Hidden Representation after Reinforcement Learning of Hand Reaching Movement with Variable Link Length

Katsunari Shibata
Dept. of Electrical & Electronic Eng.
Oita University
Oita, 870-1192, Japan
Email: shibata@cc.oita-u.ac.jp

Koji Ito
Dept. of Comp. Intelli. and Sys. Sci.
Tokyo Institute of Technology
Yokohama, 226-8502, Japan
Email: ito@dis.titech.ac.jp

Abstract— **Iriki et al. reported interesting results regarding the visual receptive field of two kinds of neurons in the parietal cortex of a monkey. A monkey did a task to reach its hand or tool to a target. The receptive field of one kind of neuron was enlarged when the monkey used the tool grasped by its hand. The receptive field of the other type of neurons moved together with its hand even though the hand was hidden under an opaque plate. They discussed those results in relation to high-order cognitive functions such as body image and symbolization[1]-[3].**

In this paper, a hypothesis is posited that these neurons contribute to generate the critic output (state evaluation in a given task) and are obtained through reinforcement learning. Thereby, tool use is considered to be the change of link length for simplicity; a layered neural network learns hand reaching by a manipulator based on reinforcement learning. Inputs of the network are visual sensory signals and the state of the manipulator. Outputs are the critic and joint torques as the actor. After learning, the manipulator came to move its hand toward the target on the visual sensor when the target was located within the hand's reach. Both types of neurons observed in experiments of Iriki et al. were found in the hidden layer of the neural network.

I. Introduction

Hand reaching to some object is a primitive movement for humans and monkeys. It has been well investigated for analysis of motion learning in humans. The shape of the human hand-reaching path is known to be almost a straight line; also, the associated speed profile is bell-shaped with a single peak in the case of short unconstrained horizontal movements. There are also some exceptions in the case of long-distance reaching[4]. "Minimum motion-command-change" criteria and so forth have been proposed to explain such hand trajectories as an optimization problem [5][4][6]. "Feedback error learning" has also been proposed to control the arm to follow the computed trajectory[7].

On the other hand, the authors showed that a neural network, whose inputs are visual sensory signals and the state of a manipulator and whose outputs are joint torques, can learn the hand reaching movement of a manipulator by reinforcement learning[8]. Hand dynamics were considered and no preprocessing of the visual sensory signals was executed. It is unnecessary to compute the trajectory explicitly in this model; therefore, the iterative computation to generate it is also unnecessary. The obtained hand path was almost a straight line and its speed profile was nearly bell-shaped, but not as similar to the human's as that derived from the optimization-based path planning model mentioned above. However, it can be considered that the path is obtained by learning of the entire process from sensors to motors without any knowledge of task and arm dynamics under insufficient simulation setups, such as low visual sensor resolution.

Iriki et al. reported some interesting results concerning the visual receptive field of some neurons in the parietal cortex during hand reaching tasks by a monkey using a tool; the report presented them in relation to high-order cognitive functions. Details are described in the next section. In this paper, the hypothesis is that these neurons contribute to generating critic output that represents state evaluation in a given task, and the neurons are obtained through reinforcement learning. Here, for simplicity, tool use is considered as the change of link length; a layered neural network learns the hand-reaching task through reinforcement learning. The neural network is analyzed after learning. We attempt to explain acquisition of high-order brain function by reinforcement learning using a neural network.

Conventionally, reinforcement learning has been used as the learning for motion planning: in other words, control in its wider sense. Many studies use a neural network in reinforcement learning, but their purpose is to realize continuous and non-linear state-action mapping. On the other hand, the authors believe that it can constitute learning for the whole process from sensors to motors, including recognition, attention, memory, and so forth. By training a neural network based on reinforcement learning, the whole process can be obtained purposively, adaptively and in harmony without being divided into some functional modules. This is expected to result in real intelligence that bridges the gap separating humans and modern robots[9][10]. One aspect of this research is to explain the process to show this ability.

The possibility that reinforcement learning is done in the basal ganglia in the cerebrum has been shown[11][12]. That is based on the above idea that reinforcement learning is learning for motion planning. On the other hand, Shidara et al. showed that some neurons in anterior cingulate in the frontal lobe activate in relation to reward expectancy [13]. The authors want to show the possibility that reinforcement learning is utilized for learning of a variety of functions in our living creatures, and contributes to learning of other areas in the brain.

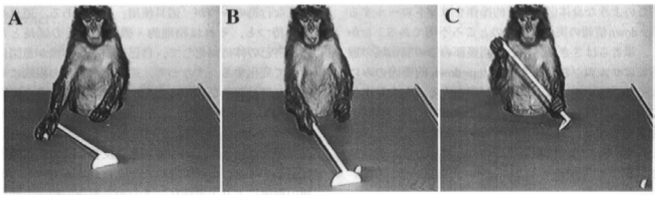


Fig. 1. Experiment of Iriki et al. This figure is copied from [2]. ©1998 by Igaku Shoin.

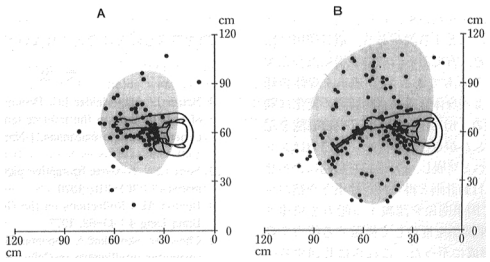


Fig. 2. Visual receptive field modification of a postcentral neuron during tool use. Dots represent positions where the neuron outputted one spike when the object scanned on the plane. Reachable areas are shaded. This figure is copied from [2]. ©1998 by Igaku Shoin.

II. Visual Receptive Field of Some Neurons in Parietal Cortex

A. Experiment by Iriki et al.

Iriki et al. trained a monkey to get food using a rake as a tool as shown in Fig. 1. After learning, when the food was just close to the monkey, the monkey got the food by hand (A). When the food was located out of the hand's reach, the monkey used the rake to get the food (B). However, when the food was located at a place where the monkey could not reach it even with the rake, it did not try to get it (C). They observed some neurons in the monkey parietal cortex. Just after using the rake, the visual receptive field of a portion of the neurons expanded from the hand-reaching area to the range where the tool could reach, as shown in Fig. 2 [1][2]. These neurons have been known to be bimodal neurons; they are activated either by somato-sensory or visuo-sensory signals.

Iriki et al. also observed another type of neurons in the same area whose receptive field moved together with the hand. When the monkey used a tool, the receptive field was formed around the tool[1][2]. Furthermore, Obayashi et al. showed that even though the hand was hidden under an opaque plate, the receptive field of such neurons still moved together with the hand as shown in Fig. 3[3]. These neurons are also bimodal neurons. They linked these results to the change of the body image and symbolic representation, and mentioned high-order cognitive functions.

B. Interpretation based on Reinforcement Learning

Though the above neurons were interpreted in relation to high-order cognitive functions, it was not mentioned how such neurons are created. However, they can be explained clearly if one attempts interpretation based on reinforcement learning.

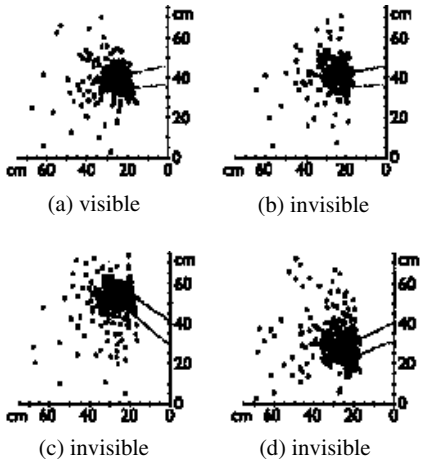


Fig. 3. The receptive field of another postcentral bimodal neuron when a opaque plate hides the hand: (a) shows the result for the case with no plate; (b),(c) and (d) show results of the invisible hand case. Dots represent the probe positions in the horizontal plane which drove the neuron to fire. This figure is copied from [3]. ©2000 by Lippincott Williams & Wilkins.

The former type of neurons in the experiment in Iriki et al. represents whether the monkey can get the food or not. In reinforcement learning, the critic output represents the state evaluation of the time to get the reward in a single reward task. When the food is located within a range where it can be reached with the hand or rake, the monkey can get the food quickly; when the food is located out of the range, the monkey cannot get the food. Therefore, the critic output should change drastically at the boundary determining whether the food can be obtained or not. The critic output is computed from sensory inputs and formed through many experiences. In other words, through many experiences, the monkey became able to judge from visuo-sensory signals whether the food was attainable. Accordingly, the neurons can be inferred to contribute to critic output generation; such neurons are expected to be obtained through reinforcement learning.

If the food is located close to the monkey, it can get the food immediately; therefore, the latter type of neuron represents whether or not the monkey can get the food immediately. The critic output becomes larger when the food is closer. Accordingly, this type of neuron can also be inferred to contribute to critic output generation. This explanation also clarifies that the receptive field expanded around the tool when the monkey used it; also, neurons activate around the hand even though the hand is hidden under the opaque plate.

III. Reinforcement Learning

In this paper, actor-critic architecture[14] is employed and implemented in a four-layered neural network as shown in Fig. 4. There are three outputs: one for the critic and two for the actor.

Temporal smoothing, or TS, learning is employed for learning of the critic [15][9]. It is very similar to temporal difference, or TD, learning; its details can be found in [15][9].

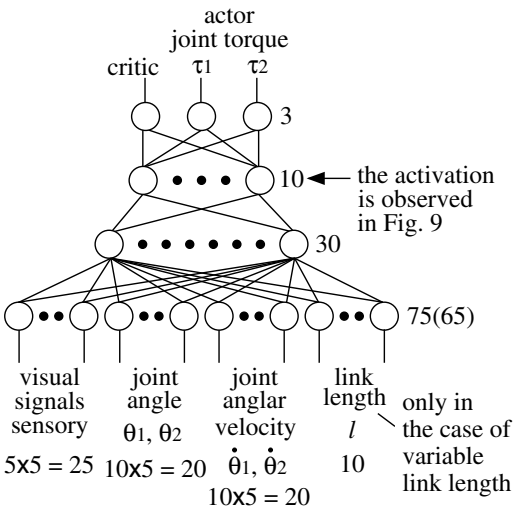


Fig. 4. The neural network structure.

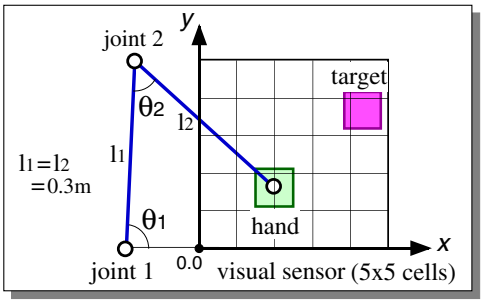


Fig. 5. The 2-link arm robot hand-reaching task.

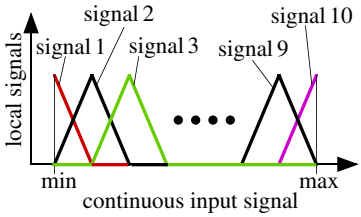


Fig. 6. Localization of a continuous input signal into 10 local signals.

The training signal for the critic p_s is computed as

$$p_s(s_{t-1}) = p(s_t) - (P_{max} - P_{min})/N_{max}, \quad (1)$$

where p is the critic output, s is the state (sensory inputs), and P_{max} and P_{min} are the maximum and minimum of the ideal critic value range; here, $P_{max} = 0.4$ and $P_{min} = -0.4$ because the value range of sigmoid function employed for each neuron's output function is from -0.5 to 0.5. N_{max} is the maximum number of time steps to the goal; it decays gradually for adaptation as

$$N_{max}[i] = \begin{cases} N[i] & if N[i] > \lambda N_{max}[i-1] \\ \lambda N_{max}[i-1] & otherwise. \end{cases} \quad (2)$$

The slope of the critic along with the time axis is trained to be a constant P_{range}/N_{max}. Accordingly, the critic output changes as a straight line in TS learning, while it changes as an exponential curve in TD learning. They are the same at the point that the critic output monotonically increases when the reward is not given. $(P_{max} - P_{min})/N_{max}$ serves as a discount factor in TD learning, but changes adaptively according to learning performance.

The actor is trained to gain more critic value. The training signals for the actor $\mathbf{m_s}$ are

$$\mathbf{m}_s(t-1) = \mathbf{m}(t-1) + \mathbf{rnd}(t-1)\{p(t) - p(t-1)\}, \quad (3)$$

where $\mathbf{m}$ is the actor output vector and $\mathbf{rnd}$ is a random vector that is added to $\mathbf{m}$ as a trial and error factor for actual motion.

IV. Simulation

A. Basic Task Setting

Here, a two-link arm as shown in Fig. 5 is supposed; the task is to learn the hand reaching movement to reach the object on the visual sensor. The visual sensor comprises $5 \times 5 = 25$ sensory cells; the receptive field of each cell is a non-overlapping square. Cell output is the area ratio occupied by a projected object. Size of the hand and target are given to be identical to one sensory cell. Hand and target images cannot be distinguished on the visual sensor. Each joint angle and angular velocity is a continuous signal, but is localized into 10 signals as shown in Fig. 6 to facilitate learning of non-linear mappings [16][9][17]. As for joint angular velocities, to achieve higher resolution around 0, the continuous value is transformed by a sigmoid function at first, then localized into 10 signals as well. In total, 65 signals comprise the layered neural network inputs.

The output function of each neuron is a sigmoid function with a value range from -0.5 to 0.5. The two actor outputs are used as torques for joint 1 and joint 2, respectively, after linear transformation to the range in Table I. The critic output is used with no transformations. The network has two hidden layers; also, the lower layer has 30 hidden neurons and the upper layer has 10.

Initial hand and target locations are decided randomly at each trial. When the hand is overlapped with the target and the hand tangential velocity is less than $0.23m/s$, the critic output is trained to be 0.4 as a reward. When one joint angle becomes less than 0 degree or the joint 1 angle is more than 90 degree, the trial is stopped and the critic output is trained to be -0.4 as a penalty.

At the early phase of learning, the target is located close to the hand; then initial distance from the hand is gradually increased according to learning performance. Concretely, learning comprises 128 stages. In the first 127 stages, the hand is located randomly in the range where the whole target can be viewed on the visual sensor. The target is located randomly within the range where the distance between the hand and the target in either the x or y direction is less than

$stage/127 \times 0.3m$; also, the whole target can be seen on the visual sensor. In the last stage, the hand is located randomly, even outside of the visual field. When the hand reaches the target within 5.1 sec successfully without any help, the learner can progress one stage. If one joint exceeds the limit, even if the target is located within the hand's reach, the target is moved to the hand gradually in the following two trials. If the hand cannot reach the target within $N_{max} \times 1.5$ five times, even if the target is located within the range where the hand can reach, the target is also moved to the hand gradually in the following trial.

B. Arm Dynamics

Arm dynamics are identical to [4]; they are

$$\begin{aligned}\tau_1 &= (I_1 + I_2 + 2M_2 l_1 s_2 cos\theta_2 + M_2 (l_1)^2 \ddot{\theta}_1 \\ &\quad +(I_2 + M_2 l_1 s_2 cos\theta_2)\ddot{\theta}_2 \\ &\quad -M_2 l_1 s_2 (2\dot{\theta}_1 + \dot{\theta}_2)\dot{\theta}_2 sin\theta_2 + B_1 \dot{\theta}_1 \qquad (4) \\ \tau_2 &= (I_2 + M_2 l_1 s_2 cos\theta_2)\ddot{\theta}_1 + I_2 \ddot{\theta}_2 \\ &\quad +M_2 l_1 s_2 (\dot{\theta}_1)^2 sin\theta_2 + B_2 \dot{\theta}_2 \qquad (5)\end{aligned}$$

where τ_i represents torque for joint i, and M_i, l_i, s_i, I_i are mass, length, distance between a joint and center of gravity, and inertia of the link i, respectively. If joint angle 2 exceeds 180 degree, the angle is fixed at 180 degree and dynamics are computed as one link. Each parameter is set as shown in Table I. The differential equation is solved numerically by the Runge-Kutta method with sampling time of 0.02 sec.

TABLE I

PARAMETERS USED IN THE DYNAMIC ARM MODEL.

Parameter		link1	link2
mass	M_i (kg)	2.0	
length	l_i (m)	0.3	
center of mass	s_i (m)	$l_i/2$	
rotary inertia	I_i (kg m^2)	$M_i * l_i^2 / 3.0$	
viscosity	B_i (kg m^2/s)	0.4	0.2
maximum torque	$\tau_{max\,i}$ (N m)	4.0	2.0

C. Variable Link Length

At first, a simulation in which the tool use is considered as variable link length is introduced. The length of the second link is varied continuously between 0.3 m and 0.45 m, and is decided randomly at every trial. The link length variable l $(0.0 \leq l \leq 1.0)$ is appended to the inputs of the network after localizing into 10 signals as shown in Fig. 6. The arm and visual sensor are set as Fig. 7. When the link length is the shortest, the hand cannot reach the target on the right side of the visual field. Even when the link length is the longest, the hand cannot reach the target at the upper right corner. Two broken lines in the figure indicate the boundary of the target center whether the hand can reach or not for the case of the shortest and longest link length respectively.

Some examples of hand paths after learning are shown in Fig. 8 for cases of the shortest link [Fig. 8(a)] and the longest link [Fig. 8(b)]. When the target is located within hand's reach, the hand moves toward the target and reaches it [path (1)(4) in Fig. 8 (a) and path (2)(5) in Fig. 8(b)]. While the hand moves to the other direction when it cannot reach the target [path (2)(5) in Fig. 8(a) and path (3) in Fig. 8(b)]. Even though the hand cannot reach the target with the shortest hand [path (2)(5) in Fig. 8(a)], it was able to reach the target at the same location with the longest link [path (2)(5) in Fig. 8(b)]. In [1][2], when the monkey obviously could not reach the target, it did not move its hand. Even in this simulation, the system seems to know whether the hand can reach or not before the hand moves. If energy consumption cost is introduced, useless hand motion is expected to be suppressed.

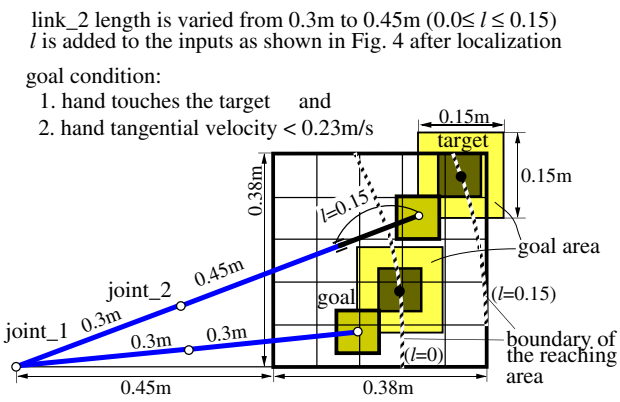
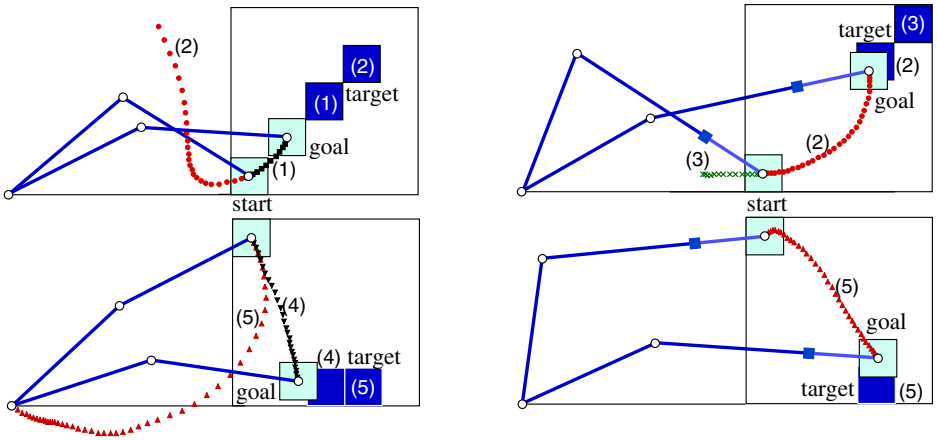


Fig. 7. The task setting in the case of variable link length. This figure is drawn for the case that the joint_2 is stretched ($\theta_2 = \pi$), i.e. the hand is farthest from the joint_1.

Figure 9 shows output distributions of the critic (row 1) and two of the upper hidden neurons (rows 2 and 3) as functions of target location for each of the shortest (two left-most columns) and longest (two right-most columns) link lengths and for each of the lower left (columns 1 and 3) and upper left (columns 2 and 4) hand locations. Joint angular velocities are set to 0.0. The sign of hidden neuron 1's output value is shown inversely for clarity, but the represented information is not different. The critic output on the first row of the figure is small beyond the boundary of reaching area that is indicated by the broken line; it is independent of hand position and length. In the other area, the critic output increases when distance between the hand and target decreases. Hidden neuron 1's output on the second row changes its value almost discretely around the reaching area boundary; it does not depend on the distance between the hand and target. It can be inferred that this neuron is coding whether the hand can reach the target or not. Such a neuron can be found in the experiment of Iriki et al. [1][2]. Four of 10 upper hidden neurons show such distribution. Hidden neuron 2 on the bottom row is similar to the critic output, but the output depends more on the distance between the hand and target. It can be inferred that this neuron corresponds to the neuron whose receptive field is formed around the hand and moves together with it. It can also explain that during tool use, the receptive field expands around the tool[1][2].



(a) when the link length is the shortest (l=0.0) (b) when the link length is the longest (l=0.15)

Fig. 8. Difference of hand trajectories depending on link length. A number in parentheses indicates the target position and the corresponding path.

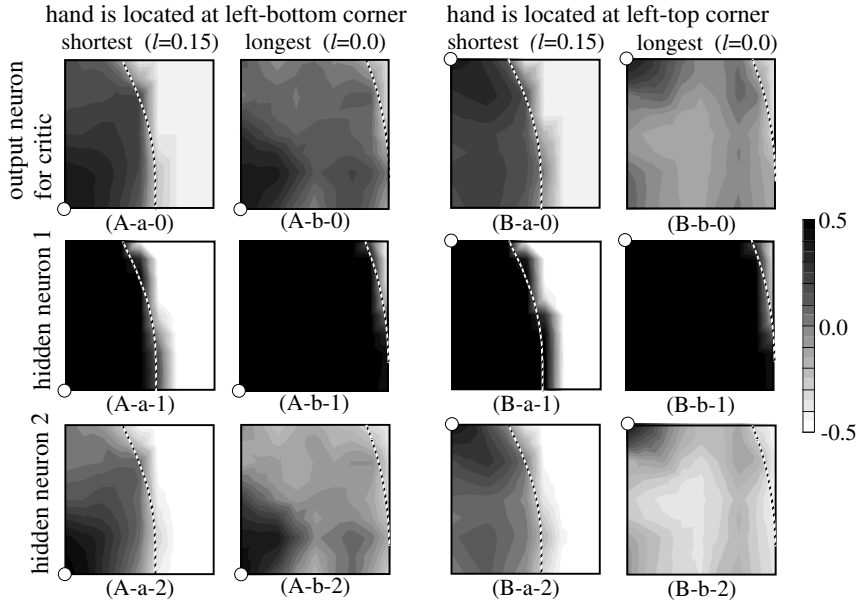


Fig. 9. Distribution of critic output and two hidden neurons' outputs for each of two hand positions and for each of two link lengths as a function of target location in the visual sensor. The small circles indicate hand position.

D. Invisible Hand

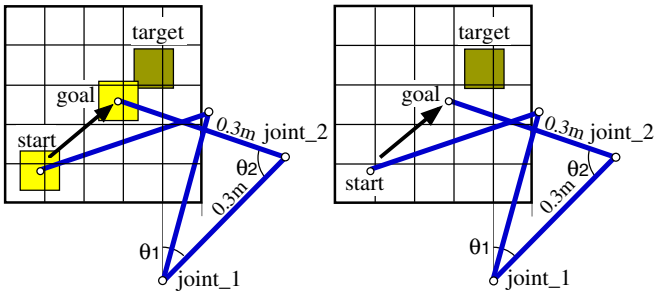
Here, it is supposed that the hand image sometimes disappears even if the hand is located under the visual sensor as shown in Fig. 10; also, learning of the reaching task is done. Random numbers determine whether the hand image appears or not with probability of 0.5, but information as to whether the hand is visible or not is not given. Link length is fixed in this simulation.

After learning, the hand can reach the target independent of whether the hand is visible or not with few exceptions. Figure 11 shows an example of the hand path after learning for both visible and invisible hand cases. The paths are slightly different, but the hand can reach the target in each case. In most cases, these paths do not differ much, but in some cases they differ more.

Figure 12 shows the distribution of one hidden neuron's activation for two cases of hand positions. It can be seen that the receptive field moves together with its hand, independent of whether the hand is visible or not; distribution of activation is almost identical for these two cases.

E. Discussion

In this model, there remain many characteristics whose validity should be considered, such as simplification of tool use into variable link length, the form of input signals, the model of each neuron, and so forth. However, even if they were changed, the interpretation described in subsection II.B may still be valid, and similar results as the above are expected to be obtained.



(a) when the hand is visible (b) when the hand is invisible

Fig. 10. Task setting when the hand image disappears out of the visual field even if the hand is located below the sensor.

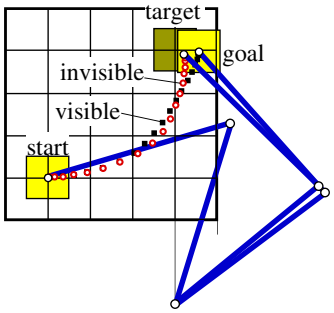
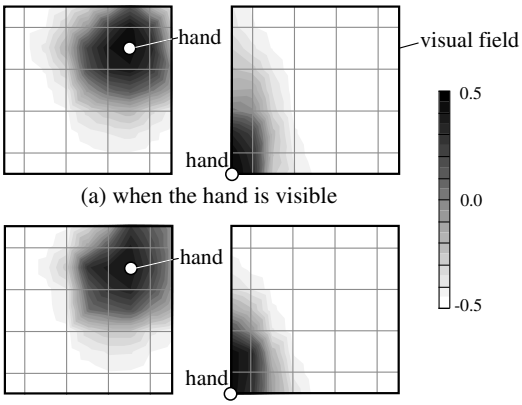


Fig. 11. Hand trajectory for both visible and invisible hand cases.



(a) when the hand is visible

(b) when the hand is invisible

Fig. 12. Distribution of a hidden neuron's output as a function of target location for visible and invisible hand cases. The small circles indicate hand position.

V. Conclusion

The hand-reaching task is trained by a combination of reinforcement learning and neural network for cases of variable link length and invisible hand. Two types of hidden neurons were found. The first one represents whether the hand can reach the target or not. The receptive field of the other one was formed around the hand, and moves together with it. Further, even if the visual sensor was not able to catch the hand image, the receptive field was still formed around the hand position. These results match experimental results using a monkey that was introduced as a neuron activation to represent high-order cognitive functions in the brain, such as body image and symbolization, by Iriki et al. We infer that we can show the possibility that a combination of reinforcement learning and neural networks can explain high-order brain functions.

Acknowledgment

I would like to thank Dr. Iriki and Dr. Obayashi for their interesting work and kind permission to use their figures (Figs. 1, 2 and 3). This research was partially supported by the Sci. Res. Found. of the Japanese Ministry of Edu., Sci., Sports and Culture (#10450165, #13780295) and by "The Japan Society for the Promotion of Science" as "Biologically Inspired Adaptive Systems" (JSPS-RFTF96I00105) in "Res. for the Future Program".

References

[1] Iriki, A., Tanaka, M. & Iwamura, Y. (1996) Coding of modified body schema during tool use by macaque postcentral neurons, *Neuroreport*, **7**: 2323–2330

[2] Ikiri, A. (1998), Monkey tool use and the body image, *Shinkei Kenkyu no Shinpo (Advances in Neurological Sciences)*, **42** (1): 98-105 (in Japanese)

[3] Obayashi, S., Tanaka, M. & Iriki, A. (2000) Subjective image of invisible hand coded by monkey intraparietal neurons, *Neuroreport*, **11** (16): 3499–3505

[4] Uno, Y., Kawato, M. & Suzuki, R. (1989) Formation and control of optimal trajectory in human multijoint arm movement, *Biol. Cybern.*, **61**: 89–101

[5] Flash, T. & Hogan, N. (1985) The coordination of arm movements: An experimentally confirmed mathematical model, *J. Neurosci.*, **5**: 1688-1703

[6] Kawato, M. (1992) Optimization schemes and neural network models for formation and control of coordinated movement, *Attention and Performance XIV*, MIT Press: 821–849

[7] Kawato, M., Furukawa, K. & Suzuki, R. (1987) A hierarchical neural-network model for control and learning of voluntary movement, *Biol. Cybern.*, **57**: 169–185

[8] Shibata, K., Sugisaka, M. & Ito, K. (2000) Hand reaching movement acquired through reinforcement learning, *Proc. of The 2000 KACC (Korea Automatic Cont. Conf.)*, 90rd (CD-ROM)

[9] Shibata, K., Ito, K. & Okabe, Y. (2001) Direct-Vision-Based Reinforcement Learning Using a Layered Neural Network - For the Whole Process from Sensors to Motors -, *Trans. of SICE*, **37**(2): 168–177 (in Japanese).

[10] Shibata, K. (2001) Reinforcement Learning and Robot's Intelligence. - Can the intelligence be formed by carrots-and-sticks ? -, *Proc. of Annual Conf. of JSAI*, panel discussion, 2A1-05 (in Japanese).

[11] Schultz, W., et al. (1993) Response of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning of delayed response task, *J. Neurosci.*, **13**: 900-913.

[12] Houk, J. C., Adams, J. L. & Barto, A. G. (1994) A model of how the basal ganglia generate and use neural signals that predict reinforcement, In Houk, J. C., Davis, J. L. & Beiser, D. G., editors, *Models of Information Processing in the Basal Ganglia*, MIT Press, 249–270.

[13] Shidara, M. & Richmond, B. J. (2002) Anterior cingulate: Single neuronal signals related to degree of reward expectancy, *Science*, **296**:1709–1711.

[14] A. G. Barto, R. S. Sutton & C. W. Anderson (1983) Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems, *IEEE Trans. SMC*, **13**: 835-846

[15] Shibata, K., Ito, K. & Okabe, Y. (1998) Direct-Vision-Based Reinforcement Learning in "Going to an Target"Task with an Obstacle and with a Variety of Target Sizes, *Proc. of Int'l. Conf. on Neu. Net. & Their Appli.(NEURAP) '98*: 95–102.

[16] Shibata, K., Sugisaka, M. & Ito, K. (2001) Fast and Stable Learning in Direct-Vision-Based Reinforcement Learning, *Proc. of Artificial Life and Robotics (AROB)'01*, **1**: 200–203.

[17] Shibata, K. & Ito, K. (1999) Gauss-Sigmoid Neural Network, *Proc. of Int'l Joint Conf. on Neu. Net. (IJCNN)'99*, #747 (CD-ROM).