Artif Life Robotics (2003) 7:102–106
DOI 10.1007/s10015-003-0243-y

ORIGINAL ARTICLE

Masaru Iida · Masanori Sugisaka · Katsunari Shibata

Direct-vision-based reinforcement learning in a real mobile robot

Received and accepted: March 31, 2003

Abstract It was confirmed that a real mobile robot with a simple visual sensor could learn appropriate motions to reach a target object by direct-vision-based reinforcement learning (RL). In direct-vision-based RL, raw visual sensory signals are put directly into a layered neural network, and then the neural network is trained using back propagation, with the training signal being generated by reinforcement learning. Because of the time-delay in transmitting the visual sensory signals, the actor outputs are trained by the critic output at two time-steps ahead. It was shown that a robot with a simple monochrome visual sensor can learn to reach a target object from scratch without any advance knowledge of this task by direct-vision-based RL.

Introduction

Reinforcement learning (RL) is an attractive learning method in autonomous robots, and has been utilized to obtain an appropriate mapping from state space to action space. By combining reinforcement learning and a neural network, continuous states and actions can be handled, because neural networks are able to approximate nonlinear functions with continuous input and output values. This combination has successfully been applied to nonlinear control tasks[1,2] and games.[3]

Among many types of robot sensor, a visual sensor contains the most sensory cells, thereby giving considerable information about the environment to the robot. Humans also largely rely on visual information in order to know the environment state. Previously, Asada et al.[4] applied reinforcement learning to soccer robots that were equipped with a visual sensor. For this experiment, processing of the visual sensory signals was given to the robot beforehand to divide the state space into a number of discrete states. The robot learned appropriate actions for each state by Q-learning. However, the processing of the visual sensory signals needs the knowledge of the task and intelligence actually.

On the other hand, in order to realize intelligence in a robot, we think that it is important that humans do not provide knowledge for the given tasks to the robot, and the robot obtains the function to achieve the task by itself. Based on this idea, direct-vision-based RL has been proposed.[5,6] During learning, the whole process from interpreting the sensory signals to moving the motors are computed by a layered neural network. Raw visual sensory signals are sent directly to a layered neural network, which is then trained using back propagation. Then the training signal is autonomously generated within the robot using reinforcement learning. This synthetic learning is not only for the motion planning, but also for a series of processes from sensors to motors, including recognition and other functions. It has been reported that when a robot learns the actions needed to reach a target, spatial information is adaptively represented on the hidden layer.[5,6] Moreover, it was found that learning is faster and more stable than when preprocessed information is used as inputs.[5,7] The effectiveness of direct-vision-based RL mentioned above has been confirmed only by some simulations.[5,8]

Here, it is shown that a real mobile robot with a monochrome visual sensor can learn appropriate motions from scratch without any advance knowledge in a "going to a target" task.

M. Iida · M. Sugisaka · K. Shibata (✉)
Department of Electrical and Electronic Engineering, Oita University, 700 Dannoharu, Oita 870-1192, Japan
Tel. +81-97-554-7832; Fax +81-97-554-7832
e-mail: shibata@cc.oita-u.ac.jp

This work was presented in part at the 7th International Symposium on Artificial Life and Robotics, Oita, Japan, January 16–18, 2002

Actor–critic architecture

Figure 1 depicts the concept of direct-vision-based RL. Here, an actor–critic architecture[9] has been employed. The actor, a motion command generator, and the critic, a state

evaluator, are composed of one layered neural network, as in previous simulations.[5,6] That means that both the actor and the critic use the hidden layer in common. The temporal difference (TD) method is used for critic learning. The TD error is defined as

$$\hat{r}_t = r_t + \gamma P_t - P_{t-1} \quad (1)$$

where γ is a discount factor, r_t is a reward, and P_t is the critic output. The critic output at the previous time P_{t-1} is trained by the training signal:

$$P_{s,t-1} = P_{t-1} + \hat{r}_t = r_t + \gamma P_t \quad (2)$$

where $P_{s,t-1}$ is the training signal for the critic output. The motion commands of the robot are the sum of the actor output vector $\mathbf{a}_t$ and the random number vector $\mathbf{rnd}_{t-1}$ that represents the trial and error factor. The actor output vector $\mathbf{a}_{t-1}$ is trained by the training signal:

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

Finally, the neural network is trained by back propagation according to Eqs. 2 and 3. In this way, motion commands are trained in order to gain more critic output.

Experimental system and environment

Figure 2 shows the robot and the monochrome visual sensor (Khepera and the K213 vision turret) that was used in this work. The specifications of Khepera and the K213 vision turret are as follows.

Height: 33 mm
Diameter: 55 mm
Interface with PC: RS232C (serial port)
Transmission rate: 38400 bps
Sensor cells: 64
Gradation: 256 gray scale
Visual field: 36°

This visual sensor is composed of two parts, the image perception optics and the light intensity detector optics, as shown in Fig. 2b. The light optics detect the intensity of the light around the robot, and then the image perception optics adjust the image sensory outputs according to the light intensity. Therefore, when the light intensity is not strong enough, the image perception optics attempt to compensate, resulting in all the pixel values becoming almost white. As a result, the robot is unable to distinguish bright points and dark points. Therefore, when a black target object is located just in front of, and on the right side of, the robot, the robot loses the target.

Figure 3 shows the experimental environment. The action area was 70 × 70 cm and was surrounded by a 10 cm-tall white paper wall. A fluorescent light was installed to maintain an adequate amount of light intensity. The target was 8 cm tall and 2.5 cm in diameter, and was wrapped in black paper.

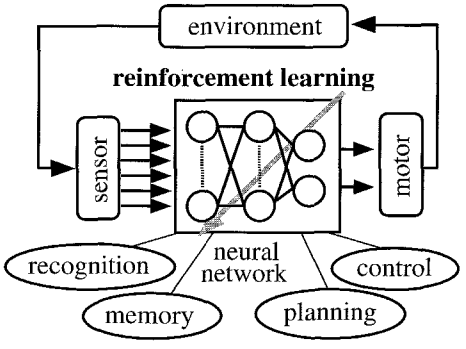


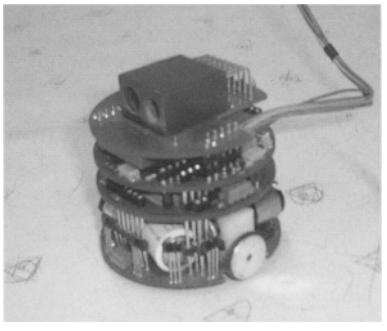
Fig. 1. The concept of direct-bision-based reinforcement learning

Application to a real robot

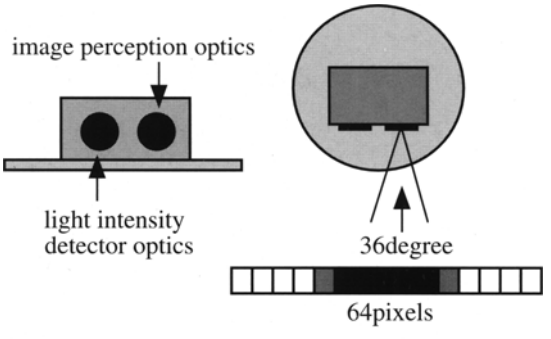
Coping with a time delay

When direct-vision-based RL is applied to a real robot, a time delay should be considered, although this does not have to be considered in simulations. The PC receives visual sensory signals from the real robot through an RS232C serial port, and its transmission rate is not fast enough. The approximate time needed to execute each process is as follows.

Transmission of visual sensory signals: 90 msec
Transmission of motion commands: 10 msec



a



b

Fig. 2. **a** Khepera and the K213 vision turret that was mounted on the Khepera. **b** K213 vision turret

Computation of neural network: less than 1 msec (for both forward computation and learning)

Considering the measurement interval of the visual sensor, the sampling time is set to 300 msec. If the motion commands are transmitted to the robot just after the transmission of the visual sensory signals, the robot continues to move according to the previous motion command during the transmission of the visual sensory signals. Then the robot location obtained from the visual sensory signals is different from the location when the next motion command is transmitted. Here, in order to reduce this influence, the visual sensory signals are transmitted just after the motion command. Figure 4 gives the timing of system events in this experiment. It can be seen that P_t is influenced by the action of $\mathbf{a}_{t-2}$ on behalf of $\mathbf{a}_{t-1}$. The critic learns according to Eq. 2, as in the simulation. On the other hand, the motion command, which is sent two steps earlier, is trained by the training signal as

$$\mathbf{a}_{s,t-2} = \mathbf{a}_{t-2} + \hat{r}_t \cdot \mathbf{rnd}_{t-2} \tag{4}$$

on behalf of the signal as in Eq. 3.

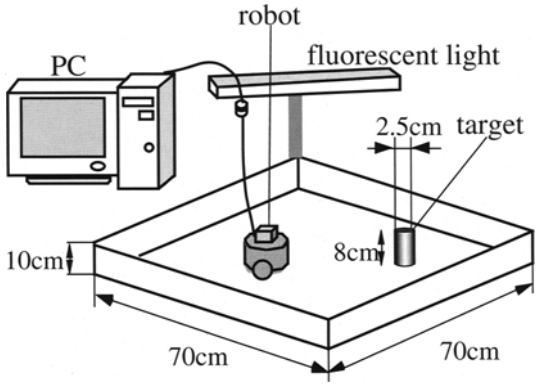


Fig. 3. Experimental environment

Discrete actions

Since the motion command for the Khepera must be an integer, the continuous motion value is discretized into a "speed" integer by Eq. 5.

$$\begin{aligned} &\text{speed}_t = (\text{int})8 \cdot (\mathbf{a}_t + \mathbf{rnd}_t) \\ &\text{If}(\text{speed}_t \leq -3) \quad \text{speed}_t = -3 \\ &\text{If}(\text{speed}_t \geq 3) \quad \text{speed}_t = 3 \end{aligned} \tag{5}$$

$-3 \leq \text{speed}_t \leq 3$, $-0.5 \leq \mathbf{a}_t \leq 0.5$, $-0.2 \leq \mathbf{rnd}_t \leq 0.2$, where speed is the motion command for the robot.

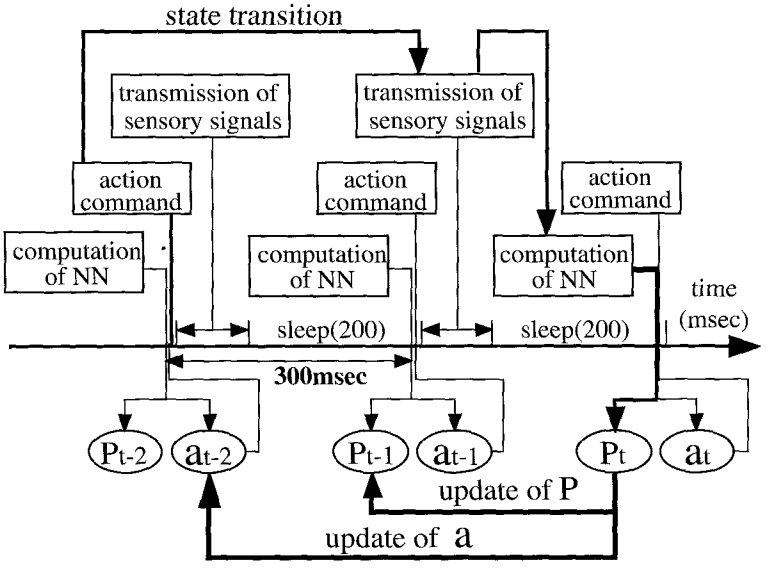
Experiment

Task

The goal of the real mobile robot with the monochrome visual sensor is to reach the target. Here, a three-layered neural network with 64 input units, 30 hidden units, and 3 output units was used. The neural network inputs were gray-scale values of the raw visual sensory signals. One of the outputs was for the critic, and the other two were for the actor. Before learning, the input-hidden connection weights were small random numbers, and all of the hidden-output connection weights were 0.0. The output function for each hidden or output neuron is a sigmoid function with an output range of from −0.5 to 0.5.

After the visual sensory signals have been transmitted, they are binarized using a boundary value of 85. The "width" of the target was defined as the number of dark pixels in the robot's view. The "center pixel" was defined as the center pixel number in the dark area. When $30 \leq$ width, and $21 \leq$ center pixel ≤ 41, the critic output is trained to be 0.9 as a reward. When the robot loses the

Fig. 4. Timing chart in a real robot



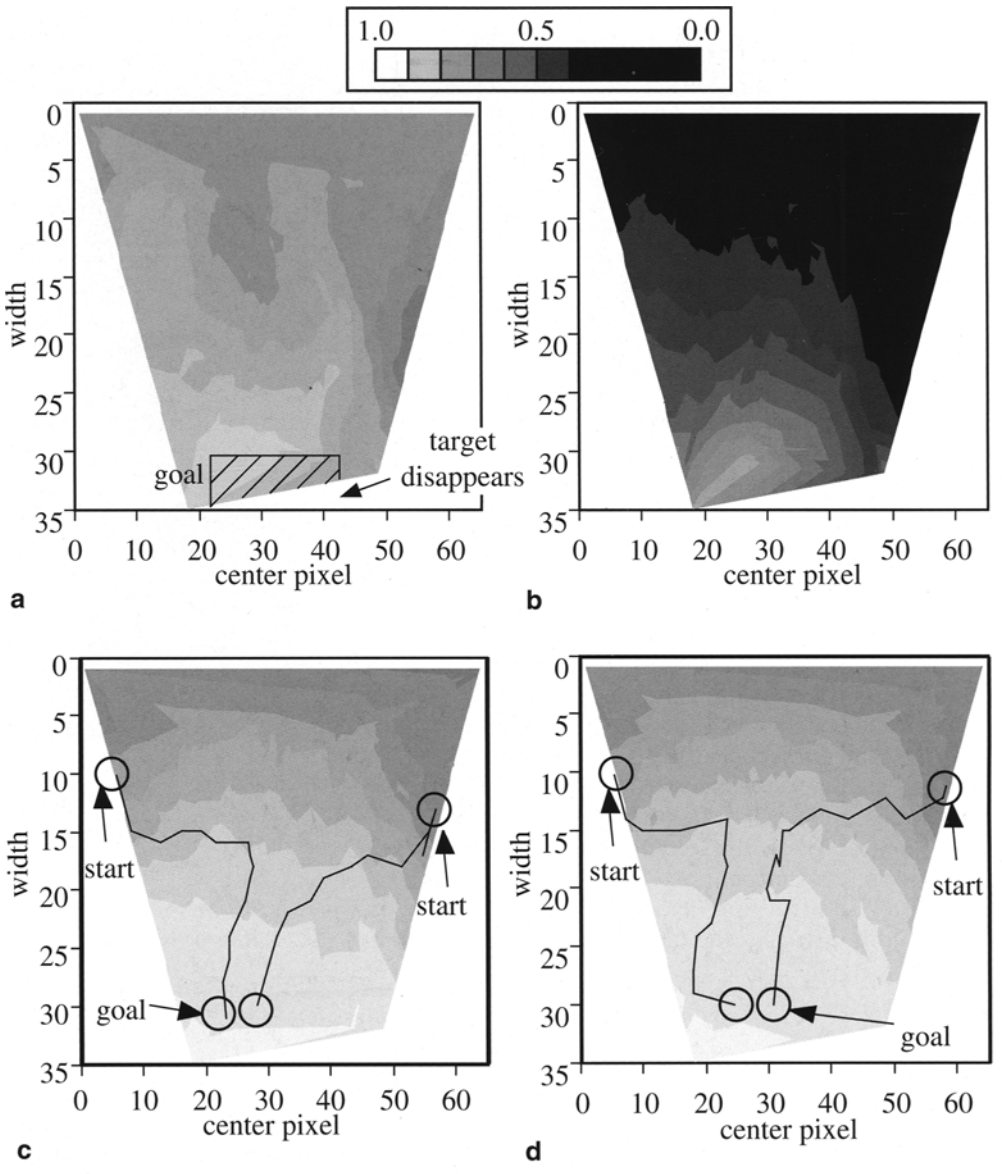


Fig. 5. Distribution of the evaluation values. **a** 400 trials; **b** 1000 trials; **c** 3200 trials; **d** 6000 trials

target, it is trained to be 0.1 as a penalty. Otherwise, the critic is trained according to Eq. 2 at every time-step with $r = 0$. Each trial is stopped after 150 time-steps, even if the robot has not reached the target object. For transforming the critic output of the neural network into the actual critic value, 0.5 is added to the critic output. The discount factor γ is 0. 99.

We now explain how the robot's initial positions are determined. The initial positions of the width and center pixel are randomly chosen from the ranges $5 \leq$ width ≤ 29 and $5 \leq$ center pixel ≤ 59. This means that the robot can always find the entire target in its initial view. The robot then goes to the initial position autonomously, according to a program which is provided beforehand. When the robot is first starting to learn, it is initially placed close to the target, since it moves only according to the random numbers. As the learning progresses, however, the initial robot location range gradually becomes wider.

Learning results

Figure 5 shows the critic distribution after learning. The loci of the target are also shown for two different initial positions in Fig. 5c and d. Figure 6 shows the loci of the robot using the absolute coordinates after 6000 trials (in Fig. 5d). The critic distribution is drawn by computing the neural network off-line for 1338 sample sets of visual sensory signals. The vertical axis in Fig. 5 indicates the width and the horizontal axis indicates the center pixel of the target object in the robot's view. It can be seen that as the learning progresses, the slope of the critic distribution is formed and beomces steep at first, and then becomes more gradual. This is because the number of time steps needed to reach the target becomes smaller, while the discount factor is always the same. It can also be seen that the critic output is low on the right side of the figures, especially in Fig. 5b. After that, the critic distribution becomes symmetrical again. The rea-

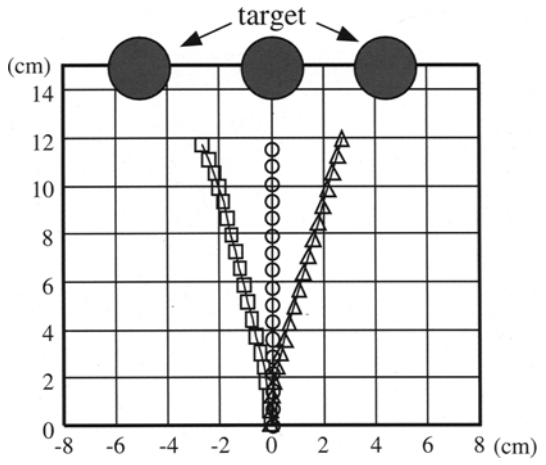


Fig. 6. The track of the robot on the field after 6000 trials

son is that the target disappears from the robot's view when it is placed at the location shown in Fig. 5a owing to the sensor characteristics mentioned in Sect. 3. When the robot loses the target object near the goal, the critic value for the previous states becomes low. After that, since the robot learns that it rotates first, and then goes forward after catching the object around the center, the critic distribution becomes symmetrical again.

Figure 7 shows a time-series of photographs that show how the robot reaches the target object exactly after 6000 trials.

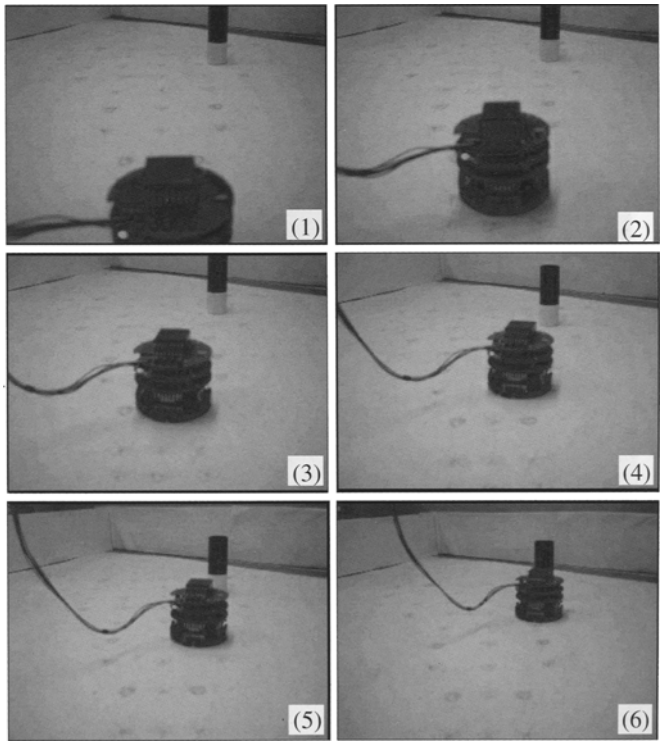


Fig. 7. The robot succeeded in reaching the target object after 6000 trials

Conclusion

Direct-vision-based RL was applied to a real mobile robot with a linear monochrome visual sensor. After considering the time-delay in transmitting the visual sensory signals, it was proposed that the actor outputs are trained using the critic output at two time-steps ahead. It was shown that a real mobile robot could reach a target object by learning from scratch without any advance knowledge or help from humans.

Acknowledgments A part of this research was supported by the Science Research Foundation of the Ministry of Education, Culture, Sports, Science and Technology of Japan (No. 13780295), and the Plan and Coordination Council of Exchange among Industry, Academy and Government in Oita.

References

1. Anderson CW (1989) Learning to control an inverted pendulum using neural networks. IEEE Control Syst Mag 9:31–37
2. Morimoto J, Doya K (2001) Aquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. Robotics Autono Syst 36:37–51
3. Tesauro GJ (1992) Practical issues in temporal difference learning. Mach Learn 8:257–277
4. Asada M, Noda S, Tawaratsumida S, et al. (1996) Purposive behavior acquisition for a real robot by vision-based reinforcement learning. Mach Learn 24:279–303
5. Shibata K, Okabe Y, Ito K (2001) Direct-vision-based reinforcement learning using a layered neural network. For the process from sensors to motors (in Japanese). Transaction of SICE 37(2):168–177
6. Shibata K, Okabe Y (1997) Reinforcement learning when the visual signals are directly given as inputs. Proceedings of ICNN Houston 3:1716–1720
7. Shibata K, Sugisaka M, Ito K (2001) Fast and stable learning in direct-vision-based reinforcement learning. Proceedings of AROB 6th'01, Tokyo, pp 200–203
8. Shibata K, Ito K, Okabe Y (1998) Direct-vision-based reinforcement learning in "Going-to-a-target task" with an obstacle and with a variety of target sizes. Proceedings of NEURAP'98, Marseille, pp 95–102
9. Barto AG, Sutton RS, Anderson CW (1983) Neuron-like adaptive elements that can solve difficult learning control probrems. IEEE Trans SMC 5:835–846