The Eighteenth International Symposium on Artificial Life and Robotics 2013 (AROB 18th '13),
Daejeon Convention Center, Daejeon, Korea, January 30-February 1, 2013

Application of an actor–critic method to a robot using state representation based on distance between distributions

Manabu Gouko

Dept. of Mechanical Engineering and Intelligent Systems,Faculty of Engineering,
Tohoku Gakuin University, Japan
(Tel&Fax: 81-97-594-0181)

gouko@tjcc.tohoku-gakuin.ac.jp

Abstract: In this study, an actor–critic learning method was applied to a mobile robot. The method adopts a state representation based on distances between probability distributions. This state representation is less affected by the environment i.e., sensor signals maintain an identical state even under certain environmental changes. A simulation was performed and verified that the mobile robot can learn action relationship in the suite state using the actor–critic method.

Keywords: Reinforcement learning, actor–critic method, state representation, mobile robot

1 Introduction

Over the past few decades, several researches have been conducted on autonomous robots. Given the wide variety of external environments, robot's adaptability has become primarily importance. In a robot system, it is important to determine how the outside environment is processed as a state from sensor information.

In a previous study, a state representation was developed on the basis of noisy sensor data using distances between probability distributions[1]. This state representation is insensitive to the environment, i.e., sensor signals maintain an identical state even under certain environmental changes. Sensor signals are assumed to be expressed by probability distributions, and states are defined in terms of distances between distributions.

In the previous study, reinforcement learning to the autonomous mobile robot was applied using the proposed state representation. Then, by repeated trial and error, it was confirmed that the robot can learn a suitable state–action relationship that helps it to perform a given task. Specifically, the robot was trained by *Q*-learning method to move forward along walls. However, *Q*-learning cannot usually be applied to discrete state and action spaces. In such cases, the discrete state and the action of the robot must be defined prior to robot learning.

In this study, an actor–critic method was applied to a mobile robot, which uses the proposed state representation. Actor–critic is a reinforcement learning algorithm that can process a continuous state and action space, demanding the need to define the discrete state and action prior to robot learning.

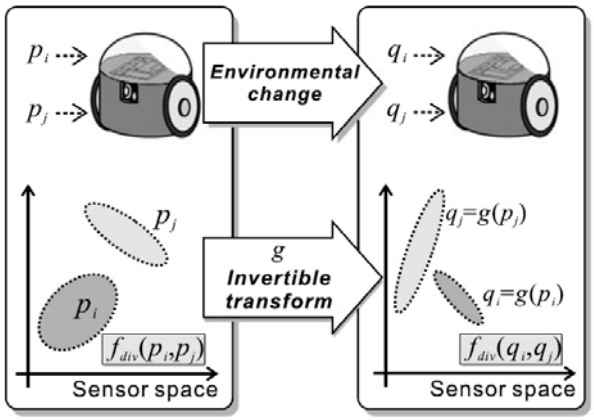

Fig. 1. State representation using distance between distributions [1].

A simulation was performed and verified that the mobile robot can learn action relationship in the suite state using the actor–critic method. The robot's ability to perform the moving task was confirmed.

2 State representation using distance between distributions

In this section, the proposed state representation was defined. In theoretical information and statistics, the distance between two probability distributions is expressed in several ways. *f*-divergence (*f*-div) is a family of measures introduced by Csiszár and Shields [2] that includes the well-known Kullback–Leibler divergence. The *f*-div of probability distribution $p_i(x)$ from $p_j(x)$ is defined as

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

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 showed that all invariant measures are of the *f*-div form. Furthermore, they applied the invariant of measures to speech recognition [4].

3 Mobile Robot Application and Behavior Learning

Subsection 3.1shows how the proposed state representation is applied to a mobile robot. Subsection 3.2 describes behavior learning by the actor–critic method.

3.1 Mobile Robot Application

Figure 2 shows the autonomous mobile robot, named e-puck, used in our experiments.

The robot has eight infrared distance sensors. In our experiments, I used the six sensors labeled in Fig. 2. The next section describes an experiment in which the mobile robot performs a wall-following task. It was assumed that the robot's sensors respond only to differences in the wall color and not to wall position. Figure 3 shows how the state representation is acquired. First, while a robot moves in time Δt, each sensor memorizes M data. Next, the distances between the distributions of each sensor's data are calculated. In this study, the sensor distribution i $(1, \cdots, i, \cdots I)$is assumed to be Gaussian with mean μ_i and standard deviation σ_i. I used the Bhattacharyya distance (*BD*), a widely used measure of *f*-div, as the distance between two distributions. The *BD* between the distributions of the sensors i and j ($p_i$andp_j) is given by the following formula

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

BD is calculated from the sensor signal distributions acquired by the moving robot from time $t - \Delta t$ to t. The state vector at time t contains distances and is defined as follows:

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

where $\boldsymbol{v}_{i,j}$ is $BD(p_i, p_j)$. In this formulation, when an object is outside the sensing range of a sensor and the distribution is 0, the distance between distributions cannot be calculated. In such a situation, the distance between the distributions of that sensor and other sensors is set to 0.

3.2 Behavior learning by actor–critic method

This subsection explains behavior learning using reinforcement learning [5]. In the reinforcement learning framework, a robot learns a suitable state–action mapping without prior knowledge of its dynamics or environment.

The actor–critic learning method is applied, which is a reinforcement learning algorithm that can handle a continuous state and action spaces. This method needs a critic, which estimates a reward expectation from a state. It also needs an actor as a controller. The actor outputs a motor commanding response to the state.

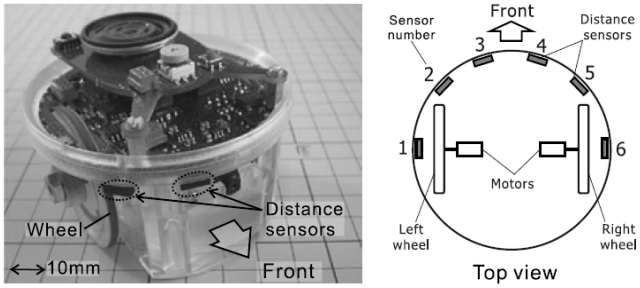

Fig. 2. Autonomous mobile robot e-puck and its infrared distance sensors [1].

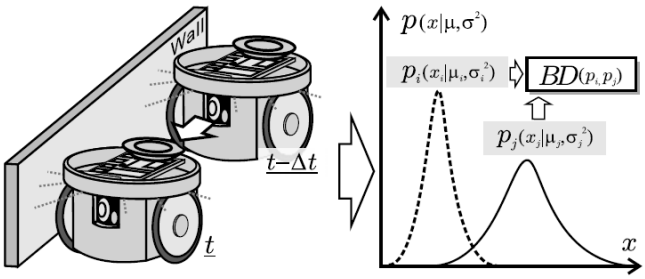

Fig. 3. Proposed state representation using e-puck. While a robot moves in time Δt, Mdata are memorized by every sensor. Next, the distances between the distributions of each sensor's data are calculated [1].

4 Experimental Result and Discussions

In this section, I present and discuss the experimental r esult. The robot executed behavior learning and obtaine d state–action mapping. After the learning process, the sensor signal was artificially transformed and the robot can perform a task using the acquired mapping was ver ified.

The robot was then placed in the experimental environment shown in Fig. 4.

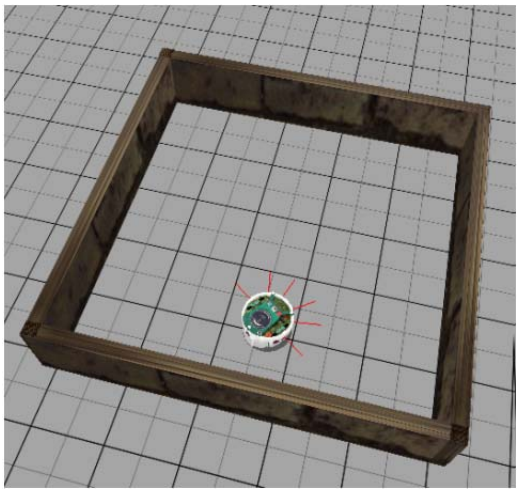

Fig. 4. Experimental environment.

Behavior learning was assessed in a wall-following task. When all conditions given below were satisfied at time *t*, the reward was defined as follows,

$$r_t = x_{t,6} - x_{t,3} + m_l + m_r \,. \qquad (4)$$

Here $s_{t,6}$ and $s_{t,3}$ are signal outputs by sensors 6 and 3, respectively, and m_l and m_r are the respective motor commands of the left and right wheel. In this experiment, Δt was set to 1 sec and M was set to 20.

The learning time was 10,000 steps (one step = Δt). The robot was placed near the wall at every 500 steps during learning. After learning, I confirmed the success of the learning behavior. Figure 5 shows the obtained reward data for the three nonlinear transformations shown below (Equations 5, 6, and 7).

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

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

$$x'_{t,i} = x_{t,i}^2 \quad (\mathrm{i} = 1, \cdots, 6) \qquad (7)$$

In this figure, the total rewards accumulated over 200 steps are normalized by the total rewards obtained by the robot with normal sensor signals.

The performance of the robot using our proposed state representation showed minimal degradation. Note that these nonlinear transformations did not correct for physical environmental changes. The results indicate that the proposed state representation model is applicable to all invertible transformations, including nonlinear transformations.

5 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 the environment, i.e., sensor signals maintain an identical state even under certain environmental changes. A simulation was performed and verified that the mobile robot can learn action relationship in the suite state using the actor–critic method.

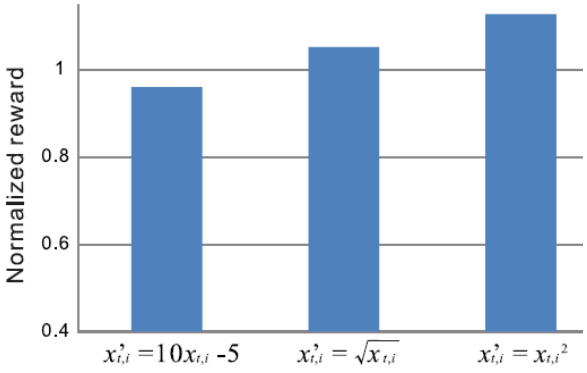

Fig. 5. Normalized rewards.

Acknowledgements

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 Journal of Social Robotics, DOI:10.1007/s12369-012-0164-9.
[2] Csiszár I, 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.z
[4] Qiao Y. and Minematsu N. (2010) A study on invariance of *f*-divergence and its application to speech recognition. IEEE TransSignal Process, 58 (7): 3884–3890.
[5] Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. The MIT Press, Cambridge.