Acquisition of Flexible Image Recognition by Coupling of Reinforcement Learning and a Neural Network

Katsunari SHIBATA * and Tomohiko KAWANO *

Abstract : The authors have proposed a very simple autonomous learning system consisting of one neural network (NN), whose inputs are raw sensor signals and whose outputs are directly passed to actuators as control signals, and which is trained by using reinforcement learning (RL). However, the current opinion seems that such simple learning systems do not actually work on complicated tasks in the real world. In this paper, with a view to developing higher functions in robots, the authors bring up the necessity to introduce autonomous learning of a massively parallel and cohesively flexible system with massive inputs based on the consideration about the brain architecture and the sequential property of our consciousness. The authors also bring up the necessity to place more importance on "optimization" of the total system under a uniform criterion than "understandability" for humans. Thus, the authors attempt to stress the importance of their proposed system when considering the future research on robot intelligence. The experimental result in a real-world-like environment shows that image recognition from as many as 6240 visual signals can be acquired through RL under various backgrounds and light conditions without providing any knowledge about image processing or the target object. It works even for camera image inputs that were not experienced in learning. In the hidden layer, template-like representation, division of roles between hidden neurons, and representation to detect the target uninfluenced by light condition or background were observed after learning. The autonomous acquisition of such useful representations or functions makes us feel the potential towards avoidance of the frame problem and the development of higher functions.

Key Words : reinforcement learning, neural network, function emergence, robot, image recognition.

1. Introduction

The authors have long-held misgivings about so-called "intelligent robots". "Are intelligent robots which are provided many pieces of knowledge by humans actually intelligent?" Such skepticism led us to pursue an extreme target, that is, how little knowledge robots need to acquire their process through learning, including those usually considered unworthy of learning, such as image processing. We have then proposed a very simple autonomous learning system consisting of one neural network (NN), whose inputs are raw sensor signals and whose outputs are directly passed to actuators as control signals, and which is trained by using reinforcement learning (RL) [1]–[3].

Unlike recognition and control, which are closer to sensors and actuators respectively, in the flexible acquisition of higher functions in robots, it is very difficult to decide in advance what the inputs and outputs should be. This difficulty might be the main reason why the research on higher functions has made less progress than the other. We believe that, without a significant change in thinking, an intelligent system that enables the emergence of higher functions cannot be achieved. It also becomes important that learning not be specific only for one purpose, but flexibly applicable for the emergence of various functions according to the necessity and situation. From this viewpoint, the autonomy, flexibility, harmony, and generality in learning should be evaluated together with the acquired function itself.

The focus of recent research for developing more intelligent robots or agents based on autonomous learning has centered on "prediction", and some interesting works have been done [4]–[7]. Some of them use or consider using RL, but "prediction" is performed separately from RL. However, if the number of sensor signals is large especially when a visual sensor is used, it might be impossible to predict all sensor signals at every future time step. Therefore, the information that is predicted should be picked up from all sensor signals, and some relevance criterion is necessary for it. Discovering an appropriate criterion is the process needs high intelligence. As a criterion for prediction, linear independence has been proposed [8], but we think that purposive compression is necessary for consistency with the system purpose and effective compression.

Currently, the autonomous learning ability of RL and parallel and flexible learning in NNs are widely accepted. However, many researchers are still positioning RL as learning for actions in the total process, and the NN as just a function approximator. No one has ever tried to apply RL to a NN with a large number of inputs in a real-world-like environment without explicit state space construction. The prevalent idea seems to be that the learning system in which regular RL and a NN are just coupled is too simple to work on complicated tasks in the real world.

We believe that in such tasks, our proposed system can make more significant contributions than alternative methods because it is increasingly difficult for humans to design appropriate processing. In this paper, in order to support the belief and attempt to stress the significance of our proposed system, we put forth two necessities for the future research on human-like intelligent robots. Those are the necessity to introduce autonomous learning of a massively parallel and cohesively flexible system with massively inputs from the consideration about the brain

* Department of Electrical and Electronic Engineering, Oita University, 700 Dannoharu, Oita 870-1192, Japan
E-mail: shibata@cc.oita-u.ac.jp, shibata@oita-u.ac.jp (future)
(Received April 21, 2008)
(Revised September 4, 2008)

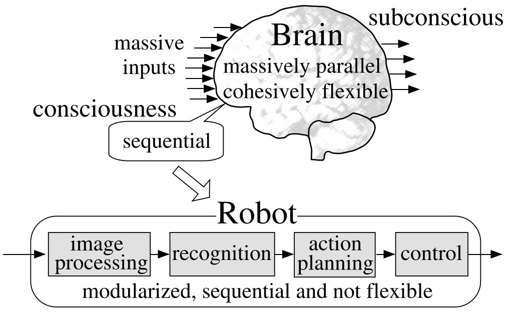


Fig. 1 The comparison between brain and robot processing.

architecture and the sequential property of our consciousness, and also the necessity to place more importance on "optimization" of total system under a uniform criterion than "understandability" for humans. After that, we introduce one successful experimental result in a flexible image recognition task in a real-world-like environment, and we hope that the analysis result of hidden neurons' representation in the NN shows the autonomous learning ability and the potential towards the development of higher functions in our proposed system.

2. The Marriage of Reinforcement Learning (RL) and Neural Network (NN)

Here we introduce two concepts for aiming at developing a system that enables the emergence of higher functions in the future research on robot intelligence, and try to support the significance of our proposed learning system.

2.1 The Necessity of Massively Parallel and Cohesively Flexible System with Massive Inputs

When comparing the processing between humans and robots, it is easily noticed that our brain is a massively parallel and cohesively flexible system with huge sensory inputs, while the processing in robots seems modularized, sequential and generally not so flexible as shown in Fig. 1. We can also notice that our consciousness seems sequential even though the brain it originates from is a massively parallel system. It might be true that much of our processing is performed subconsciously in our brain. For example, we are not aware of the orientation selectivity in the visual cortex. This means that we do not have the means to understand the processing in our brain exactly, and we can do nothing more than use sequential consciousness to try to guess the processing of the brain. Accordingly, we understand the brain function by dividing it into modules, and they are arranged sequentially when the robot processes are designed. That might cause the "frame problem" [9],[10]. The modularized system is not so flexible because the input and output of each module have to be defined in advance. We do not think that higher functions can be developed as an extension of such a sequential and modularized system.

Imaging and electrical recording of brain activity seem insufficient to understand the exact process of the whole brain. Even though they provide sufficient information, it is difficult for us to understand the mechanism and to apply it to robots. "Optic illusions" and the "choice blindness" [11] also seem to suggest the gap between what we do and what we are conscious of.

While maintaining its harmony, our brain is nonetheless very flexible. For example, when one sensor cell dies, other cells and neurons that exist already compensate for the lack of the dead cell by their growth and learning.

From the above discussion, we put forth the necessity of a massively parallel and cohesively flexible system with massive inputs for the emergence of higher functions.

2.2 The Necessity of Optimization of Entire Process under One Uniform Criterion

The Subsumption Architecture [12] proposed by Brooks has had a great significance in discovering the problem of cumulative sequential processing in the conventional approach and in introducing a novel way of employing a parallel architecture. Actually, such architecture has shown the usefulness of avoiding the "frame problem" [9],[10]. However, he states that the decomposition of a complex system into parts is necessary in all engineering endeavors [12]. He also puts "understandability" for humans before "optimization" for creatures. This notion results in the introduction of layers, in other words, functional modules although they are connected in parallel. He suggests developing useful programs for the layers and designing the interaction between them even though he expresses concerns about the extensibility of his system and development of higher functions. Its fixed topology network and increasing complexity in the interaction design between layers are the main problems with his approach.

Here, let us reconsider what the processing in robots should be. In this context, the processing consists in deriving the outputs, as actuator commands, from the sensor signals given as inputs. The objective is generally to obtain appropriate outputs for some purpose under the condition of the inputs. In other words, the objective is to solve an optimization problem as shown in Fig. 2(a). As Brooks suggested [12], researchers have often placed "understanding" before "optimization". However, as long as the processing in robots is not modified manually, the objective should be "optimization" rather than "understandability". A radical idea asserts that all the functions humans have are obtained as a result of optimization of the outputs under the condition of given inputs. For example, better recognition leads to better actuator commands. Thus, considering the discussion in the last subsection, we suggest to introduce some parallel processing system, and to optimize it to obtain appropriate outputs for some purpose. We should not interfere with the robots' processes, but leave everything to their own optimization as much as possible.

It is important that the optimization should be useful for the future states, but in the real world, creatures do not encounter exactly the same states as those encountered in the past. Nevertheless, humans can perform appropriate actions by referring to past experiences. The key properties realizing this ability are "generalization" and "abstraction". "Generalization" usually means the property of having similar outputs corresponding to similar inputs. On the other hand, humans, for example, can grasp the same meaning from either '2' or 'II' on a clock face even though the images of the characters are not similar to each other. In this meaning, construction of a useful abstract space and generalization on that space is important [13],[14].

Brooks states that abstraction is the essence of intelligence and the hard part of the problems being solved [12]. Most researchers might agree that abstract information is the important information extracted from sensor signals. However, what is necessary is to discover the criterion to decide which infor-

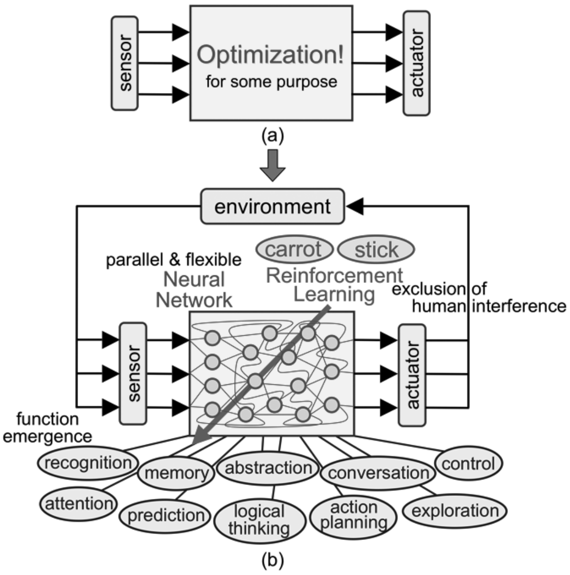


Fig. 2 (a) The objective in robots is to optimize the process from sensors to actuators for some purpose. (b) Parallel and cohesively flexible learning of the whole process from sensors to actuators by the coupling of RL and NN and the emergence of necessary functions.

mation is important. Feature extraction in image processing can be considered as one type of abstraction. However it is not acquired autonomously based on some relevance criterion, but is provided by humans. One possible relevance criterion can be "how well the sensor signals are reconstructed from the low-dimensional abstract signals". It is used to compress high-dimensional signals using principle component analysis or other method like [15]. A bottleneck NN that learns identical mapping with fewer hidden neurons [16],[17], is also based on the criterion. However, the abstraction is performed without considering the purpose of the system. Therefore, we are afraid that purposive and effective compression cannot be expected. If optimization is the objective to develop the process in robots, the criterion for abstraction should be consistent to it. The discussion is the same as the case of "prediction" in Introduction.

From the above discussion, we put forth the idea that whole the system should be optimized under one uniform criterion, and leave everything to the optimization in robots as much as possible for the emergence of higher functions.

2.3 Proposed System and Learning

As aforementioned, we have previously proposed a system in which a NN, whose input are sensor signals and whose output are passed directly as actuator commands, and which is trained by using RL as shown in Fig. 2(b) [1]–[3]. The NN is a parallel and cohesively flexible processing system. RL provides the optimization towards a given purpose. A NN requires training signals to achieve purposive learning. If they are provided by humans in advance, they can be a constraint for the creature. However, RL can always generate and provide the training signal to the NN autonomously, and that enables purposive optimization of the total system under one uniform criterion. If it is assumed that a NN can always realize the optimization through learning, the system has very suitable properties for emergence of intelligence. Through purposive learning, various functions such as recognition, planning and control are expected to emerge according to the necessities of a given task. It has also been verified that by using a recurrent NN, "attention and associative memory" [18], "communication" [19], and "contextual behavior" [20] emerge in very simple tasks. Moreover, we expect that "prediction" to which many researchers are giving much attention in developing intelligence, as mentioned in Introduction, can be obtained through learning.

Since the NN trained by RL is optimized on the criterion through learning, the hidden representation in NN is suitable to be called abstract information. NN trained by RL has an excellent ability to realize such purposive abstraction and knowledge transfer is effectively done through learning [21]. Such ability is essential for the learning in the real world and leads to higher functions. From this viewpoint, the discovery of useful hidden neurons' representation will be focused in the next "experiment" section.

The combination of NN and RL sometimes causes instability of learning [22]. However, if a continuous state space is divided into some local spaces, and each input signal represents the information in only one of the local spaces, learning becomes stable. The reason for this is that the learning for an input pattern does not influence so much to that for the other input patterns. The mechanism is similar to that of the stability in CMAC [23] or RBF-based network including NGnet(Normalized Gaussian Network) [24]. The regular NN has one or more hidden layers that can represent abstract information, and that is different from the case of CMAC or RBF-network. However, the stability was shown empirically even in such cases, and abstract space can be reconstructed very flexibly in the hidden neurons if each input signal represents local information [2],[25],[26].

In the proposed system, RL can be either Q-learning [27] or actor-critic [28]. Since Q-learning is used and the task is episodic in the experiment in this paper, the way of train a NN using Q-learning in an episodic task is explained. Q-learning is usually represented by the update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right] \quad (1)$$

where $Q(s_t, a_t)$ is Q-value for state-action pair (s_t, a_t), r_t is a reward, and γ is a discount factor. According to this update, $Q(s_t, a_t)$ moves towards $r_{t+1} + \gamma \max_a Q(s_{t+1}, a)$. When a NN is used, the number of outputs is equal to the number of the actions, and each output is dealt with as the Q-value for the corresponding action. At a non-terminal state with a new sensor signal vector S_{t+1}, the training signal $T_{a_t,t}$ for the output of the executed action a_t with no provided reward is calculated as

$$T_{a_t,t} = \gamma(\max_a O_a(S_{t+1})), \quad (2)$$

where $O_a(S_{t+1})$ is the a-th output of the NN when S_{t+1} is entered. When S_{t+1} is a terminal state, it is

$$T_{a_t,t} = r_{t+1}. \quad (3)$$

The NN is trained by Back-Propagation [16] using the training signal. However, to calculate $T_{a_t,t}$ at non-terminal state, $O_a(S_{t+1})$ is necessary. Accordingly after forward computation for the input S_{t+1} and calculation of $\max_a O_a(S_{t+1})$, forward computation for the input S_t is done again and training by Back-Propagation using $T_{a_t,t}$ is done.

3. Experiment

3.1 Setup

Here we introduce an experiment using two AIBO robots to show that the autonomous learning ability of our proposed system works effectively in real-world-like environment. The environment is shown in Fig. 3. Two AIBOs are placed face-to-face 43cm from one another. The black AIBO can rotate its head around the vertical axis and catch sight of the white AIBO using its color camera located at its nose. The horizontal angle of view of the camera is 57 degree. The number of pixels is around 350,000, but when the AIBO sends the image to the computer, it is reduced to $52 \times 40 = 2080$ by pixel skipping. The resolution and range of the head motion are 5 and ±20 degrees respectively. That means that there are a total of 9 head states where the location of the white AIBO in the camera image is different. However, here, no explicit state space construction is given, and raw sensor signals are entered to the NN directly.

The procedure that the black AIBO has to perform in the task is shown in Fig. 4. There are three actions that the black AIBO can choose. They are "rotate right", "rotate left", and "bark". The aim is to bark after rotating its head until it catches the white AIBO at the state 0 that indicates the white AIBO is located at the center of the captured image. When it barks correctly, it receives a reward, but if it barks at an incorrect state (state −4,..,−1, 1,.., 3 or 4), a small penalty is imposed. ϵ-greedy ($\epsilon = 0.13$ here) is employed for action selection.

As shown in Fig. 5, the NN has 5 layers, and the inputs are the raw image signals (2080 pixels × 3 colors (RGB) = 6240) after inverting each pixel value and normalizing it between 0.0 and 1.0. There are 3 output neurons, each of which represents Q-value and is corresponding to one of the 3 actions. The number of neurons in each layer is 6240-600-150-40-3 from input layer to output layer. The training signal is generated by using Q-learning [27] and the NN is trained by Error Back Propagation [16]. When the action of "rotate right" or "rotate left" is chosen at time t, the head is rotated and forward computation of NN at time $t+1$ is performed with the new input S_{t+1} that can be obtained after the head rotation. After that, the input S_t are inputted into the NN again, and only the output for the executed action at time t is trained by the training signal as

$$T_{a_t,t} = \gamma(\max_a O_a(S_{t+1}) + 0.4) - 0.4 \qquad (4)$$

on behalf of Eq. (1). Here, the sigmoid function used as the output function of each neuron ranges from −0.5 to 0.5. To adjust the offset between Q value and output, 0.4 is added and subtracted in Eq. (4). That means that Q-value 0.0 corresponds to output −0.4 and Q-value 0.8 corresponds to output 0.4. Here, the discount factor γ is 0.8. When the black AIBO barks at time t, the trial terminates, and the training signal is provided as

$$\begin{aligned} T_{a_t,t} &= 0.4 \qquad\qquad \textit{at state}0 \\ &= O_{a_t}(S_t) - 0.02 \qquad\qquad (5) \\ &\qquad\qquad \textit{at all the other states.} \end{aligned}$$

After that, another trial (episode) starts from a state chosen randomly. Since learning takes so long using a robot in real time, learning was performed using some images captured beforehand. The images were captured under various backgrounds and light conditions. 312 patterns (312 × 9 states = 2808 samples) were used for learning, and 92 patterns (92×9 states = 828 samples) were used for testing. If one image input is considered as one state, the number of states is 2808 in the learning phase and when the test patterns are used, none of the training states appears. At each state transition, one image is chosen randomly from images for the new head state. This means that the state transition is not deterministic. Furthermore, because of the use of a real robot, the position of AIBO in the image is sometimes shifted slightly from what is expected.

Figure 6 shows the variety of captured images due to the light condition and background. We can see that it is not so easy to write a program to recognize the location of AIBO. The face and shoulder areas of Fig. 6 (a) and (b) are cropped and enlarged as shown in Fig. 7. The number beside each image indicates average brightness. Note that the brightness is almost the same between the black area in the bright condition and the white area in the dark condition even though they are the opposite colors.

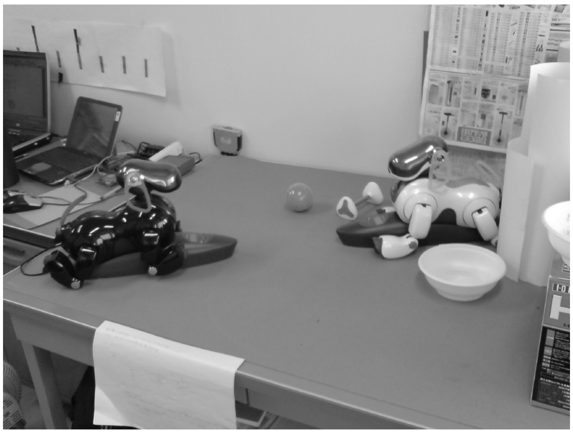


Fig. 3 The environment for the experiment.

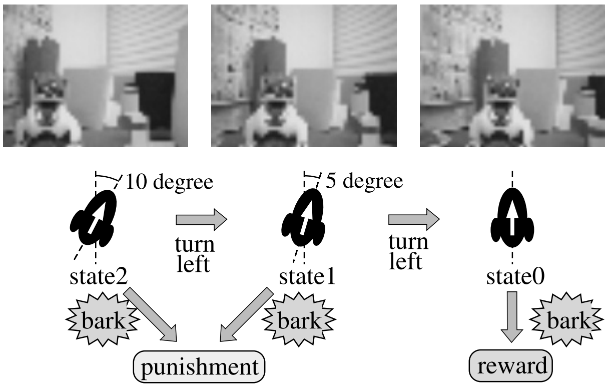


Fig. 4 Sample of the state transition in the experiment.

3.2 Learning Results

Figure 8 shows the learning curve. At every 100 trials, success ratio over 50 trials with $\epsilon = 0$ is observed for the cases of both learning patterns and test patterns. The initial head location in each trial is +20 degree or −20 degree in the test phase. The success ratio rose steeply at around 5000 trials in both cases, and finally reached around 95% for the learning patterns and 90% even for the test patterns that were not used in learning. After 20,000 learning trials, the performance was also examined on the real robot. The success ratio for 100 trials with 5 different backgrounds for each of daytime and night is shown

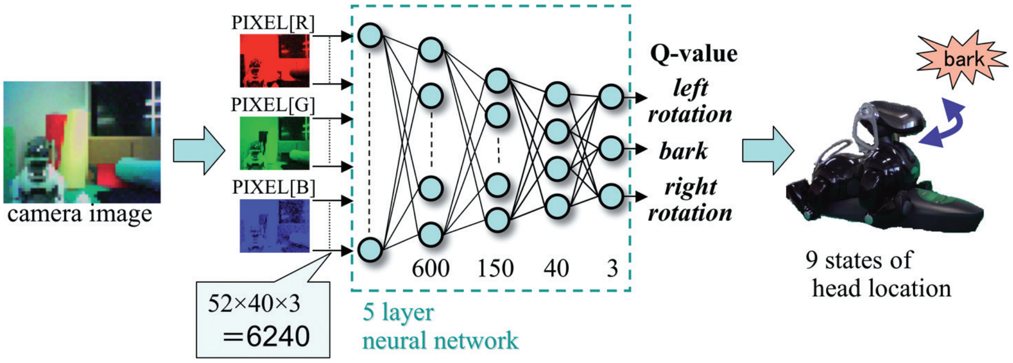


Fig. 5 The processing in the black AIBO.

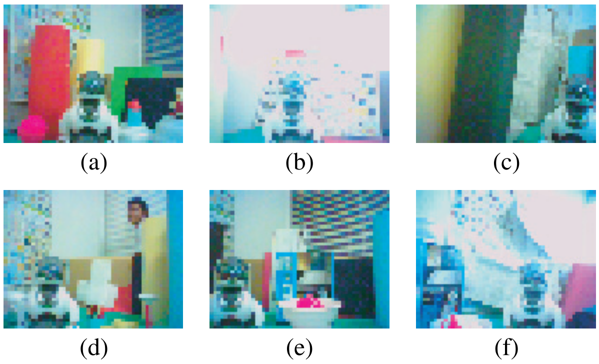


Fig. 6 Some sample images that were captured under different light conditions and with different backgrounds and were used in learning.

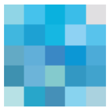
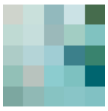
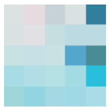
	light condition	
	(a) dark (night)	(b) bright (daytime)
face (black area)	65	 138
shoulder (white area)	 148	 185

Fig. 7 Four magnified images each of which is the black (face) part or white (shoulder) part of AIBO in one of the two images (a) and (b) in Fig. 6. The number beside each image indicates the average brightness.

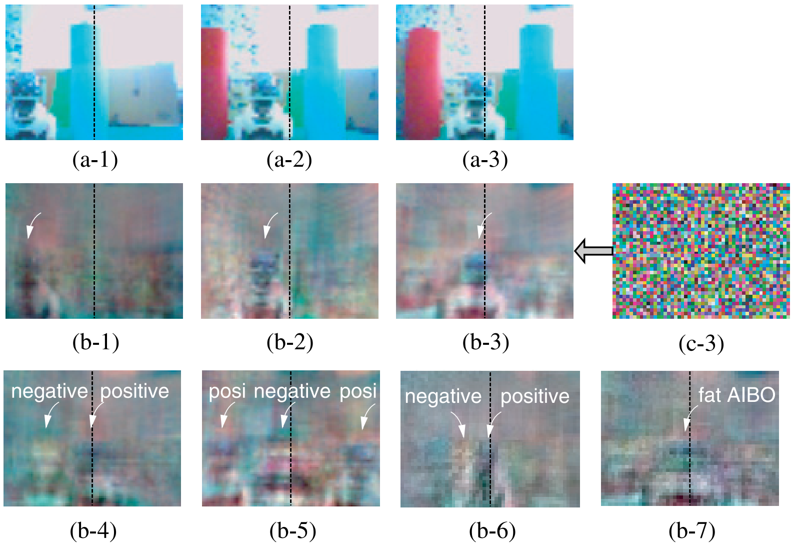


Fig. 9 (a) Samples of actual image. (b-1,..,6) The images that visualize the change of the connection weights from the input layer for 5 lowest hidden neurons through RL. AIBO images can be found at the place where small white arrows are pointing. (b-7) Visualized connection weight change of one neuron in the middle hidden layer. (c-3) Visualized connection weights for the same hidden neuron as (b-3).

in the right part of Fig. 8. The ratio was equal to or more than 90%. The large success ratio for the test patterns indicates that the acquired recognition function has generalization ability.

3.3 Visualization of Hidden Representation

Next, the connection weights of each of the 600 neurons in the lowest hidden layer were observed. Because the number of connections in each lowest hidden neuron is the same as the number of inputs, the weight pattern after some linear transformation can be observed as a color image whose size is 52×40. The weight patterns seem random as shown in Fig. 9 (c-3) by the influence of initial connection weight that was determined randomly. However, revealing patterns could be observed only when the change of each weight from the initial value was observed as shown in Fig. 9 (b-3). The linear transformation from each weight value to the corresponding pixel value $f_{c,i,j}$ that ranges from 0 to 255 is as

$$f_{c,i,j} = (int)\left(\frac{Wa_{c,i,j} - Wb_{c,i,j}}{\max_{c,i,j}\left|Wa_{c,i,j} - Wb_{c,i,j}\right|} \times 127\right) + 128, \qquad (6)$$

where Wa, Wb indicate the weight after and before RL respectively, c indicates the color and can be R, G, or B, and i, j indicate the raw and column number of a pixel in the image respectively. By this transformation, if the value of a connection weight increases during learning, the pixel corresponding to it becomes bright, and if the value decreased, the pixel becomes dark. The maximum absolute value of the weights for the lowest hidden neuron that is shown in Fig. 9 (b-3) and (c-3) is 0.127 before RL, and 0.131 after RL. While, the maximum absolute value of the change of weights between before and after RL is as small as 0.012.

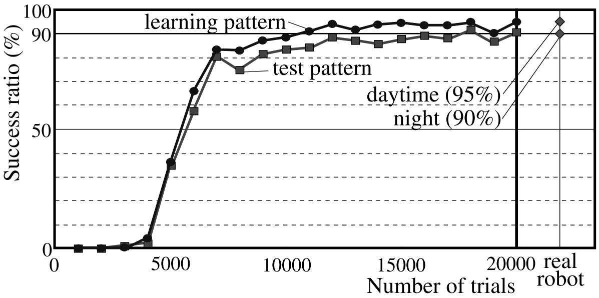
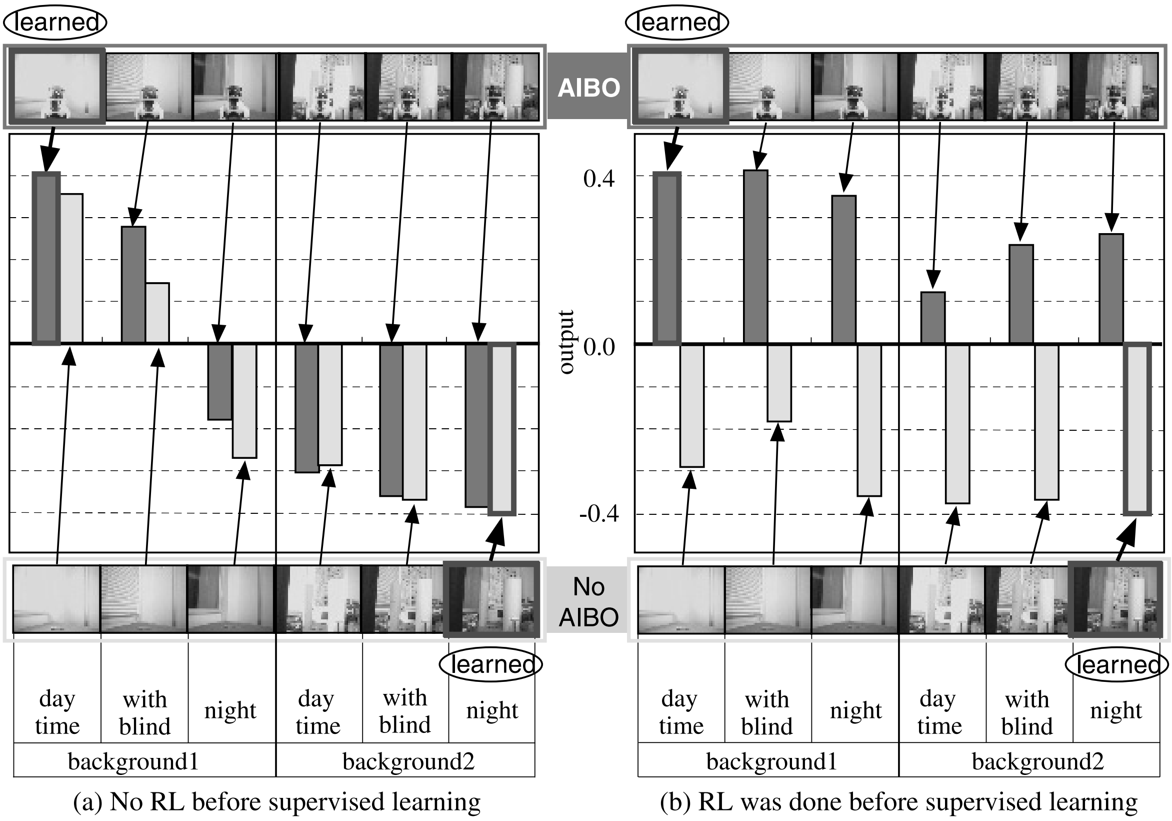


Fig. 8 Learning curve for both training and test patterns, and the success ratio when using the real AIBO robot after learning.

In most of the weight images, AIBO silhouette could be found. Some of them are shown in Fig. 9 (b-1,..,6). Some real pictures are shown in Fig. 9 (a-1,2,3) for reference. The location of the AIBO image appears in (b-1,2,3) is almost the same as that in the real pictures (a-1,2,3). It can be inferred that each of the three neurons in the lowest hidden layer works like a template and plays a role in detecting whether the AIBO is caught at a particular location on the image. What is significant is that such division of roles among hidden neurons can be realized autonomously through RL even though either input images or propagated error signals that are provided to each neuron are not controlled by anyone. In Fig. 9 (b-4), one becomes aware that one AIBO located at the left half reveals a negative pattern and the other AIBO around the center reveals a positive one. In Fig. 9 (b-5), one can find a blurred negative AIBO image around the center and a positive one around each of the left and right ends. In Fig. 9 (b-6), one half of the AIBO image looks positive and the other half looks negative. The appearance of both positive and negative AIBO images possibly helps in the recognition uninfluenced by light condition.

The weight-change images of the neurons in the middle hidden layer are observed by normalizing the weighted sum of the weight-change in the lowest hidden neurons by the connection weights between the lowest and middle hidden neurons. In



(a) No RL before supervised learning

(b) RL was done before supervised learning

Fig. 10 One output neuron was added to the NN after RL, and supervised learning was performed using only two input images labeled as "learned". The training signal for one pattern is 0.4, while it is −0.4 for the other. After that, the output for some test input images was observed. The light conditions, background, and also the presence of AIBO are varied in the input images. The outputs are compared with (a) the case that no RL was done before this learning and (b) the case that RL was done before this learning. In the latter case, the output changes according to the presence of AIBO.

most of them, AIBO silhouette could be also found. It is interesting that a fat AIBO silhouette as shown in Fig. 9 (b-7) could be found very often. It is not certain, but it might reduce the influence of head displacement on recognition. Our guesses might be only a part of the functions that the NN acquired, as we cannot understand completely how the brain functions when we see the excitation pattern of real neurons in the brain.

3.4 Observation of Hidden Representation through Additional Supervised Learning

Finally, in order to show that the NN had an internal representation to represent the AIBO location without being influenced by the background or light condition, additional supervised learning is performed to the NN after RL. The output neurons were removed and a new output neuron with 0 connection weight from all the neurons in the upper hidden layer was added. 12 new images were captured after putting the two AIBOs at a different place where the learning and test patterns were captured. There are 6 pairs of images, and in each pair, only the difference is whether the white AIBO appears on the center of the image or not. In 3 pairs, no object was located in the background, while in the other 3 pairs, some objects were located. One of the 3 pairs was captured in daytime, another was captured also in daytime but shaded with a blind, and the other was captured at night. In this learning, among 12 images, only the two images that are labeled as ”learned” in Fig. 10 were used in learning. In one of them, AIBO appears in the image with nothing placed behind it, and the image is bright because it was captured during daytime. In the other, the AIBO is not present, and some objects are placed. The image is not bright because it was captured at night. One of the two images was chosen randomly and presented as input at each step during learning. For the first image, the training signal 0.4 is given, and for the other, −0.4 is given.

After the output for two training images was almost equal to each training signal, the output for the other 10 images is observed as shown in Fig. 10. The output is compared with two cases when no RL was done before this supervised learning (Fig. 10 (a)) and when supervised learning was done after RL (Fig. 10 (b)). The initial connection weights were the same between them. When no RL was done before, the output changes according to the background and also light conditions, but does not change so much according to the presence of the AIBO. This is logical from the viewpoint of generalization because the value in more pixels changes according to the background and light conditions than according to the presence of AIBO. The AIBO occupies only a small part of the image.

However, when supervised learning was done after RL, the sign of the output is determined by whether the AIBO is present or not. This means that through learning, the NN obtained the internal representation to express whether the AIBO appears in the center of the image or not without being influenced by the background or light condition. Generalization worked on the hidden neurons’ representation that is acquired through RL rather than on the input sensor signals’ space.

In this experiment, no one told the black AIBO to recognize the white AIBO location nor did anyone teach the black AIBO how to recognize the white AIBO location without being influenced by the background or light condition. However, the black AIBO acquired the ability to recognize the white AIBO autonomously through RL. It can be said that the black AIBO could learn not only appropriate actions, but also could learn autonomously which information is important to achieve the given task.

4. Discussion

If we consider that this task is to classify 2808 images consisting on 6240 pixels into 3 categories, and the head state is limited in only 9 states even though its control is not so precise, the task seems very easy. (The first report about the result of a more difficult walking task will appear in [29].) However, we know in advance what the target of the task is and that the head location is limited. Also, we know what the template matching technique is, and we can guess that it probably works well after some compensation of light condition because the head location is limited. For these reasons, we think the task should be easy. On the other hand, the AIBO did not know them at all before learning. Any classification samples were not provided during learning, and the AIBO had to learn what is an appropriate classification during learning. The reward for successful bark and punishment for incorrect bark are the only things provided to the AIBO during learning. The architecture and learning are quite simple and general, and are not designed specially for this task. Nevertheless, the AIBO finally discovered the template-like hidden representation, and furthermore, it is not influenced much by light conditions or background. This suggests us that the autonomy and flexibility of the learning system is excellent. Even though the task is learned as a classification task by SVM (Support Vector Machine), such abstraction cannot be achieved. This is because SVM aims at classification problems, but does not have a way to form the intermediate representation flexibly and purposively through learning.

It has been pointed out that in developing intelligent robots, the frame problem [9],[10] becomes serious. Brooks showed that it is useful to introduce a parallel architecture on behalf of conventional sequential processing for the frame problem [12]. However, as he pointed by himself, it is difficult to design each module and how to connect them in parallel. We cannot see the way to higher functions in this approach. In our proposed learning system, functions autonomously emerge in the NN that is a parallel processing system and no special technique is added for special purposes. This would suggest that the system is free from the “frame problem” fundamentally and has a potential towards the emergence of higher functions thanks to its autonomous, flexible, parallel, harmonious and general property of learning.

5. Conclusion

From the viewpoint of our insufficient ability to understand and express the parallel and flexible mechanism of our brain and the notion of “optimization” of the system, the authors recommend interfering with the couple of reinforcement learning and a neural network as little as possible to obtain intelligence. In the experiment in a real-world-like environment, template-matching-like image processing, autonomous division of roles among hidden neurons and also internal representations that are not influenced by the background or light conditions could be observed in the neural network that is parallel and flexible as is our brain even though no prior knowledge about the task or image processing was given to the robot. This work shows the potential of the couple in complicated tasks in the real world and evidence towards the emergence of higher functions.

Acknowledgments

This research was supported by JSPS Grant in-Aid for Scientific Research #15300064 and #19300070. The authors would like to express their gratitude to the reviewers who gave us very helpful comments to organize the early part of this paper.

References

[1] K. Shibata and Y. Okabe: Reinforcement learning when the visual signals are directly given as inputs, *Proc. ICNN '97*, Vol. 3, pp. 1716–1720, 1997.

[2] K. Shibata, Y. Okabe, and K. Ito: Direct-vision-based reinforcement learning using a layered neural network, *Trans. SICE*, Vol. 37. No. 2, pp. 168–177, 2001 (in Japanese).

[3] K. Shibata and M. Iida: Acquisition of box pushing by direct-vision-based reinforcement learning, *Proc. SICE Annual Conf. 2003*, 0324.pdf, pp. 1378–1383, 2003.

[4] J. Schmidhuber: *Exploring the predictable, Advances in Evolutionary Computing*, pp. 579–612, Springer, 2002.

[5] J. Tani: Learning to generate articulated behavior through the bottom-up and the top-down interaction processes, *Neural Networks*, Vol.16, No.1, pp. 11–23, 2003.

[6] R. S. Sutton, E. J. Rafols, and A. Koop: Temporal abstraction in temporal-difference networks, *Advances in Neural Information Processing Systems*, Vol. 18, pp. 1313–1320, 2006.

[7] P.-Y. Oudeyer, F. Kaplan, and V.V. Hafner: Intrinsic motivation systems for autonomous mental development, *IEEE Trans. Evolutionary Computation*, Vol.11, No.1, pp. 265–286, 2007.

[8] P. McCracken and M. Bowling: Online discovery and learning of predictive state representations, *Advances in Neural Information Processing Systems*, Vol. 18, pp. 875–882, 2006.

[9] J. McCarthy and P.J. Hayes: Some philosophical problems from the standpoint of artificial intelligence, *Machine Intelligence*, Vol.4, pp. 463–502, 1969.

[10] D. Dennett, Cognitive Wheels: The frame problem of AI, *The Philosophy of Artificial Intelligence*, M. A. Boden, Ed., pp. 147–170, Oxford University Press, 1984.

[11] P. Johansson, L. Hall, S. Sikstrom, and A. Olsson: Failure to detect mismatches between intention and outcome in a simple decision task, *Science*, Vol. 310, pp. 116–119, 2005.

[12] R. A. Brooks: Intelligence without representation. *Artificial Intelligence*, Vol. 47, pp. 139–159, 1991.

[13] K. Shibata and K. Ito: Adaptive space reconstruction on hidden layer and knowledge transfer based on hidden-level generalization in layered neural networks, *Trans. SICE*, Vol. 43, No. 1, pp. 54–63, 2007 (in Japanese).

[14] K. Shibata and K. Ito : Reconstruction of visual sensory space on the hidden layer in a layered neural networks, *Proc. ICONIP '98*, Vol. 1, pp. 405–408, 1998.

[15] L. Saul and S. Roweis: Think globally, fit locally: Unsupervised learning of nonlinear manifolds, *Journal of Machine Learning Research*, Vol. 4, pp. 119–155, 2003.

[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams: Learning internal representation by error propagation, *Parallel Distributed Processing*, Vol. 1, pp. 318–364, MIT Press, 1986.

[17] B. Irie and M. Kawato: Acquisition of internal representation by multilayered perceptrons, *Electronics and Communications in Japan*, Vol. 74, pp. 112–118, 1991.

[18] K. Shibata and M. Sugisaka: Dynamics of a recurrent neural network acquired through the learning of a context-based attention task, *Artificial Life and Robotics*, Vol. 7. No. 4, pp. 145–150, 2004.

[19] K. Shibata: Discretization of series of communication signals in noisy environment by reinforcement learning, *Proc. ICANNGA'05*, pp. 486–489, 2005.

[20] H. Utsunomiya and K. Shibata: Contextual behavior and internal representations acquired by reinforcement learning with a recurrent neural network in a continuous state and action space, *Proc. ICONIP2008, LNCS*, Springer-Verlag, 2009 (to appear).

[21] K. Shibata: Spatial abstraction and knowledge transfer in reinforcement learning using a multi-layer neural network, *Proc. Fifth Int'l Conf. Development and Learning*, 36, 2006.

[22] J. A. Boyan and A. W. Moore: Generalization in reinforcement learning, *Advances in Neural Information Processing Systems*, Vol. 7, pp. 370–376, The MIT Press, 1995.

[23] R.S. Sutton: Generalization in reinforcement learning: Successful examples using sparse coarse coding, *Advances in Neural Information Processing Systems*, Vol. 8, pp. 1038–1044, MIT Press, 1996.

[24] J. Moody and C.J. Darken: Fast learning in networks of locally-tuned processing units, *Neural Computation*, Vol. 1, pp. 281–294, 1989.

[25] A. Maehara, M. Sugisaka, and K. Shibata: Reinforcement learning using gauss-sigmoid neural network, *Proc. AROB 6th*, Vol. 2, pp. 562–565, 2001.

[26] K. Shibata, M. Sugisaka, and K. Ito: Fast and stable learning in direct-vision-based reinforcement learning, *Proc. AROB 6th*, Vol. 1, pp. 200–203, 2001.

[27] C.J.C.H. Watkins: *Learning from delayed rewards*, PhD thesis, Cambridge University, Cambridge, U.K., 1989.

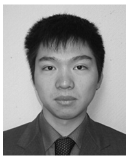
[28] A.G. Barto, R.S. Sutton, and W. Anderson: Neuronlike adaptive elements can solve difficult learning control problems, *IEEE Trans. SMC*, Vol. 13, No. 5., pp. 834–846, 1983.

[29] K. Shibata and T. Kawano: Learning of action generation from raw camera images in a real-world-like environment by simple coupling of reinforcement learning and a neural network, *Proc. ICONIP2008, LNCS*, Springer-Verlarg, 2009 (to appear).

Katsunari Shibata (Member)

He received his B.E., M.E., and D.E. degrees from The Univ. of Tokyo, Japan, in 1987, 1989, and 1997, respectively. He worked at Hitach Co. Ltd., The Univ. of Tokyo, and Tokyo Inst. of Tech. Since 2000, he worked at Oita Univ., where he is currently an Assoc. Professor. In 2005, he was a visiting professor of Univ. of Alberta. His research has focused on function emergence based on autonomous learning especially using reinforcement learning and neural networks. He is a member of SICE, IEICE, JNNS and IEEE.

Tomohiko Kawano



He received his B.E. and M.E. degrees from Oita University, Japan, in 2006 and 2008, respectively. He is now working in Kyushu Toshiba Engineering Co. Ltd. His research interests include autonomous learning system and robotics.