Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 20 (2007) 723–735
Neural Networks
www.elsevier.com/locate/neunet

Reinforcement learning for a biped robot based on a CPG-actor-critic method

Yutaka Nakamura[a,b], Takeshi Mori[a], Masa-aki Sato[c], Shin Ishii[a,*]

[a] *Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan*
[b] *Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan*
[c] *ATR Computational Neuroscience Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan*

Received 4 April 2006; received in revised form 16 January 2007; accepted 16 January 2007

Abstract

Animals' rhythmic movements, such as locomotion, are considered to be controlled by neural circuits called central pattern generators (CPGs), which generate oscillatory signals. Motivated by this biological mechanism, studies have been conducted on the rhythmic movements controlled by CPG. As an autonomous learning framework for a CPG controller, we propose in this article a reinforcement learning method we call the "CPG-actor-critic" method. This method introduces a new architecture to the actor, and its training is roughly based on a stochastic policy gradient algorithm presented recently. We apply this method to an automatic acquisition problem of control for a biped robot. Computer simulations show that training of the CPG can be successfully performed by our method, thus allowing the biped robot to not only walk stably but also adapt to environmental changes.

Keywords: Reinforcement learning; Central pattern generator; Biped walking; Actor-critic model; Policy gradient method

1. Introduction

Rhythmic movements are fundamental to animal behaviour; for example, locomotion and swimming are crucial abilities for many animals to survive. These rhythmic movements are characterized by their rapid adaptability to changing environments, and the mechanism realizing such adaptability has been extensively studied both in biological science and in engineering (Doya & Yoshizawa, 1992; Grillner, Wallen, Brodin, & Lansner, 1991). Recent studies on biped locomotion have enabled humanoid robots to walk in the real world (Hirai, Hirose, Haikawa, & Takenaka, 1998). Despite these advances, however, further studies are still necessary because human locomotion movements are much more flexible and robust than those of the current robots (Morimoto & Atkeson, 2003).

Neurobiological studies have revealed that rhythmic motor patterns, such as swimming by a fish or crawling by a reptile, are controlled by neural circuits called central pattern generators (CPGs), which exist in the spinal cord of vertebrates and output rhythmic signals (Grillner et al., 1991). It has also been suggested that sensory feedback signals play an important role in stabilizing rhythmic movements by coordinating the physical system with the CPGs. Ekeberg (1993) investigated the spinal cord of a lamprey, constructed a model of CPG, and simulated swimming locomotion of a model lamprey composed of ten rigid links. In the field of engineering, many studies have focused on legged robots controlled by CPG controllers, for example, hexapod (Cruse, Kindermann, Schumm, Dean, & Schmitz, 1998; Lewis, Fagg, & Bekey, 1993), quadruped (Fukuoka, Kimura, & Cohen, 2003) and single legged (Wadden & Ekeberg, 1998). Among those, Taga, Yamaguchi, and Shimizu (1991) derived a model of a human lower body (a biped robot) and a CPG controller and then applied these to simulate human-like biped walking. Although there have been some studies of designing a CPG controller (Matsuoka, 1985), the determination of CPG parameters is still difficult, since they are dependent on both the target physical system (robot) and the environment.

* Corresponding author.
E-mail addresses: nakamura@ams.eng.osaka-u.ac.jp (Y. Nakamura), tak-mori@is.naist.jp (T. Mori), masa-aki@atr.jp (M. Sato), ishii@is.naist.jp (S. Ishii).

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.01.002

To determine these parameters, many studies have employed a genetic algorithm (GA). For example, a salamander that 'walks' efficiently (Ijspeert & Cabeluguen, 2003) and a human-like biped walking that conserves much energy but achieves long-distance walking (Ogihara & Yamazaki, 2001) have been successfully simulated. A GA is a model of an evolution process of animal species, and the parameter is basically updated by selecting individuals with higher fitness, where the fitness is determined by interactions with the environment during all lifetime of each individual. Meanwhile, a reinforcement learning (RL) tries to model the learning mechanism of individual animals and the parameter is basically updated for each interaction with the environment. Then, the time scale of trial-and-error is conceptually different between GA and RL. In this article, we propose an RL method for a CPG controller that generates stable rhythmic movements.

In the field of RL, value-based RL methods like TD(λ) learning have been successfully applied to various Markov decision processes (MDPs) with finite state and action spaces (Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 1998). In these methods, the value of each state (and action) is estimated so that the policy is updated to increase the value. On the other hand, policy-based methods have recently been attracting much attention because of their robustness. In the latter methods, the gradient of the performance indicator with respect to the policy parameter is estimated probabilistically, and the policy parameter is updated according to the gradient (Williams, 1992). These "stochastic policy gradient" methods are proven to converge under certain conditions (Konda & Tsitsiklis, 2003; Sutton, McAllester, Singh & Manour, 2000), and combinations of the stochastic policy gradient and value learning have been successfully applied to playing the game of Tetris (Kakade, 2001), balancing a cart-pole (Kimura & Kobayashi, 1998; Peters, Vijayakumar, & Schaal, 2005), motor control of a humanoid robot (Schaal, Peters, & Ijspeert, 2004) and quasi passive dynamic walking (Tedrake, Zhang, & Seung, 2004).

Nevertheless, standard policy gradient methods (Kimura & Kobayashi, 1998; Konda & Tsitsiklis, 2003; Sutton et al., 2000) are not suited to training the CPG controller, which is an instance of recurrent neural networks, because the control signal generated by the CPG controller should depend on the past states of the controller itself; namely, the control should incorporate its context, which makes the policy of the CPG controller non-stationary. To deal with this contextual-dependency problem, we propose in this study a new RL scheme called the "CPG-actor-critic" method. We apply this method to the biped robot simulator used in Taga et al. (1991). Computer simulations show that training of the CPG can be successfully performed by our RL method, thus allowing the biped robot to stably walk in the sagittal plane and, moreover, to adapt to environmental changes.

2. Central pattern generator

To develop a framework for autonomously obtaining a control for a robot with a large number of degrees of freedom (DOF), such as a humanoid robot, it is necessary to employ some specific devices, in order to overcome the "curse of dimensionality" inevitably involved in problems with a large DOF. For example, imitating human movements is beneficial for restricting the possible high-dimensional control space (Bentivegna, Ude, Atkeson, & Cheng, 2002; Inamura, Nakamura & Toshima, 2004).

As for other devices, many studies have employed a CPG to generate rhythmic movements. In those studies, the CPG controller and the robot interacted with each other and were eventually entrained into a limit-cycle attractor. The characteristic of this entrainment is attributed to connection weights in the CPG controller and those from sensory feedback signals. The basic rhythmic character of the CPG controller is determined by mutual connections among the CPG neurons, and the sensory feedback connections coordinate the phase relation of the rhythm between the CPG controller and the robot. When these weight values are set to produce a stable attractor, the CPG-based control methods are robust against disturbances from the environment. Among these studies, a human-like biped walking was successfully simulated by using a CPG controller (Taga et al., 1991).

Other approaches to realizing human-like walking, for example, control methods based on "zero moment point (ZMP)", have allowed humanoid robots to walk in the real world (Hirai et al., 1998; Lim, Yamamoto, & Takanishi, 2002). In these methods, a target trajectory is calculated based on the precise information of the environment, suggesting that a new target trajectory should be re-calculated when a disturbance occurs. In addition, energy consumption is usually high, and the gait pattern can be quite different from that of a human.

Passive dynamic walking (Asano, Yamakita, Kamamichi, & Luo, 2004; McGeer, 1990) is an efficient and human-form of walking, and it is realized by designing the robot to be entrained mechanically into an attractor. This walking is, however, known to be unstable in the event of disturbances. Quasi-passive dynamic walking was also proposed to make walking robust and to allow the robot to walk on various terrains, and Tedrake et al. (2004) applied a stochastic policy gradient method to automatic control of the quasi-passive dynamic biped walking. In this study, the actor (a controller) observes a state of the biped robot and outputs a parameter of a nonlinear feedback controller once in each step; the controller with the selected parameter then produces control signals until the next step. Because the motion pattern of the biped robot is determined by the characteristics of the dynamical system and the parameter output by the actor, this learning scheme can be viewed as learning the selection of motion patterns. Their study focused on the stabilization of walking by utilizing the property of the target system because a passive dynamic walking robot is inherently suitable for walking but unstable with disturbances, while our study focuses mainly on the acquisition of stable walking by utilizing the rhythmic patterns inherently possessed by the controller.

We assume that a physical system is controlled by a CPG controller, as depicted in Fig. 1. The CPG controller is implemented as a neural oscillator network that outputs control