future work concerning motion generation.

II. Learning Algorithm

This section describes a method that enables robots to deal with dynamic features of sensory information during active sensing. It is well known that 'statistical techniques' represented by the hidden Markov model (HMM) can process time-sequence data efficiently. However, these methods require huge amounts of data for learning. It is quite laborious for real robot systems to carry out experiments for collecting such a lot of data due to the durability problem. Moreover, the HMM can deal with only "known" objects. This could be a fatal problem in the adaptability to the real dynamic environment. Therefore we tried to use a 'deterministic model' represented by an artificial neural net (ANN) technique to solve this problem. For example, it is well known that RNN can self-organize (acquire) contextual information [6].

We use the FF-model (forwarding forward model) proposed by Tani [7]. This model is also called the RNN with parametric bias (RNNPB) model. It articulates complex motion sequences into motion units, which are encoded as the limit cycling dynamics and/or the fixed-point dynamics of the RNN. We have already reported the study of human-robot interaction based on *quasi-symbols* acquired by the RNNBP [8].

A. RNNPB Model

The RNNPB model has the same architecture as the conventional Jordan-type RNN model [9] except for the PB nodes in the input layer. Unlike the other input nodes, these PB nodes take a constant value throughout each time sequence and are used to implement a mapping between fixed length values and time sequences. The network configuration of the RNNPB model is shown in Figure 1.

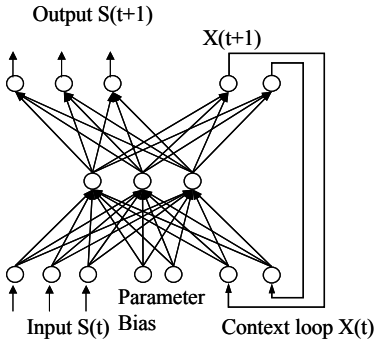


Fig. 1 Network Configuration of RNNPB

Like the Jordan-type RNN model, the RNNPB model learns data sequences in a supervised manner. The difference is that in the RNNPB model, the values that encode the sequences are self-organized in the PB nodes during the learning process. The common structural properties of the training data sequences are acquired as connection weights by using the back propagation through time (BPTT) algorithm [10], as also used in the conventional RNN. Meanwhile, the specific properties of each individual time sequence are simultaneously encoded as PB values. As a result, the RNNPB model self-organizes a mapping between the PB values and the time sequences.

B. Learning of PB Vectors

The learning algorithm for the PB vectors is a variant of the BPTT algorithm. The step length of a sequence is denoted by l. For each of the sensory-motor outputs, the back-propagated errors with respect to the PB nodes are accumulated and used to update the PB values. The update equations for the ith unit of the parametric bias at the t in the sequence are as follows.

$$\delta\rho_t = k_{bp} \cdot \sum_{t-l/2}^{t+l/2} \delta_t^{bp} + k_{nb}(\rho_{t+1} - 2\rho_t + \rho_{t-1}) \quad (1)$$

$$\Delta\rho_t = \varepsilon \cdot \delta\rho_t \quad (2)$$

$$p_t = sigmoid(\rho_t / \zeta) \quad (3)$$

In Eq. (1), the δ force for updating the internal values of the PB p_t is obtained from the summation of two terms. The first term represents the delta error, $\delta_t^{bp,}$ back propagated from the output nodes to the PB nodes: it is integrated over the period from the t–l/2 to the t+l/2 steps. Integrating the delta error prevents local fluctuations in the output errors from significantly affecting the temporal PB values. The second term is a low-pass filter that inhibits frequent rapid changes of the PB values. Internal value ρ_t is updated using the delta force, as shown in Eq. (2). And k_{bp}, k_{nb}, and ε are coefficients. Then, the current PB values are obtained from the sigmoidal outputs of the internal values. After learning the sequences, the RNNPB model can generate a sequence from the corresponding PB values.

Furthermore, the RNNPB model can be used for recognition processes as well as for sequence generation processes. For a given sequence, the corresponding PB value can be obtained by using the update rules for the PB values (Eqs. (1) to (3)), without updating the connection weight values. This inverse operation for generation is regarded as recognition.

The other important characteristic of the RNNPB model is that relational structure among the training sequences can be acquired in the PB space through the learning process. This generation capability enables the RNNPB model to generate and recognize unseen sequences without any additional learning. For instance, by learning several cyclic time sequences of different frequencies, it can generate novel time sequences of intermediate frequencies.

III. Active Sensing by Moving Objects

A. Addition of New Functions to Robovie-IIs

We refined the humanoid robot Robovie-IIs [11] as a platform of our experiments. Robovie-IIs itself is the refined model of Robovie-II developed at ATR [12]. The original Robovie-II has three DOF (degrees of freedom) on the neck and four DOF on each arm. It also has two CCD cameras on the head. The characteristic of Robovie-IIs is tactile sensors in soft silicon covering its whole body. The tactile sensor can discriminate three kinds of contact: *hit*, *rub*, and *touch* by detecting the pressure velocity.

Furthermore, we added some functions to Robovie-IIs: two external ears on the head and two 1-DOF hands on the arms for the experiment on "active sensing". Figure 2 shows a photograph of the head with external ears and the hand of our Robovie.

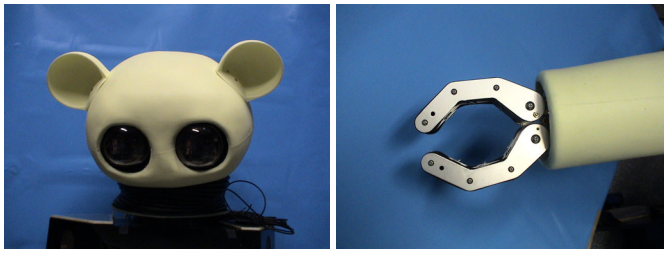


Fig. 2 The Head with External Ears and The Hand

B. Motion of Active Sensing and Target Objects

Perception with only static features such a visual image is not enough to discriminate objects that have similar sizes, shapes, and colors. Also such a recognition framework cannot be applied to dynamic motion planning. The perception should be designed in the sense of "sensory-motor coordination" [13].

We focused on active sensing motion that a robot is moving/hitting an object on the table with its own arm. Infants often touch and hit unknown object in front of them, and they acquire the skill to manipulate objects through such experiences. The motion of a moving object is the behavior for exploiting the dynamic features of the object, like tactile pressure to move it, actual trajectory, and sound pattern generated by collision with the table etc.

Figure 3 shows an actual experiment. Robovie touched and moved the object by rotating the shoulder motor (roll axis) with constant velocity (60 deg/s). While the robot was moving each object, the sound, object trajectory, and touch pressure were collected by its own microphones, cameras, and tactile sensors.

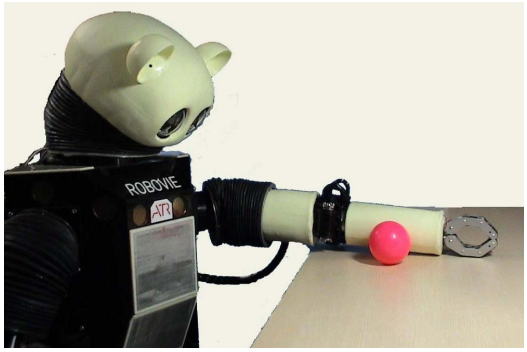


Fig. 3 Experiment of Active Sensing

In total, 20 kinds of objects were used as the recognition target as shown in Figure 4: *a rubber ball, plastic ball, ceramic cup, plastic cup (2 kinds), glass, can, moneybox, stuffed doll, Rubik's cube, toy-car, funnel, pen tray, scrub brush, soft brush, water dumbbell,* and *shampoo container.* In particular, the *water dumbbell* and *shampoo container* had two conditions, "full" and "empty". These two patterns cannot be discriminated just by the static features like visual images.

Fig. 4 Target Objects for Recognition

C. *RNNPB Configuration and Learning*

The following sensory data were normalized ([0-1]) and synchronized (9 frame/s) between different modalities for use by the RNNPB model.

1) Audio Information (5 units): The audio signal was detected by the microphones in the external ears (48 kHz). The five signal features were extracted using a Mel Filter Bank.

2) Visual Information (4 units): The center position (*x, y*) and the color (*R, B*) were detected by a CCD camera with resolution of 320 x 240 pixels (30 frame/s).

3) Tactile Information (1 unit): The input voltage from the skin tactile sensor was used (4.3 Hz).

The system diagram is shown in Figure 5. The designed RNNPB works as a prediction system whose input is current sensory data $s(t)$ and output is next sensory state $s(t+1)$. It consists of only 42 neurons: 10 neurons in the input layer, 20 neurons in middle layer, 10 neurons in context layer, and 2 neurons as parametric bias.

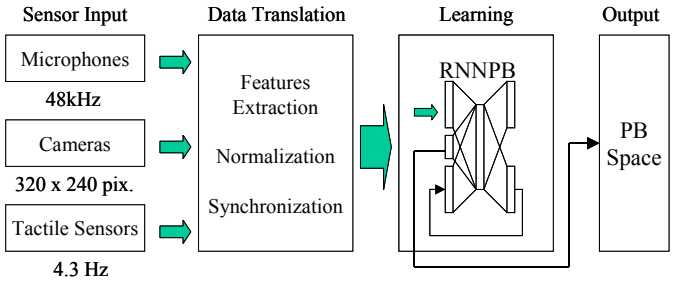


Fig. 5 System Diagram of Object Recognition

The training sequence of the RNNPB was segmented when the change values of all sensory input were less than the threshold. In the experiments, the sensory sequence lengths L_s were 15 to 40 steps.

Our goal is to acquire the specific parameter values corresponding to each object for recognition and motion generation. Therefore, in order to fix the parameter values during the sensing motion, Eq. (1) was simplified in our RNNPB model training as follows.

$$\delta\rho_t = k_{bp} \cdot \sum_{0}^{L_s} \delta_t^{bp} \qquad (4)$$

Also, Eq. (3) that normalizes the parameter values was not used in our experiments to make analysis of the acquired PB spaces ease.

IV. Experiments and Results

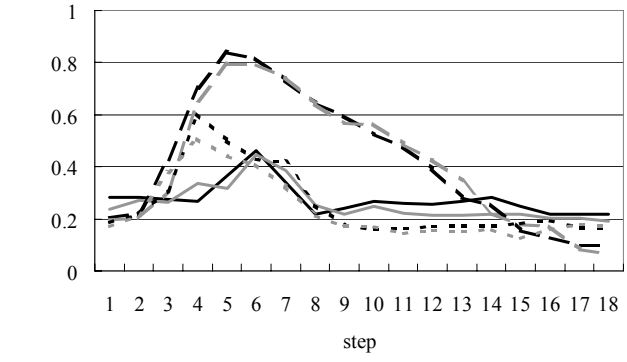
A. *Self-Organization of PB Space and Modality Differences*

We carried out the experiment using the 20 kinds of objects described in the previous section in order to confirm the clustering capability of the proposed method using object's dynamic features. Robovie moved these objects five times (20 x 5 = 100 sequence data), and the RNNPB was trained by a collecting data 100,000 times which required approx 1 hour using Pentium IV, 2.8 GHz.

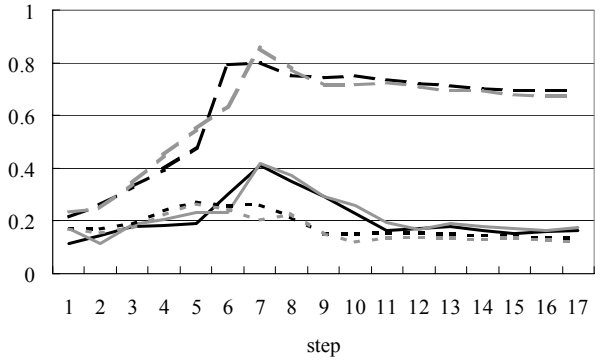
Figure 6 shows the sequences of the tactile pressure, the object position (x coordinate), and the sound power, when Robovie moved (a) *glass* and (b) *scrub brush.* The black lines describe the RNNPB input (real value) and the gray lines show the RNNPB output (prediction). We confirmed that the

RNNPB predicts each sequence well. The average prediction error is less than 1.5%.

Figure 7 shows the PB space acquired by each sensor modality. Two parametric values in the RNNPB before normalization correspond to the X-Y axes in the space. The characteristics of each space are as follows.



(a) An Example of Glass



(b) An Example of Scrub Brush

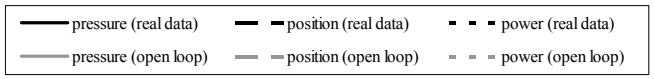


Fig. 6 Sensor Flow and Prediction output of the RNNPB

1) PB space acquired by tactile sensor: Figure 7-(a) shows the PB space when only tactile sensors were used. Though most objects were not categorized, there was a tendency for heavy objects to be mapped in the upper part in the space.

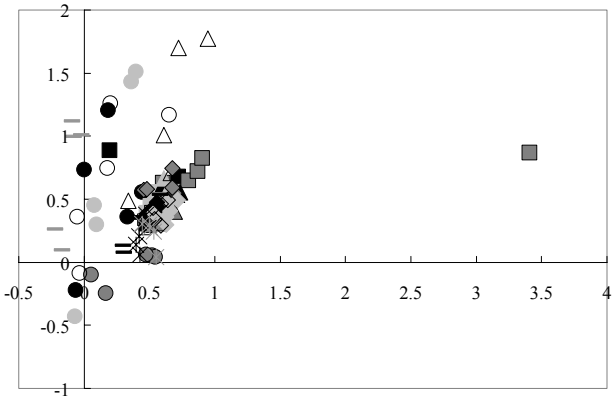
2) PB space acquired by sound signal: Figure 7-(b) shows the space when only sound signal was used. In this space the objects that did not make a sound were mapped in the right-upper area. Though some vague clusters can be seen, the sharpness of separation is quite low.

3) PB space acquired by visual data: Figure 7-(c) shows the space when only visual information was used. In this space, almost all objects could be separated. However, the objects with similar trajectories, such as "*can*/*glass*" and "*moneybox*/*pen tray*", were not separated.

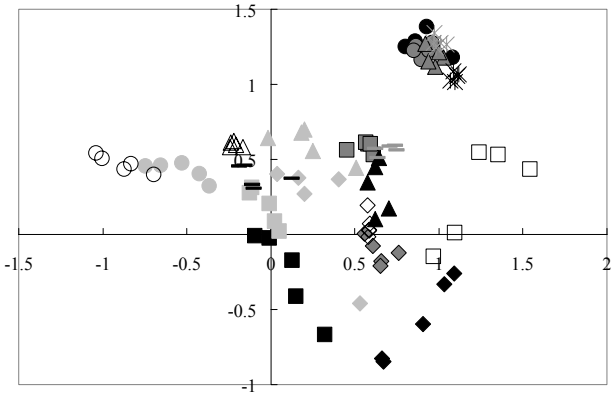
4) PB space acquired by all sensory modalities: Figure 7-(d) shows the PB space self-organized when all sensory modalities were used. We confirmed that the RNNPB could acquire the clusters for all kinds of objects.

B. Clustering of Unknown Objects

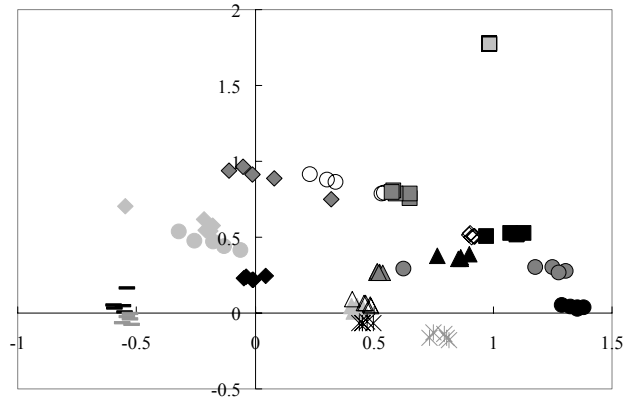
We carried out other experiments to confirm the ability of generalization of our method by recognizing unknown (not trained) objects. In this experiment, we used 8 objects: *a rubber ball, glass, moneybox, pen tray, scrub brush, soft brush,* and *shampoo container* (empty and full).



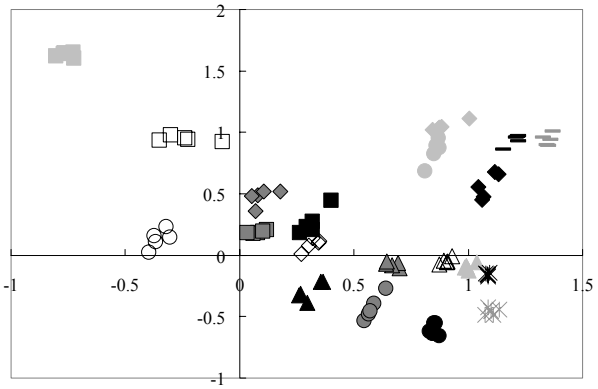
(a) PB Space Acquired by Tactile Sensor



(b) PB Space Acquired by Audio Signal



(c) PB Space Acquired by Visual Image



(d) PB Space Acquired by All Sensor Modality

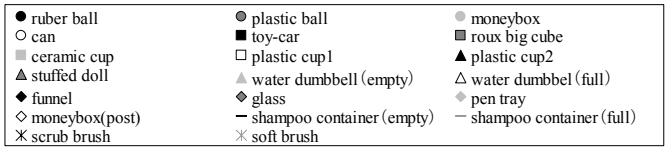
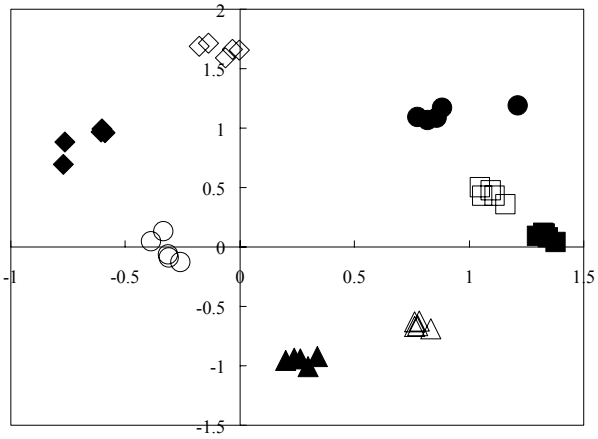


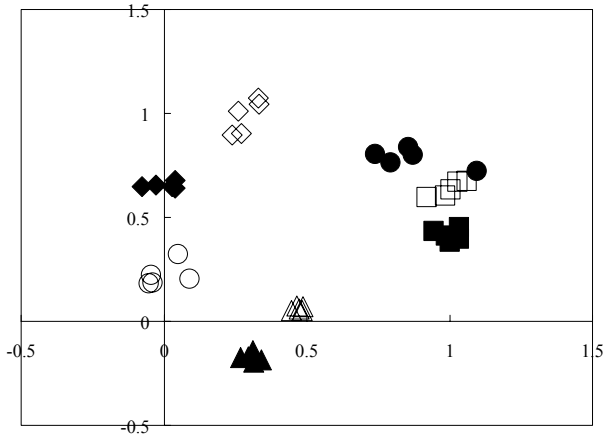
Fig. 7 PB Spaces Acquired by Sensor Modality

The RNNPB was trained in two different ways. The RNNPB-1 was trained using all multi-modal sensory data in active sensing motion. The RNNPB-2 was trained using only *the rubber ball, glass, scrub brush,* and *shampoo container* (empty). For this RNNPB-2, *moneybox, pen tray, soft brush,* and *shampoo container* (full) were unknown objects. Both RNNPB were trained 100,000 times.

Figure 8 shows the PB spaces of RNNPB-1 (a) and RNNPB-2 (b) respectively. The trained/untrained objects are shown by white/black plots respectively. Here, the parameter values of RNNPB-2 corresponding to unknown objects were determined by renewing only the parameter values without updating the synaptic weights (recognition process). Renewing the PB value only 1000 times completed the recognition.



(a) PB Space Acquired of RNNPB-1
Trained by 8 objects



(b) PB Space Acquired of RNNPB-2
Trained by 4 objects and recognition results of untrained 4 objects

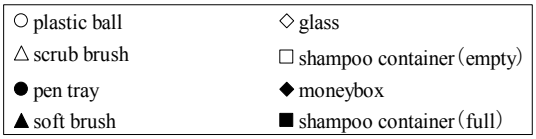


Fig. 8 Comparison between two RNNPB
(Generalization Analysis)

It can be observed that the clusters were self-organized corresponding to all objects in the PB space of the RNNPB-1. Specifically,

1) Objects that moved easily were mapped in the upper-left area,
2) Objects making sound were mapped in the upper area,
3) Blue objects were mapped in the upper-right area.

Furthermore, Figure 8-(b) demonstrates that RNNPB-2 acquired almost the same map as Figure 8-(a) of RNNPB-1, even though it had been trained with the data of only four objects. This means that the RNNPB-2 could configure the PB space with a similar structure to that of the RNNPB-1 except the distribution of each cluster was different between two PB spaces.

V. Discussion

A. Motion Design and Multi-Modality

It is difficult to prepare a motion pattern that can be applied to various kinds of objects. Most research concerning active sensing selected touch and/or grasping motions which focused on tactile and joint-degree sensing. Though these motions guarantee to obtain reliable data about the shape, size, and weight of objects, such motions require elaborate skills for detecting the accurate position of objects in order to pick them up. However, it is well known that even human infants have difficulty manipulating objects with their own hands.

We selected the motion of object moving/hitting by the robot's arm for the active perception. This motion can be completed without any accurate sensor data and it is possible to extract many kinds of dynamic features of objects including

moving trajectory and sound. In particular, the 'sound' signal reflects many properties of objects such as shape, material, and internal structure. Visual device alone cannot obtain these properties simultaneously as shown in Figure 7-(c).

A large number and variety of motion patterns of moving/hitting objects in different speeds and directions could enable the extraction of a greater variety of dynamic features of objects.

B. Clustering Ability of RNNPB

As mentioned in Section I, most conventional studies concerning active sensing have dealt with few objects as the recognition target. Hence it is also difficult to prepare a recognition system that can handle various kinds of objects. For example, typical neural networks for time-sequence data processing represented by TDNN need an impractically large number of neurons and learning times for the problem treated in this paper, because it is designed to store all time-sequences of sensory data in the input layer. In contrast, the number of neurons in our RNNPB was only 42 because it uses self-organizing contextual information in the context-layer.

C. Generalization Ability of RNNPB

We confirmed that RNNPB has superior generalization capability for clustering dynamic sequences. It can express various objects and their relationships in the PB space self-organized through training with a few objects. As mentioned in Section II, our method is better for robot learning than stochastic learning methods because real robot systems have fatal limitations of hardware-durability.

There is another deterministic learning system called "mixture of experts" represented by MOSAIC [14] which also works well to deal with multiple dynamic-patterns (attracters). This type of system usually consists of several dynamic recognizers which categorize and learn target sequences individually (local expression). In contrast, RNNPB acquires multi-attracters in overlapping fashion in a single network by changing parameters that represent the boundary condition (distributed expression). In RNNPB, all neurons and synaptic weights participate in representing all trained patterns.

In local expression, interference is minimized between patterns because it allocates a novel pattern in an additional recognizer. However in a distributed expression, memory interference will occur since the memories share the same network resources. Nevertheless, as a result of embedding multiple attractors in a distributed network, we got a global structure that handles learned patterns as well as unknown (unlearned) patterns. We think this is why RNNPB could show the generalization ability in recognizing unknown objects in Section IV-B.

VI. Conclusions and Future work

This paper proposed an active sensing method using a humanoid robot with a recurrent neural net to solve the problem of object recognition. Specifically, the RNNPB model with only 42 neurons was trained with the data of sounds, trajectories, and tactile senses generated while a humanoid robot was moving/hitting an object with its own arm. The clusters of 20 kinds of objects could be self-organized in the parametric bias space (PB space). Also experiments using unknown (not trained) objects demonstrated that the proposed method could configure these unknown objects in PB space appropriately, which proves its *generalization* capability**.**

An interesting challenge for future work is to achieve the robot motion planning using our method. The configuration of the RNNPB can be redesigned for treating the motor output easily. We expect that our robot will be able to generate arm motion by using the RNNPB output. For example, the RNNPB could associate the arm motion patterns with observed object trajectories and sounds. This association could be related to the discussion of "imitation" based on behavioral primitives corresponding to the parametric bias in our study.

Acknowledgment

This research was supported by RIKEN Brain Science Institute and the Scientific Research on Priority Areas: "Informatics Studies for the Foundation of IT Evolution".

References

[1] R. Bajcsy, "Active Perception," *IEEE Proceedings, Special issue on Computer Vision*, Vol. 76, No. 8, pp. 996-1005, 1988

[2] K. Noda, M. Suzuki, N. Tsuchiya, Y. Suga, T. Ogata, and S. Sugano, "Robust modeling of dynamic environment based on robot embodiment," *IEEE ICRA 2003*, pp. 3565-3570, 2003.

[3] A. Arsenio and P. Fitzpatrick, "Exploiting cross-modal rhythm for robot perception of objects," *Int. Conf. on Computational Intelligence, Robotics, and Autonomous Systems*, 2003.

[4] P. Dario, M. Rucci, C. Guadagnini, and C. Laschi, "Integrating Visual and Tactile Information in Disassembly Tasks," *Int. Conf. on Advanced Robotics*, pp. 191-196, 1993.

[5] T. Kohonen, "Self-Organizing Maps," *Springer Series in Information Science, Vol. 30*, Springer, Berlin, Heidelberg, New York, 1995.

[6] L. Lin and T. Mitchell, "Efficient Learning and Planning within the Dynamic Framework," *SAB'92*, pp. 281-290, 1992.

[7] J. Tani and M. Ito, "Self-Organization of Behavioural Primitives as Multiple Attractor Dynamics: A Robot Experiment," *IEEE Transactions on SMC Part A*, Vol. 33, No. 4, pp. 481-488, 2003.

[8] T. Ogata, M. Matsunaga, S. Sugano, and J. Tani, "Human Robot Collaboration Using Behavioral Primitives," *IEEE/RSJ IROS 2004*, pp. 1592-1597, 2004.

[9] M. Jordan, "Attractor dynamics and parallelism in a connectionist sequential machine," *Eighth Annual Conference of the Cognitive Science Society* (Erlbaum, Hillsdale, NJ), pp. 513-546, 1986.

[10] D. Rumelhart, G. Hinton, and R. Williams, "Learning internal representation by error propagation," *in D.E. Rumelhart and J.L. McLelland, editors, Parallel Distributed Processing (Cambridge, MA: MIT Press)*, 1986.

[11] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, and R. Nakatsu, "Robovie: an interactive humanoid robot." *Int. Journal of Industrial Robotics*, Vol. 28, No. 6, pp. 498-503, 2001.

[12] T. Miyashita, T. Tajika, K. Shinozawa, H. Ishiguro, K. Kogure and N. Hagita, "Human Position and Posture Detection based on Tactile Information of the Whole Body," *IEEE/RSJ IROS 2004 Work Shop*, 2004

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

[14] M. Haruno, D. Wolpert, and M. Kawato, "MOSAIC model for sensorimotor learning and control," *Neural Computation* 13, pp. 2201-2220, 2001.