Learning of Reaching a Colored Object Based on Direct-Vision-Based Reinforcement Learning and Acquired Internal Representation

Kazuyoshi Yuki Masanori Sugisaka Katsunari Shibata

Dept. of Electrical and Electronic Engineering,Oita University.
700 Dannoharu Oita 870-1192 Japan. shibata@cc.oita-u.ac.jp

Abstract

In this paper, it is shown that a mobile robot with a color line sensor could learn to reach a colored target object for which a reward was given, based on Direct-Vision-based Reinforcement Learning. In the learning, non-processed color sensor signals were the input of the neural network. On the hidden layer of the network, global information of the target location was represented by integrating the local sensor signals. The representation was almost the same between the colors of the rewarded target objects, even though the input signals differ completely by color even when the target location is the same.

1. Introduction

In recent years, reinforcement learning (RL) has been focused on as an autonomous learning for autonomous robots. By using a neural network in RL, continuous state and continuous action can be handled. The method has been applied to some control problems [1] [2].

On the other hand, the authors have proposed Direct-Vision-Based RL (D-V-B RL)[3][4] in which non-processed visual sensor signals are the input of a neural network and the network is trained based on RL when a robot learns its actions autonomously. Through the learning, continuous state space is formed in the hidden layers, and not only planning the actions but also the whole process from sensors to motors can be learned in harmony.

Then, a task in which a real mobile robot with a monochrome visual sensor reached and pushed a black box made of paper was employed, and visual sensor signals were inputted to a neural network directly. It was confirmed that the robot could learn to approach and push the box without giving any knowledge about the task and image processing in advance[4]. It was also shown that even if each input signal represents only local information, global spatial information of the object position is represented in the hidden layer of the neural network by integrating the signals.

In this paper, color visual sensor signals are used as input. At first, it is verified whether the robot can learn to reach the object as well as the case of monochrome sensor or not. Moreover, it is observed how the object position information and the color information are represented in the hidden layer.

It has been said that color information and position information are processed separately in the brain[5]. The possibility that the acquisition of the separate processing can be accounted for based on the combination of RL and neural network is also considered.

2. Direct-Vision-Based Reinforcement Learning (D-V-B RL)[3][4]

Fig.1 shows the concept of D-V-B RL. Here, actor-critic architecture is employed, and the actor and critic are composed of one layered neural network. Visual sensor signals are inputted to the neural network directly. The motor commands are generated from the actor output of the neural network. The training signal is generated based on RL, and the network is trained by the signal.

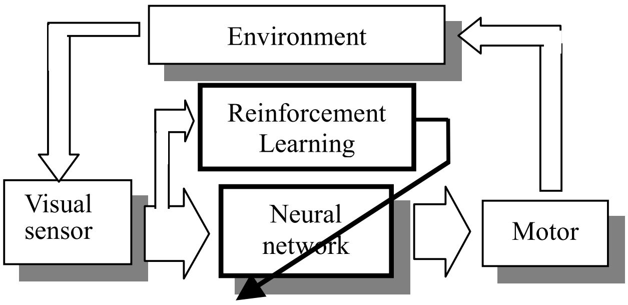


Fig.1 Direct-Vision-Based Reinforcement Learning.

TD (Temporal Difference) learning is applied for the critic learning. TD error is defined as

$$\hat{r}_t = r_t + \gamma P(\mathbf{x}_t) - P(\mathbf{x}_{t-1}) \qquad (1)$$

where γ: a discount factor, r_t: a reward, P: the critic output and $\mathbf{x}_t$: observation vector. TD error is calculated from the present state value and the previous one. The previous critic output $P(\mathbf{x}_{t-1})$ is trained by the training signal as

$$P_{s,t-1} = P(\mathbf{x}_{t-1}) + \hat{r}_t = r_t + \gamma P(\mathbf{x}_t), \qquad (2)$$

On the other hand, the motor command vector of the robot is the sum of the actor output vector $\mathbf{a}(\mathbf{x}_t)$ and a random number vector $\mathbf{rnd}_t$ as trial and error factors. The actor output vector $\mathbf{a}(\mathbf{x}_{t-1})$ is trained by the training signal as

$$\mathbf{a}_{s,t-1} = \mathbf{a}(\mathbf{x}_{t-1}) + \hat{r}_t\,\mathbf{rnd}_{t-1}. \qquad (3)$$

The neural network is trained by Back Propagation (BP) method according to Eq. (2) and (3). In this way, when $\hat{r}_t$ is positive, $\mathbf{a}(\mathbf{x}_{t-1})$ is reinforced in the direction of $\mathbf{rnd}_{t-1}$.

3. Simulations

3.1 Task and environment

In this paper, a task is employed in which a mobile robot with a color line sensor reaches a colored target object. Fig.2 shows the simulation environment. Three colored objects, red (R), green (G), and blue (B) one, are prepared. At each trial, one of them is chosen randomly and is located randomly in front of the robot. Fig.3 shows the robot and the color line sensor that is mounted on the robot.

The robot height is 4.0, and the distance between left and right wheels are 3.8. The motor command for each wheel is continuous value from –0.3 to 0.3, and is used as the forward speed of the wheel. The color line sensor has 64 pixels, and the view angle is 180 degree. The sensor is installed a little backward from the robot center. The image is determined depending on the object color and the relative object position from the robot. The number of input signals is 192, since each of 64 pixels contains the information of the three colors, R, G and B. Each signal is continuous and ranges from 0 to 1. The background of the image is always black that means that the pixel values for the three colors are all 0.0. Therefore, when the color that a sensor cell detects is different from the object color, the pixel value is always 0 not depending on the object position. For simple analysis, only the primary colors are used as the object color.

Here, 4-layer neural network with 192 input units, 30 and 10 hidden units, and 3 output units was used. One of the outputs was for critic, and the other two were for actor. The output function of each neuron was a sigmoid function that ranges from –0.5 to 0.5. Since the critic ranges from 0 to 1 here, 0.5 was added to the critic output of the neural network. Each element of the random vector **rnd** was a cubed uniform random number that ranges from -1.0 to 1.0. Since the range of each motor command is from –0.3 to 0.3, the actor output was multiplied by 0.6. All the initial connection weights between the upper hidden layer and the output layer were 0. The other initial weights were small random numbers whose range was from –0.5 to 0.5. The discount factor γ was 0.99.

The goal of the robot is that the robot center passes through the object whose diameter is 2.0. When the robot reached the blue and green object, a reward is given. The training signal of the critic $P_{s,t}$ is 0.9 as a reward according to Eq. (2) with $r_t = 0.9$ and $P_t = 0.0$. When the robot reached the red object, the critic is not trained. Moreover, when the object goes to the back of the robot without passing through the robot, the training signal of the critic $P_{s,t}$ is 0.1 as a penalty. Otherwise, the critic is trained according to Eq. (2) with $r_t = 0$.

At every trial, the robot was returned to a starting position initially, and the object color and position are on determined randomly. At the early phase of learning, the object position range is set near from and in front of the robot. As learning progresses, the range is gradually extended to the area far from the robot and the area at the side of the robot. The maximum distance from the robot to the object is 30 and the maximum angle from the front of the robot to the object is 90 degree.

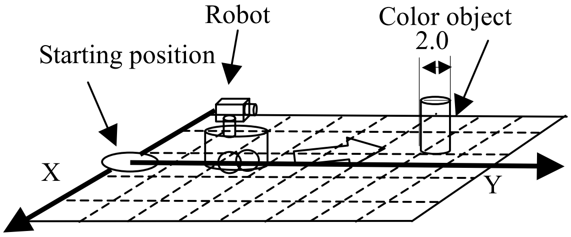


Fig.2 Simulation environment

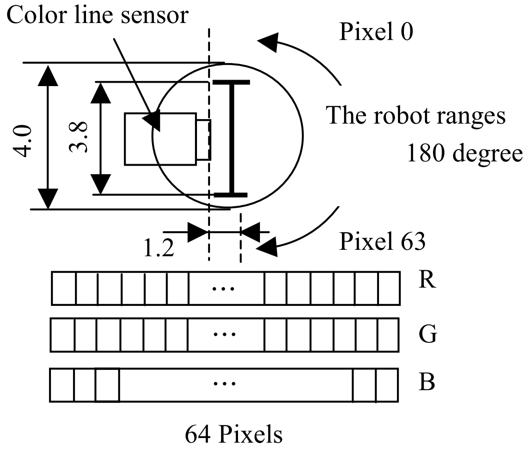


Fig.3 The robot and color line sensor

3.2 Results

Fig.4 shows the loci of the robot for some object locations for each object color after 30000 trials of learning. Fig.5 shows the distribution of the state value (critic output) as a function of the object position after learning. The robot is fixed at the position (0, 0) in this figure. The figure is drawn from the values for 1474 object positions on a lattice with a constant interval of 1. In case of the green or the blue object, the state value became larger gradually as the object was closer to the robot. It can be seen that the robot could reach the object not depending on the object position. It can be also seen that the loci and the state value were similar between the green and blue objects.

On the other hand, in case of the red object, though the robot was sometimes punished, it never got a reward. Therefore, the state value of the red object was small not depending on the object position. The action to reach the red object was not learned.

3.3 The analysis of the hidden representation

It was observed how the color information and the position information were represented in the hidden layer through learning. Using the weight after 30000 trials of learning, the output distribution of each neuron in the hidden layer that is closer to the output layer was observed.

An agreement index between two colors for each hidden neuron is defined to know how much the output of the neuron agrees between two colors. The agreement index is defined as

$$M_{color1,color2} = \sqrt{\sum_{p \in P} (h_{color1,p} - h_{color2,p})^2 / N_P} \,/\, \sigma_{color1,color2} \qquad (3)$$

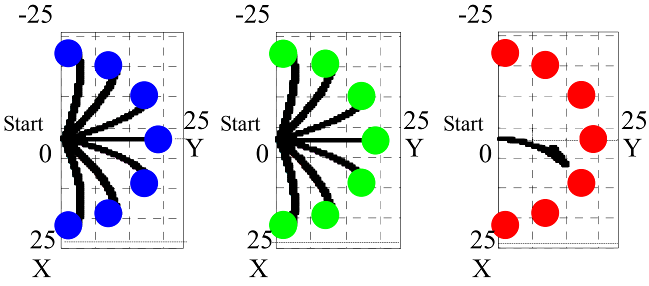
where $h_{color,p}$ is the output of the hidden neuron when the object color is $color$ and the object position is p. P is the sample set of the object positions. N_P is the number of elements of P. Here, N_P is 1474. $\sigma_{color1,color2}$ is the standard deviation of the hidden neuron's output for both $color1$ and $color2$ objects. It should be noticed that when the hidden output agrees more between two colors, the agreement index becomes smaller.

Table 1 shows the agreement indices between red and blue and that between green and blue for each hidden neuron. The connection weight to each output neuron is also shown. Before learning, it can be seen that all the agreement indices are more than 1. However, after learning, some agreement indices between green and blue are less than 1. The underlined weight values indicate that the absolute weight value is the maximum for one of the output neurons, in other words, the hidden neuron contributes the most to the output neuron.

Fig.6 shows the output distribution of the four hidden neurons. In the first three neurons, the output distribution is very similar between the green and blue objects. On the other hand, the distribution for the red object is different from that for the other colors. The neuron 3 that contributes to the critic output represents the distance to the object when the object color is blue or green. The neuron 4 and 5 that contribute to the actor outputs classify the object location into right or left for the object colors. The distance and "right or left" are both global information that are calculated by integrating local sensor signals. On the other hand, in case of red object, the distribution is influenced by the locality of the receptive field of the sensor cells.

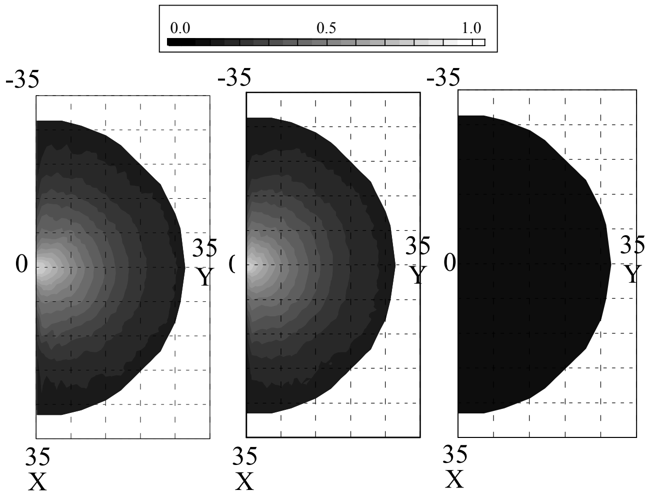
The contribution of the hidden neuron 9 to the actor1 is also large, but the agreement index between the green and blue object is not small. When the object color is green or blue, the output is almost 0.5 for all the object positions as shown in Fig.6 (d). The representation is not different actually between the blue and green objects.

From this result, it can be said that if the reward



(a) Blue object (b) Green object (c) Red object

Fig.4 The tracks of the robot after 30000 trials

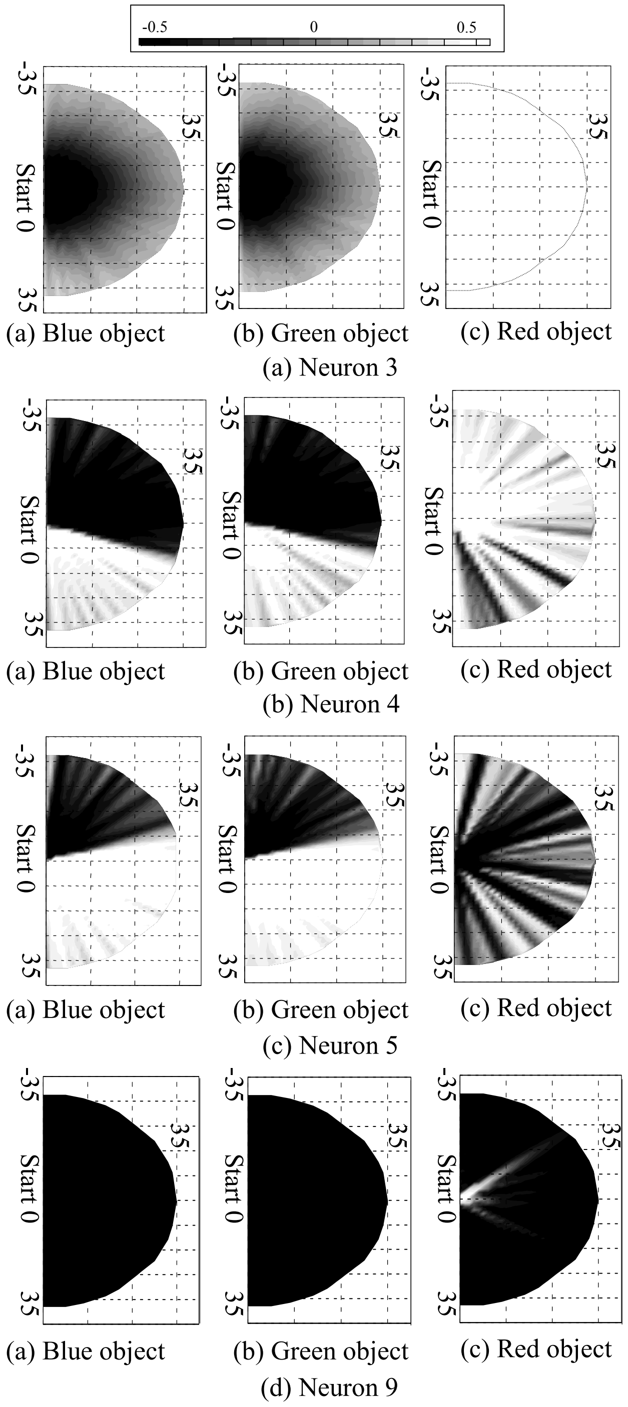


(a) Blue object (b) Green object (c) Red object

Fig.5 Distribution of the state value (critic output) as a function of the relative target location.

Table1 Agreement index and connection weight to each output neuron for each hidden neuron.

Neuron	Agreement index				Connection weight to each output neuron of each hidden neuron		
	Before learning		After learning				
	R-B	G-B	R-B	G-B	Critic1	Actor1	Actor2
1	1.37	1.41	1.14	0.59	1.02	-0.004	0.01
2	1.45	1.43	1.75	1.42	-0.93	-0.07	-0.1
3	1.49	1.49	1.88	**0.149**	**-1.67**	-0.01	0.03
4	1.44	1.48	1.67	**0.253**	0.04	0.005	**-4.4**
5	1.61	1.13	1.36	**0.274**	0.03	**4.42**	0.01
6	1.35	1.22	1.79	1.67	0.05	-0.07	-0.01
7	1.42	1.40	1.3	0.78	0.35	0.03	0.03
8	1.59	1.55	1.69	1.59	0.2	-0.05	0.01
9	1.43	1.42	1.77	1.32	0.02	**4.36**	0.07
10	1.36	1.52	1.59	0.85	0.09	-0.02	-0.01



(a) Blue object (b) Green object (c) Red object
(a) Neuron 3

(a) Blue object (b) Green object (c) Red object
(b) Neuron 4

(a) Blue object (b) Green object (c) Red object
(c) Neuron 5

(a) Blue object (b) Green object (c) Red object
(d) Neuron 9

Fig.6 The output distribution of the hidden neurons that have large connection weights to one of the output neurons.

condition is the same between two colored objects, the hidden neuron that has a strong connection weight to one of the output neurons represents the same information about object position. It has been shown that even if the input patterns are not close, the hidden representation tends to be closer if the training signals are close[6]. It is thought that the nature of the neural network causes the result in this paper.

It has been said that color information and position information are processed separately in the brain of our living things[5]. In the result of this paper, in the hidden layer, the position information not depending on the color is extracted when the reward condition is the same. In the real world, there are many situations in which the position information not depending on the object color is important. For example, the range of the object location to which a person can reach its hand does not depend on the object color. We think that the possibility that reinforcement learning contributes to process the color information and the position information separately could be suggested from this result.

4. Conclusion

The learning of the task in which a reward is given depending on the object color when a mobile robot reaches a target object was simulated. The robot learned to reach the rewarded colored object by Direct-Vision--Based Reinforcement Learning.

Moreover, it was verified in the hidden neurons that the global information of the object position not depending on the object color is represented for the rewarded colors.

Acknowledgements

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan or the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research, #13780295, #14350227 and #15300064.

References

[1] Anderson, C.W: Learning to Control an Inverted Pendulum Using Neural Networks IEEE Control System Magazine, Vol.9 pp 31-37 (1989)

[2] Morimoto, J. and Doya, K.: Acquisition of Stand-up Behavior by a Real Robot using Hierarchical Reinforcement learning, Robotics and Autonomous Systems, Vol.36,pp.37-51 (2001)

[3] K.Shibata, K.Ito and Y.Okabe Direct-Vision-Based Reinforcement Learning Using a Layered Neural Network -For the Whole Process from Sensor to Motors- ,Trans. of SICE,Vol.37, No2, p168-177 (2001) (in Japanese)

[4] K.Shibata and M.Iida: Acquisition of Box Pushing by Direct-Vision-Based Reinforcement Learning Proc.of SICE Annual Conf. 2003, 0324.Pdf, pp.1378-1383 (2003)

[5] S.Amari and K.Toyama Editor: Brain Science Encylopedia, 1.1Visual Recognition, Asakura Shoten, pp102-142 (2002) (in Japanese)

[6] K.Shibata and K.Ito: Adaptive Space Reconstruction on Hidden Layer and Hidden-level Generalization in Layered Neural Networks with Local Signal Inputs, Technical Report of IEICE NC2001-152, pp.151-158 (2002) (in Japanese)