Reinforcement Learning for Biped Robot

Yutaka Nakamura[1], **Masa-aki Sato**[2,3] **and Shin Ishii**[1,3]

[1] Nara Institute of science and technology. 8916-5 Takayama-cho, Ikoma, Nara 630-0192. yutak-na@is.aist-nara.ac.jp

[2]ATR, Human Information Science Laboratories. 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288.

[3]CREST, JST.

1. Introduction

Neurobiological studies revealed that rhythmic motor patterns are controlled by neural oscillators referred to as central pattern generators (CPGs) [2]. Inspired by these findings, human-like biped walking was successfully simulated in [3] by using the CPG controller. However, it is very difficult to determine the CPG parameter values for various robots and environments, since there is no design principle to determine the parameter values.

The main aim of this article is to study a reinforcement learning (RL) method for a CPG controller that generates stable rhythmic movements. In order to deal with the CPG controller, we propose a new RL method called the CPG-actor-critic method.

2. CPG-actor-critic model

An actor-critic model is a popular RL method. In that method, the actor is a controller that generates control signals to the physical system. The critic predicts the reward accumulation toward the future.

When we try to apply the actor-critic model to the CPG controller, there occur several difficulties. The RL task becomes a partially observable problem. In addition, the usual gradient-based learning algorithm for the actor-critic model is not suited for training the CPG controller.

In order to overcome these difficulties, the CPG controller is divided into two modules, i.e., the basic CPG and the actor. The basic CPG is a dynamical part of the CPG with fixed weights. The actor is a linear controller without any mutual feedback connection, which receives the basic CPG output and the sensory feedback signal, and outputs indirect control signal to the basic CPG. The parameter of the actor can be determined by a gradient method.

3. Experiment

We apply the CPG-actor-critic method to the biped robot simulator [3]. We assume that an immediate reward is determined by the next robot state $\mathbf{x}$: $\tilde{r}(\mathbf{x}) = 0.5 r_{height}(\mathbf{x}) + 0.02 r_{speed}(\mathbf{x})$,where $r_{height}(\mathbf{x})$ is proportional to the height of robot's hip and encourages the robot not to fall down. $r_{speed}(\mathbf{x})$ is proportional to the speed of robot's hip and encourages the robot to proceed to the forward direction.

After about 5800 learning episodes, the robot started to walk stably. The biped robot controlled by the learned parameter $\mathbf{a}_{RL}$ is able to walk on inclined or rough ground more stably than by the hand-tuned parameter $\mathbf{a}_{HT}$.

4. Concluding remarks

In this article, we proposed a new RL method called the CPG-actor-critic method and applied it to automatic acquisition of the biped locomotion. By using our method, a CPG controller that can walk in various environments more stably than the hand-tuned one was obtained.

References

[1] Grillner, S., Wallen, P. and Brodin, L. 1991. Neuronal network generating locomotor behavior in lamprey. *Annu. Rev. Neurosci.* 14:169-199

[2] Taga, G., Yamaguchi, Y., and Shimizu, H. 1991. Self-organized control of bipedal locomotion by neural oscillators in unpredictable environment. *Biol. Cybern.* 65:147-159

[3] Sutton, R. S. and Barto, A. G. 1998. *Reinforcement learning.* MIT Press

[4] Sato, M. and Ishii, S. 2000. On-line EM algorithm for the normalized Gaussian network. *Neural Computation* 12:407-432

[5] Sato, M. and Ishii, S. 1999. Reinforcement learning based on on-line EM algorithm. *NIPS 11*, 1052-1058

[6] Morimoto, J. and Doya, K. 2001. Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robot. Auton. Syst.,* 36:37-51.

[7] Ogihara, N. and Yamazaki, N. 2001. Generation of human bipedal locomotion by a bio-mimetic neuro-musculo-skeletal model. *Biol. Cybern.* 84:1-11.