Application of Actor-Critic Method to Mobile Robot Using State Representation Based on Probability Distributions

Manabu Gouko

Dept. of Mechanical Engineering and Intelligent Systems, Faculty of Engineering, Tohoku Gakuin University
1-13-1, Chuo, Tagajo, Miyagi, 985-8537, Japan
gouko@tjcc.tohoku-gakuin.ac.jp

***Abstract*- In this study, I applied an actor-critic learning method to a mobile robot which uses a state representation based on distances between probability distributions. This state representation is proposed in a previous work and is insensitive to environmental changes, i.e., sensor signals maintaining an identical state even under certain environmental changes. The method, which constitutes a reinforcement learning algorithm, can handle continuous states and action spaces. I performed a simulation and verified that the mobile robot can learn a wall-following task. Then, I confirmed that the learned robot can achieve the same task when its sensors are artificially changed.**

Keywords- Reinforcement Learning; Actor-Critic Method; State Representation; Mobile Robot

I. INTRODUCTION

Over the past few decades, several studies have been conducted on autonomous robots. Given the wide variety of external environments, robot adaptability has become increasingly important. In designing a robot system, it is important to determine how the outside environment is expressed as a state, based on sensor information.

In a previous study, I proposed a state representation using distances between probability distributions [1]. The proposed state representation is insensitive to environmental changes, i.e., sensor signals maintain an identical state even under certain environmental changes. Sensor signals are expressed as probability distributions, while states are expressed as distances between these distributions. I conducted experiments using a mobile robot equipped with distance sensors. Experimental results showed that the proposed representation correctly recognizes similar states using a converted sensor signal.

In the previous study, I applied reinforcement learning (RL) to the autonomous mobile robot. By repeated trial and error, I confirmed that the robot can learn a suitable state–action relationship that helps it perform a task; in this case, moving forward along walls. The robot was trained by *Q*-learning method, which is not generally applicable to discrete states and action spaces. Hence, prior to learning, it is necessary to define a discrete state and action of the robot.

In the present study, I apply an actor-critic method, which uses the proposed state representation, to a mobile robot. The actor-critic method is a type of RL algorithm applicable to continuous states and action spaces. This means that the method need not define the discrete state and action prior to robot learning.

A simulation study verified that the mobile robot can learn an action relationship in the suite state using the actor-critic method. In addition, the learned robot can perform the task using converted sensor signals.

This paper is organized as follows. Section II describes the proposed state representation previously. Section III explains the application of the representation to a mobile robot as well as behavior learning by the actor-critic method. Section IV presents and discusses the experimental results. Section V concludes the study and outlines future work.

II. STATE REPRESENTATION USING PROBABILITY DISTRIBUTIONS

In this section, the state representation is introduced. Information and statistical theory adopt several measures and divergences that express the distance between two probability distributions. *f*-divergence (*f*-div) is a family of measures introduced by Csisz ár and Shields [2], which includes the well-known Kullback–Leibler divergence. The *f*-div of a probability distribution $p_i(x)$ from $p_j(x)$ is defined as

$$f_{div}\left(p_i(x), p_j(x)\right) = \int p_j(x)\, f\left(\frac{p_i(x)}{p_j(x)}\right) dx, \tag{1}$$

where $f(y)$ is a convex function defined for $y > 0$ and $f(1) = 0$. Qiao and Minematsu [3] proposed that *f*-div is invariant to invertible transforms, and they not only showed that all invariant measures must be written in the form of *f*-div but also showed that this concept is applicable to speech recognition [4].

In [1], I used the invariant of *f*-div to propose a state representation for a robot (Fig. 1). The sensor signals are assumed as probability distributions. The *f*-div of the distributions of two sensor signals p_i and p_j incoming from the same environment is denoted as $f_{div}(p_i, p_j)$. As shown in Fig. 1, under an environmental change, p_i and p_j are transformed to q_i and q_j, respectively. If q_i and q_j are described as $q_i = g(p_i)$ and $q_j = g(p_j)$, respectively, and if g is an invertible transform, then $f_{div}(p_i, p_j)$ is equal to $f_{div}(q_i, q_j)$.

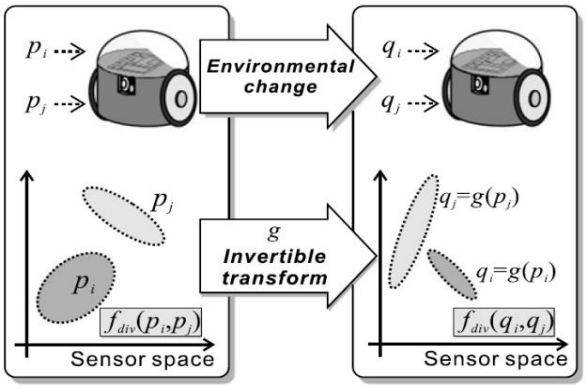

Fig. 1 State representation using distance between distributions

III. MOBILE ROBOT APPLICATION AND BEHAVIOR LEARNING

In this section, I apply the proposed state representation to a mobile robot. In addition, I describe behavior learning by the actor-critic method.

A. Mobile Robot Application

I now explain how the state representation is applied to a mobile robot equipped with multiple distance sensors. Fig. 2 shows a schematic of the autonomous mobile robot, called e-puck, used in our experiments. Six of the eight infrared distance sensors distributed on the robot were used in the experiments (indicated in Fig. 2). In the next section, I describe an experiment in which the mobile robot performs a wall-following task.

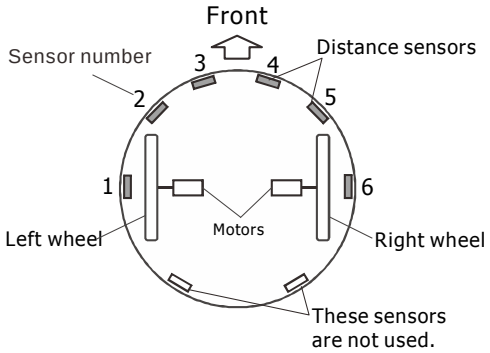

Fig. 2 Mobile robot (e-puck)

The state representation is obtained as follows. As the robot moves in time Δt, *M* sensor data are memorized for every sensor. Next, the distances between the distributions of the data collected by each sensor are calculated. Fig. 3 illustrates the data clusters for each sensor. The distribution of sensor $i\,(1, \cdots, i, \cdots I)$ is assumed to be Gaussian with mean μ_i and standard deviation σ_i. The distance between two distributions is calculated as the Bhattacharyya distance (*BD*), a fundamental equation of *f*-div. The *BD* between the distributions of the sensors i and j (p_i and p_j) is given as

$$BD(p_i, p_j) = \frac{1}{4}\frac{(\mu_i - \mu_j)}{\sigma_i^2 + \sigma_j^2} + \frac{1}{2}\ln\frac{\sigma_i^2 + \sigma_j^2}{2\sigma_i\sigma_j}. \quad (2)$$

BD is calculated from the sensor signal distributions acquired while the robot is moving between time $t - \Delta t$ and t. The distances are contained in the state vector at time t, defined as

$$\boldsymbol{s}_t = [s_{1,2}, s_{1,3}, s_{1,4}, s_{1,5}, s_{1,6}, s_{2,3}, s_{2,4}, s_{2,5}, s_{2,6}, s_{3,4}, s_{3,5}, s_{3,6}, s_{4,5}, s_{4,6}, s_{5,6}]^{\boldsymbol{T}} \quad (3)$$

where $s_{i,j}$ is $BD(p_i, p_j)$. In this method, when an object is outside the sensing range of a sensor and the distribution is 0, the distance between distributions is indeterminable. In such a situation, the distance between the distributions of that sensor and all other sensors is set to 0.

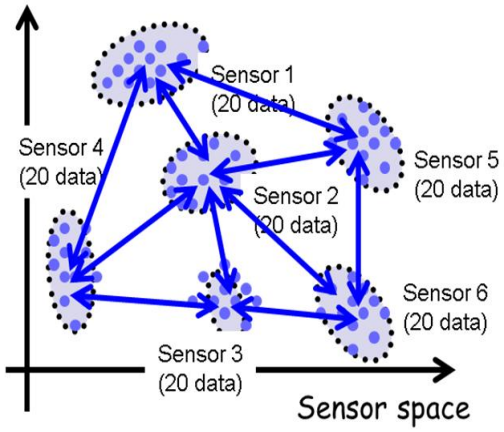

Fig. 3 Calculation of the state representation. As the robot moves in time Δt, 20(=M) sensor data are memorized for every sensor. The distances between the distributions of the data collected by each sensor are then calculated.

B. Behavior Learning by the Actor-Critic Method

In this subsection, behavior learning using RL is described [5]. In the RL framework, a robot learns a suitable state-action mapping without prior knowledge of the dynamics between itself and its environment.

I apply the actor-critic learning method, which is a RL algorithm applicable to continuous states and action spaces. This method requires a critic to estimate the reward gained from a state. It also requires an actor as a controller. An actor outputs a motor command in response to the state. In this method, actor and critic learn simultaneously. An overview of the actor-critic method is presented in Fig. 4.

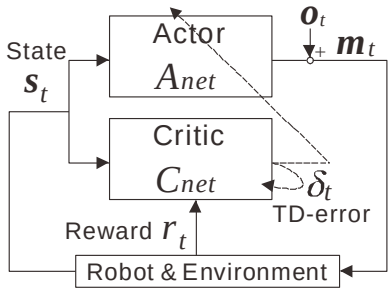

Fig. 4 Overview of the actor-critic method

The value function at time t, calculated by the critic, is defined as $(\boldsymbol{s}_t)$. If the robot moves under motor command $\boldsymbol{m}_t$, state $\boldsymbol{s}_t$ becomes state $\boldsymbol{s}_{t+\Delta t}$. The reward obtained from the environment at that time is defined as $r_{t+\Delta t}$. The value function is modified as follows:

$$V(\boldsymbol{s}_t) \leftarrow V(\boldsymbol{s}_t) + \alpha[r_{t+\Delta t} + \gamma V(\boldsymbol{s}_{t+\Delta t}) - V(\boldsymbol{s}_t)] \tag{4}$$

where α $(0 \leq \alpha \leq 1)$ is the learning rate. The reward of an environmental feature is weighted by the discount rate γ $(0 \leq \gamma < 1)$, which is a weight of reward obtained in feature. The term $[r_{t+\Delta t} + \gamma V(\boldsymbol{s}_{t+\Delta t}) - V(\boldsymbol{s}_t)]$ in Eq. 4 is called the temporal difference (TD) error, expressed as δ_t.

The critic estimates the value function from a state. In this study, the critic is implemented using an artificial neural network (C_{net}) with three layers: input, hidden, output.

A second neural network (A_{net}) represents the actor. Motor commands are calculated as

$$\boldsymbol{m}_t = A_{net}(\boldsymbol{s}_t) + \boldsymbol{o}_t \tag{5}$$

where $\boldsymbol{o}_t$ is the random noise in the motor command. The noise decreases as learning proceeds.

IV. EXPERIMENT RESULT AND DISCUSSIONS

In this section, the experimental results are presented and discussed. The robot executed behavioral learning and obtained state-action mapping. After the learning process, the sensor signal was artificially altered. The robot was able to perform the previously learned task using the obtained mapping.

The robot was placed in the experimental environment shown in Fig. 5.

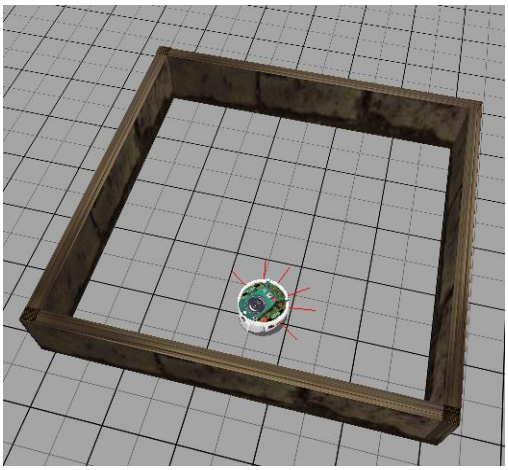

Fig. 5 Experimental environment

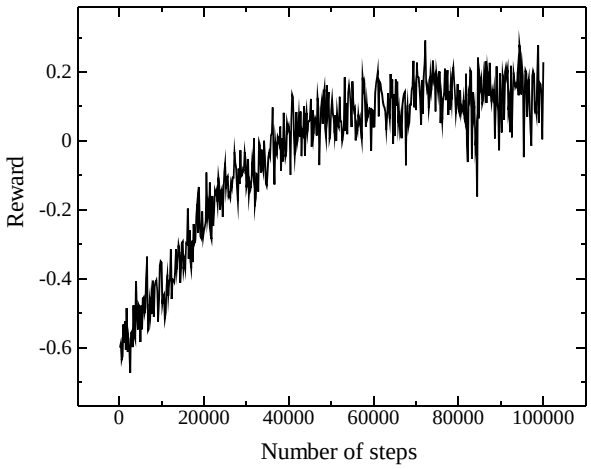

Fig. 6 Averaged reward

Behavior learning was investigated through a wall-following task. When all the conditions specified below are satisfied at time t, the reward is defined as

$$r_t = x_{t,6} - x_{t,3} + m_l + m_r \tag{6}$$

where $x_{t,6}$ and $x_{t,3}$ are signals detected by sensors 6 and 3, respectively. m_l and m_r are motor commands of the left and right wheels, respectively. Large motor commands induce forward rotation of the wheels. When the motor command is small, the wheels rotate backward. If the robot moves forward and remains close to the right-side wall, it reaps a high reward. In this experiment, Δt was set to 1 s and M was set to 20. α and γ were set to 0.7 and 0.9, respectively.

The three-layer artificial neural networks C_{net} and A_{net} were specified as follows. For C_{net}, the number of neurons of input, hidden, and output layers were 15, 20, and 1, respectively, and those for A_{net} were 15, 20, and 2, respectively.

The learning time was 100000 steps (one step =Δt). During learning, the robot was placed near the wall at 200-step intervals. The rewards in Fig. 6 are the total rewards gained over 200 steps, normalized by the total rewards obtained by the robot receiving normal sensor signals. Fig. 7 is a series of snapshots showing robot behavior after learning. The robot moves forward and remains close to the right-side wall. This result indicates successful behavioral learning by the actor-critic method.

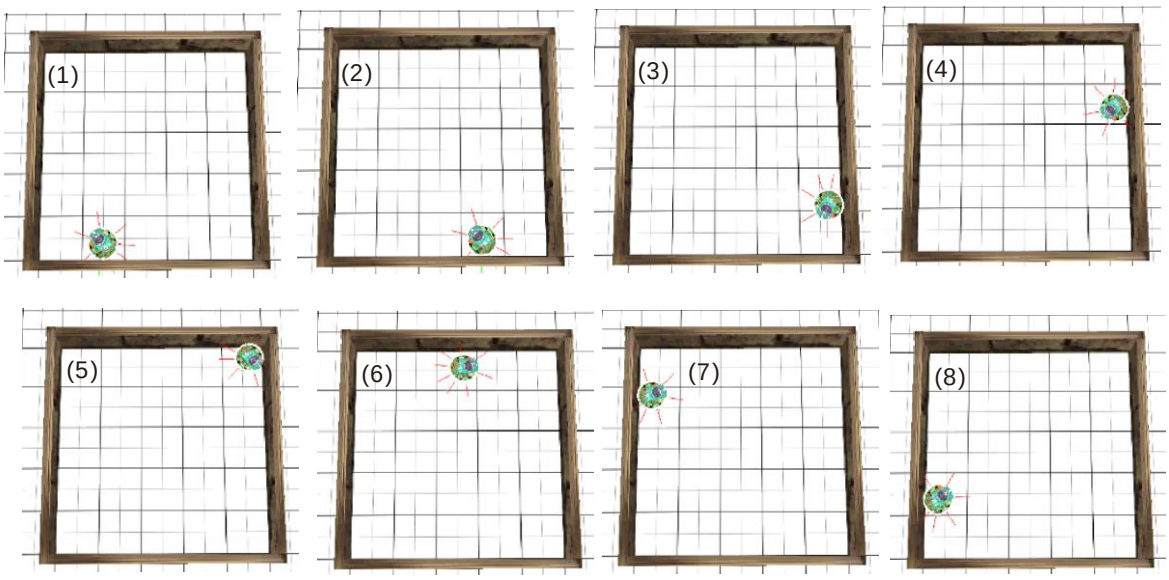

Fig. 7 Snapshots of the mobile robot following the walls of its environment

Fig. 8 shows the obtained rewards. These rewards are the total for 200 steps and are normalized by the total rewards obtained by the robot with normal sensor signals. Three different transformations (shown below) were applied:

$$x'_{t,i} = 10x_{t,i} - 5 \quad (i = 1, \cdots, 6) \tag{7}$$

$$x'_{t,i} = \sqrt{x_{t,i}} \quad (i = 1, \cdots, 6) \tag{8}$$

$$x'_{t,i} = x_{t,i}^2 \quad (i = 1, \cdots, 6) \tag{9}$$

The performance of the robot using our proposed state representation was minimally degraded under the transformation.

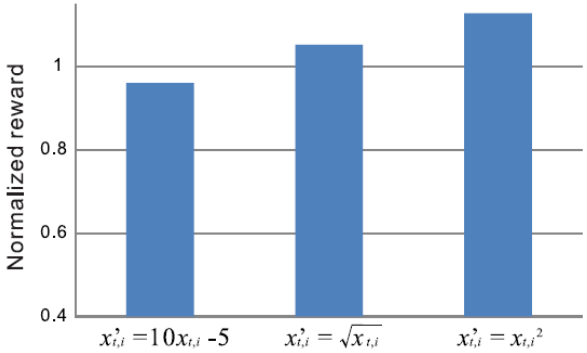

Fig. 8 Performance in terms of converted sensor signals (normalized reward).

Note that these nonlinear transformations do not represent concrete environmental changes. The results demonstrate the applicability of the proposed state representation to all invertible transformations, including nonlinear transformations.

V. CONCLUSIONS

In this study, I applied an actor-critic learning method to a mobile robot. The method uses a proposed state representation based on distances between probability distributions. This state representation is insensitive to environmental changes. Mobile robot simulations verified that the mobile robot can learn a wall-following task. The learned robot can achieve the same task when its sensors are artificially altered.

Future experiments will employ a real mobile robot. The effectiveness of the framework will be tested on various types of robots and sensors.

ACKNOWLEDGMENTS

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 24700196, 2012.

REFERENCES

[1] Gouko M. and Kobayashi Y. (2012) A State Representation Model for Robots Unaffectedby Environmental Changes. International Jounal of Social Robotics, DOI: 10.1007/s12369-012-0164-9.

[2] Csiszár I. and Shields PC. (2004) Information theory and statistics: a tutorial. Now, Boston.

[3] Qiao Y. and Minematsu N. (2008) f-divergence is a generalized invariant measure between distributions. In: Proceedings of 10th annual conference of the international speech communication association, pp. 1349-1352.

[4] Qiao Y. and Minematsu N. (2010) A study on in variance of f-divergence and its application to speech recognition. IEEE TransSignal Process, 58 (7): 3884-3890.

[5] Sutton RS. and Barto AG. (1998) Reinforcement learning: an introduction. The MIT Press, Cambridge.