The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Acquisition of Energy-Efficient Bipedal Walking Using CPG-Based Reinforcement Learning

Takita Tomoyuki*, Yoshiyuki Azuma*, and Tomoshiro Shibata*
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma, Nara, 630-0192 Japan
{tomoyuki-t, yoshiyuki-a, tom}@is.naist.jp

Abstract— **Although there have been much research on robot walking, the energy efficiency of central pattern generator (CPG)-based walking has not received much attention. This study proposes a novel method for acquiring energy-efficient CPG-based bipedal walking for a robot with knees and feet. In this method, we introduce a torque-free period for swing leg control into the swing leg control cycle. During this period, no torque is applied to the hip joint controller, and therefore no energy is consumed. When and for how long the torque-free period is inserted into the swing leg control cycle is adaptively acquired by reinforcement learning. Simulation experiments demonstrate the feasibility of our method. The energy consumed in acquiring walking is reduced by 40% compared with simple CPG-based walking without the torque-free period in the practical learning speed. Walking stability is maintained with respect to external disturbances on a level floor. Although the method is more unstable on slopes with the torque-free period, the torque-free-period can be adaptively eliminated to achieve stable walking on the slopes.**

I. INTRODUCTION

Bipedal walking is one of the major research topics in current humanoid robotics, and researchers have developed various controllers such as the quasi-passive dynamic (QPD) controllers [1], [2], [3], [4], [5], controllers based on the zero-moment point (ZMP) [6], and controllers based on central pattern generators (CPG) [7], [8], [9].

ZMP-based control usually consumes a large amount of energy to maintain the desired trajectory [10]. In contrast, QPD control is based on passive dynamic walking (PDW) control [11], which enables completely unactuated walking on a gentle decline. Because it is widely known that PDW is generally sensitive to the robot's initial posture, speed, and disturbances incurred when a foot touches the ground, QPD controllers have been proposed, in which some actuators are activated in a supplementary role to handle disturbances. Although the parameters in a QPD controller are less sensitive than those in PDW control, adjustments are still required, and this requirement becomes stricter when operating in non-stationary and/or unknown environments. Studies exist that have applied reinforcement learning to the autonomous adjustment of parameters [4], [12]. ZMP-based control and PDW control seem to be on opposite ends of the spectrum with respect to the trade-off between energy efficiency and walking stability.

Neurobiological studies have revealed that rhythmic motor patterns are produced by neural oscillators called CPGs [13]. 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. Based on these observations, Taga et al. devised a model of the lower half of a human body (a biped robot) and a CPG controller, and applied these in simulations of human-like biped walking [7]. To achieve this biped walking, however, it was necessary to determine CPG parameters which is dependent on the target physical system (robot) and the environment. Various studies have applied reinforcement learning to the autonomous adjustment of parameters [9], [8]. The energy efficiency of CPG-based walking, however, has not received much attention.

TABLE I
PHYSICAL PARAMETERS OF ROBOT

	Body	Thigh	Shank	Foot
Length [m]	0.312	0.280	0.317	0.180
Weight [kg]	6.646	0.673	0.707	0.398

This study proposes a novel method to enable a robot with knees and feet to acquire energy-efficient CPG-based bipedal walking. For energy efficiency, we introduce a torque-free period in the swing leg control during which no torque is applied to the hip joint controller. When and how long the torque-free period is inserted into the swing leg control cycle is acquired by reinforcement learning adaptable to changes in the environment.

This paper is organized as follows. In Section II, we describe the robot model, control system, and learning algorithm. Simulation settings and results are given in Section III. Finally, in Section IV we present our conclusions and future works.

II. LEARNING CPG-BASED ENERGY-EFFICIENT CONTROL

A. Robot Model

Fig. 1 and Table I specify the robot model and physical parameters used in this study. The robot model has seven links and six joints. The model consists of links called Body, Thigh, Shank and Foot, and joints called Hip, Knee and Ankle. The feet are flat. The range of the Hip joint is from -51 [deg] to 51 [deg] and that of the Ankle joint from -26 [deg] to 26 [deg]. We used the open dynamics engine (ODE) [14] for simulation.

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

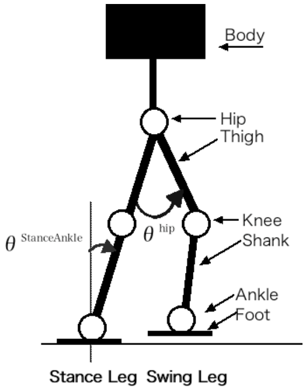

Fig. 1. Robot Model

B. Criteria for energy efficiency

To compare energy efficiency between humans and robots of different sizes, Collins et al. proposed the specific energetic cost of transport C_{et}, and the specific mechanical cost of transport C_{mt} [15]. Whereas C_{et} uses the total energy consumed by the system, C_{mt} considers only the positive mechanical work of the actuators. Throughout this paper, we refer to C_{mt}, which is defined as

$$C_{mt} = \frac{\text{Energy consumed by motors[J]}}{\text{Robot's weight[kg]} \times \text{ Walking distance[m]}}. \quad (1)$$

C. CPG Controller

In general, bipedal walking comprises an initiation phase and a maintenance phase. This study focuses on control of the maintenance phase. In the initiation phase, the robot starts walking with the predetermined initial joint angle of the Hip and with the initial horizontal velocity of the Body. The walking distance is measured as the horizontal displacement of the robot's center of mass during an episode.

In the maintenance phase, CPG control is applied at the Hip. The Knees and Ankles are controlled in accordance with the state of the CPG. We employed a CPG controller adapted from Cohen's models [16] because the parameterization thereof is designed for cyclic motion such as walking. Our CPG controller is designed as

$$\dot{\gamma} = u^r + \kappa\dot{\theta}^{StanceAnkle}, \quad (2)$$

$$\zeta = u^A \sin(\gamma), \quad (3)$$

where ζ is the output of the CPG controller and is used as the desired Hip angular velocity. This desired joint angular velocity was realized by the internal PID controller of the ODE. κ is the Ankle feedback gain, u^r specifies a constant angular velocity, and u^A is the amplitude of the CPG controller. For both Knees and Ankles, we designed two different controllers for the swing leg and stance leg, respectively. A bang-bang controller is used for the knee of the swing leg to prevent it from hitting the ground and to extend it. The Ankles are controlled to keep the feet parallel to the ground. The Knee of the stance leg is kept extended by constant torque. The stance and the swing leg are changed by ζ, cf. Fig. 2. T_n indicates the time when both feet touch on the ground at the nth walking step.

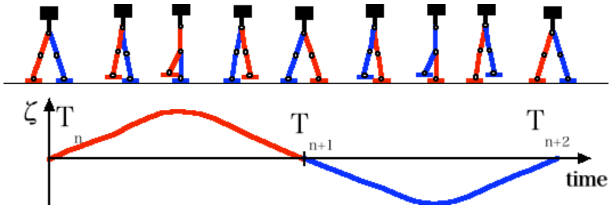

Fig. 2. Overview of CPG-output trajectory

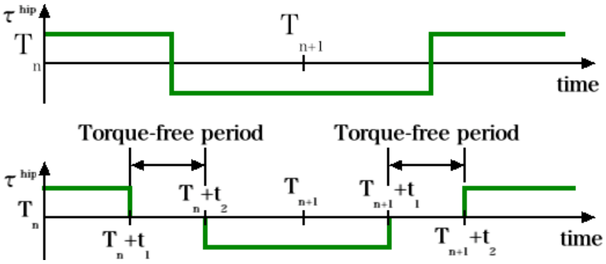

Fig. 3. Hip torque pattern without torque-free period (upper) and with torque-free period (lower)

For energy efficiency, we introduced torque-free periods as illustrated in Fig. 3. The controller does not generate any torque input to the Hip joint during the torque-free period which starts at time t_1 and ends at t_2. Because t_1 and t_2 are meant to be dependent on the walking environment, reinforcement learning was applied to acquire a t_1 and t_2 suited to the environment in an online fashion.

D. Reinforcement Learning

In this study, the robot searches for low-energy consumption walking to adjust the torque-free start and end times. We applied reinforcement learning because it enables the robot to learn adapt to variations in the environment including its body parameters without having a clear model of the robot and environment.

We applied the GARB algorithm [17]. With this algorithm, the robot selects the torque-free start time t_1^i and end time t_2^i for the ith episode by policy π, which is parameterized as vector $\overline{\mathbf{T_i}}$. While walking, the robot is in state $\mathcal{S}$, which contains the joint angle and angular velocity and the walking distance. After walking a certain distance, the robot is given a reward, r_i. The aim of the robot is to maximize the average reward by updating the policy, parameterized by $\overline{\mathbf{T_i}}$ as $\pi\left(\mathbf{T_i}|\overline{\mathbf{T_i}}\right)$. After the ith episode, the parameter vector for the next step $\overline{\mathbf{T_{i+1}}}$ is updated using reward r_{i+1}, baseline B_{i+1}, and eligibility trace $\mathbf{Z_{i+1}}$.

$$\overline{\mathbf{T_{i+1}}} = \overline{\mathbf{T_i}} + \alpha_p \nabla\overline{\mathbf{T_i}} \quad (4)$$

$$\mathbf{e}_{i+1} = \frac{\partial}{\partial\overline{\mathbf{T_i}}} \ln\left(\pi\left(\mathbf{t}_i|\overline{\mathbf{T_i}}\right)\right) \quad (5)$$

$$\mathbf{Z}_{i+1} = \mathbf{e_i} + \beta\mathbf{Z}_i \quad (6)$$

$$B_{i+1} = B_i + (r_i - B_i)/i \quad (7)$$

TABLE II

LEARNING RATE α_p

Episode	1∼100	101∼200	201∼300	301∼400	401∼
α_p	0.1	0.05	0.01	0.005	0.001

TABLE III

CPG PARAMETERS

α	u^a	u^η	u^r	κ
1.0	2.0	0.0	2.1	1.8

$$\nabla\overline{\mathbf{T_i}} = \alpha_p \left(r_{i+1} - B_{i+1}\right) \mathbf{Z}_{i+1}, \tag{8}$$

where α_p is the learning rate.

E. Learning of the torque-free period

The torque-free period start time t_1^i and end time t_2^i are generated by a stochastic policy π, which is a 2-dimensional Gaussian function with mean $\overline{\mathbf{T_i}}$ and covariance matrix $\mathbf{\Sigma}$. More specifically, the policy function is defined as

$$\pi\left(\mathbf{T_i}|\overline{\mathbf{T_i}}\right) = \frac{1}{(2\pi)\,\mathbf{\Sigma}^{(1/2)}} \exp\left\{-\frac{1}{2}\left(\mathbf{T_i} - \overline{\mathbf{T_i}}\right)^{\mathrm{T}} \mathbf{\Sigma}^{-1}\left(\mathbf{T_i} - \overline{\mathbf{T_i}}\right)\right\}, \tag{9}$$

$$\overline{\mathbf{T_i}} = \left[\begin{array}{c} \overline{t}_1^i \\ \overline{t}_2^i \end{array}\right], \tag{10}$$

where $\mathbf{\Sigma} = diag(\sigma_1^2, \sigma_2^2)$, and$\sigma_1^2$ and σ_2^2 were both set to 0.04. The eligibility $\mathbf{e}$ is defined as

$$\mathbf{e_i}' = \left[\begin{array}{c} \frac{t_1^i - \overline{t_1^i}}{\sigma_1^2} \\ \frac{t_2^i - \overline{t_2^i}}{\sigma_2^2} \end{array}\right]. \tag{11}$$

Because the update value of the policy is proportional to σ_1^2 and σ_2^2, the eligibility vector is

$$\mathbf{e_i} = \left[\begin{array}{c} t_1^i - \overline{t_1^i} \\ t_2^i - \overline{t_2^i} \end{array}\right]. \tag{12}$$

The reward function was defined as the inverse of C_{mt}.

$$\begin{aligned} r_i &= 1/C_{mt} \\ &= \frac{\text{Robot's weight[kg]} \times \text{Walking distance[m]}}{\text{Energy consumed by motors[J]}} \end{aligned} \tag{13}$$

This reward function is suitable for our purpose because the reward increases in value as the robot walks further and consumes less energy. Thus, after learning, the robot should be able to walk a longer distance while consuming less energy.

III. EXPERIMENTS

In this section, we investigate the feasibility of our method through simulation experiments.

TABLE IV

SCOPE OF THE GRID SEARCH

α	u^a	u^η	u^r	κ
-2.0∼2.0	0.0∼3.0	-1.0∼1.0	0.0 ∼ 3.0	0.0 ∼ 2.0

A. Settings

We first applied our method to a robot on level ground. One learning episode was defined in which the robot completed 10 [m] of walking or falling down. Policy π was updated as each episode terminated. The learning rate α_p was changed in a simulated-annealing fashion to stabilize learning, cf. Table II. Discount factor β for the eligibility trace was set to 0.80. Initial parameters $\overline{t_1}$ and $\overline{t_2}$ were both set to 0.5 [s], which represents half of the walking cycle. Walking started with $\theta^{Hip} = 7.6$[deg], $v = 0.175$[m/s]. The other parameters for the CPG controller are given in Table III. The CPG parameters and initial conditions (position and angular velocity of center of gravity) were determined so as to enable the robot to walk the longest distance on the flat surface. The ranges of the parameters used in the grid search are shown in Table IV. Having observed the duration of all the walking, we finally applied the CPG parameters and initial conditions under which the robot had walked the furthest.

Next, we investigated the robustness of the acquired walking against disturbances. Disturbances (20 [N] forward and backward horizontally) were applied to the robot's body for 0.1 [s] after 0.4 [s] of each trial. All other parameters remained the same as in the previous learning experiment.

Finally, we investigated the online adaptability of the controller, with the torque-free period, to an environmental change. The environment initially consisted of level ground for 2 [m], but then changed to an ascending slope of 0.02 [rad] (1.2 [deg]). We made sure that the robot was unable to climb up this slope with the torque-free period determined by the first experiment. Because we also made sure that the robot was able to climb up the slope without the torque-free period, the torque-free period was expected to be eliminated by learning. The learning algorithm and parameter α_p were not changed. The initial values were set to $\mathbf{W} = \left[\begin{array}{cc} 0.3602 & 0.6714 \end{array}\right]^T$ $\mathbf{e} = \left[\begin{array}{cc} -0.01286 & -0.0001908 \end{array}\right]^T$, and $B = \left[\begin{array}{cc} 0.05392 & -0.005343 \end{array}\right]^T$, as obtained in the first experiment.

B. Return map analysis

To quantify the stability of our nonlinear, stochastic, periodic trajectory, we measure the eigenvalues of the return map [18]. For each step we estimated the difference between $\mathbf{x_n}$, the state value of the nth step, and the equilibrium of the return map $\mathbf{x}^*$

$$(\mathbf{x}_{n+1} - \mathbf{x}^*) = \mathbf{A}(\mathbf{x}_n - \mathbf{x}^*), \tag{14}$$

where $\mathbf{x}$ represents the state value of the robot on a Poincaré section, the Hip's angle and the angular velocity of the nth

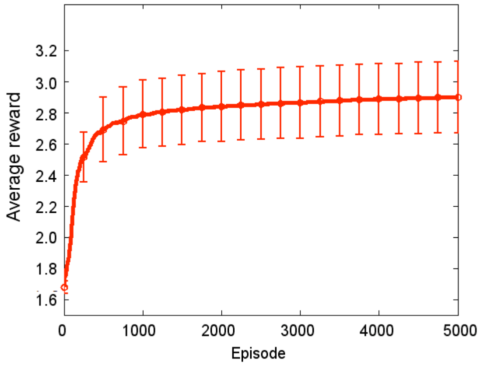

Fig. 4. Average reward

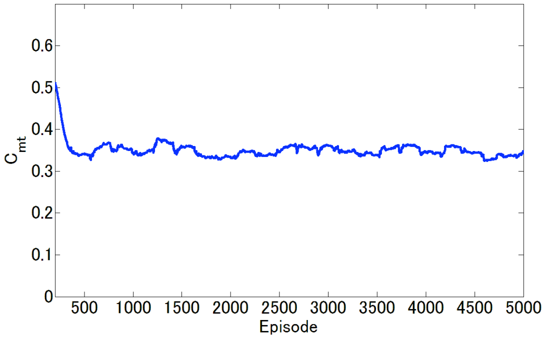

Fig. 5. Movement in C_{mt} (Moving Average over 100 episodes)

step,

$$\mathbf{x}_n = \left[\begin{array}{cc} \theta_n^{Hip} & \dot{\theta}_n^{Hip} \end{array} \right]. \tag{15}$$

In this study, we assume that $\mathbf{x}^*$ is the average of the Hip's angle and angular velocity without any disturbances during the episodes. $\mathbf{A}$ is solved by

$$\mathbf{A} = \mathbf{Y}\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}, \tag{16}$$

where,

$$\mathbf{X} = \left[\begin{array}{cc} \mathbf{x}_1 - \mathbf{x}^* & \mathbf{x}_2 - \mathbf{x}^* \cdots \mathbf{x}_i - \mathbf{x}^* \end{array} \right], \tag{17}$$

$$\mathbf{Y} = \left[\begin{array}{cc} \mathbf{x}_2 - \mathbf{x}^* & \mathbf{x}_3 - \mathbf{x}^* \cdots \mathbf{x}_{n+1} - \mathbf{x}^* \end{array} \right]. \tag{18}$$

If the eigenvalue of $\mathbf{A}$ exists within a unit circle, it means that the robot walking is stable.

C. Results

1) Learning of the torque-free period: Fig. 4 shows the average of the rewards for five trials. Error bars denote the standard deviation over 100 episodes for all the trials. Fig. 4 shows the average of C_{mt} over five trials. These figures indicate successful learning.

Fig. 6 shows the average of the torque-free start time t_1 and end time t_2. The torque-free time increases with each episode, suggesting that the energy-saving walking control is being acquired. Table V compares the energy efficiency of our method with that of other methods. According to this table, C_{mt} for our method, without torque-free times, is half that of CPG-based walking.

TABLE V
COMPARISON OF C_{mt}

	ASIMO	CPG without torque-free	CPG with torque-free	Human
C_{mt}	1.60	0.58	0.29	0.05

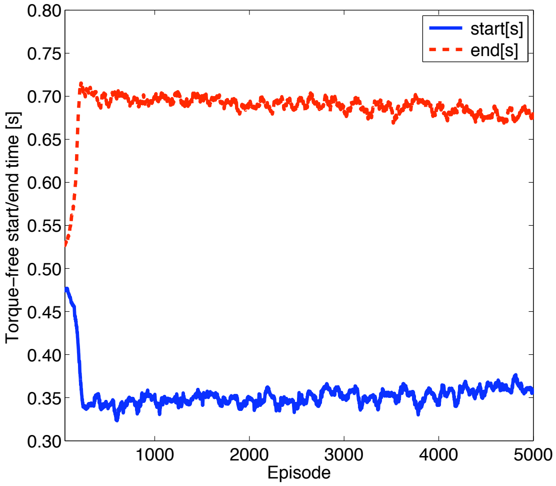

Fig. 6. Movement in Torque-free start time t_1 and end time t_2 (moving average over 100 episodes)

2) Stability analysis against disturbances: Table VI gives the eigenvalue for each disturbance pattern , shows that the CPG control with torque-free periods needs more steps to return to stable walking than that without the torque-free periods than that without the torque-free periods. Although CPG-based walking is relatively stable to our method when there are no disturbances. Our method enables more stable walking against disturbances.

Fig. 7 presents return maps. Colors are assigned corresponding to the number of steps from the time when disturbance was given. These figures depict the torque-free period contributes to the return to stable walking. The spread of plots from our method is clearly smaller than that under the forward disturbance condition. For the against condition, although both spreads appear to be the same, there is a difference. The red and green plots for the method without torque-free time spread haphazardly, whereas the same color plots for our method have similar features. Most green plots exist in the upper diagonal, and all orange plots exist under it. These results imply that torque-free time works as a stabilizer against disturbance.

3) Online adaptation to circumstances: Fig. 8 shows an example development of the torque-free period on a slope of 1.2 [deg], moving-averaged over 20 episodes, while Fig. 9 presents the corresponding development of walking distance through learning. These figures indicate that the walking

TABLE VI

EIGENVALUE OF RETURN MAP $\mathbf{A}$. ($\cdot$): ABSOLUTE VALUE

	without disturbance	forward disturbance	against disturbance
without torque-free	$0.064 \pm 0.101i$ (0.120)	0.318, -0.234	0.378, -0.109
with torque-free	-0.200, 0.129	$0.140 \pm 0.164i$ (0.216)	0.326, 0.0345

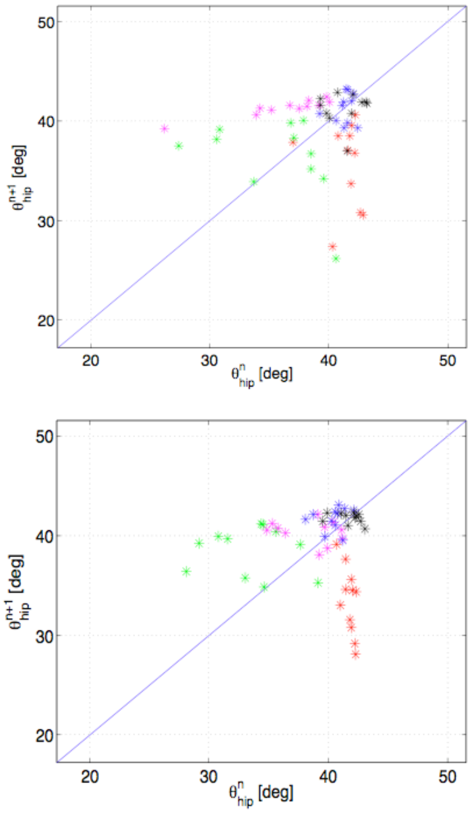

Fig. 7. Return map for a lateral disturbance: (upper) without torque-free time, and (lower) with torque-free time. Red: just after disturbance. Green: 1 step after disturbance. Purple: after 2 steps. Blue: after 3 steps. Black: after 4 steps.

distance was gradually improved, while the energy-efficiency was dramatically improved.

We additionally confirmed that the robot was able to climb up a steeper slope. These figures again indicate successful learning. Figs. 10 and 11 show the results obtained in the case of a slope of 2.0 [deg].

IV. CONCLUSION

In this paper, we proposed a method to improve the energy efficiency of a bipedal walking robot, control of which is based on CPG with reinforcement learning. We compared energy efficiency with the ZMP-based controller and conventional CPG-based controller, and showed that a conventional CPG-based walking resides between ZMP-based walking and human walking with respect to energy efficiency. We

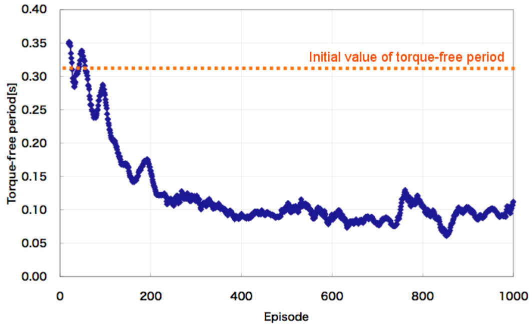

Fig. 8. Development of torque-free period (moving average over 20 episodes) on a slope

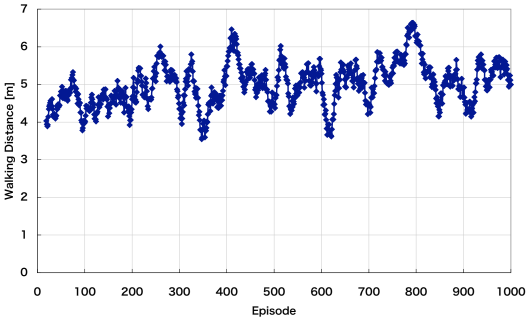

Fig. 9. Transition of walking distance (moving average over 20 episodes) on the slope

TABLE VII

CPG PARAMETERS FOR STEEPER SLOPE

α	u^a	u^η	u^r	κ
1.0	2.74	0.0	2.63	2.72

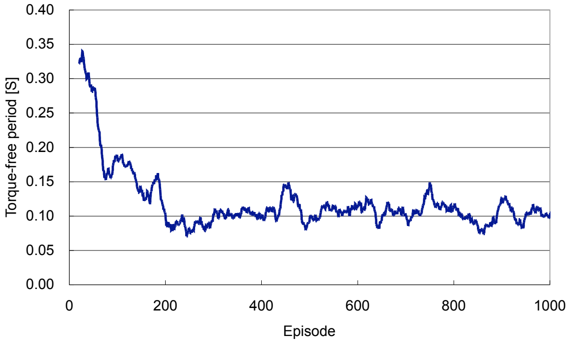

Fig. 10. Torque-free period on a slope of 2.0 [deg], moving-averaged over 20 episodes.

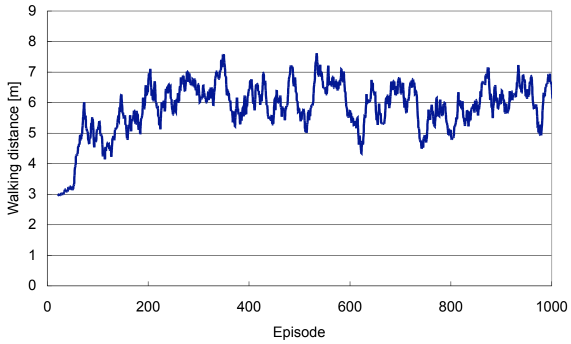

Fig. 11. Walking distance on a slope of 2.0 [deg], moving-averaged over 20 episodes.

showed that CPG-based control resides between ZMP-based control and QPD control with respect to energy efficiency. The key to our method is two-fold: (1) a torque-free period is introduced to the cycle of swing leg control, and (2) the inverse of the specific mechanical cost of transport C_{mt} is employed as the reward for reinforcement learning. We discussed its effect on energy-saving for physical reasons and demonstrated the performance of our method through simulation experiments with reinforcement learning. After learning on level ground, the value of C_{mt} was half that of the conventional CPG-based walking controller, and this is relatively close to that of a typical quasi-passive dynamic walking controller [19]. We also showed that our method did not alter the robustness against disturbance on level ground much. We further demonstrated that our method improved the ability to climb up slopes by decreasing the torque-free period, which decreased energy-efficiency but increased walking stability.

ACKNOWLEDGMENTS

This study was partly supported by Grant-in-Aid for Scientific Research (B) (No. 18300101) from the Japan Society for the Promotion of Science, and by the Artificial Intelligence Research Promotion Foundation.

References

[1] Y. Sugimoto and K. Osuka. Motion generate and control of quasi-passive-dynamic-walking based on the concept of delayed feedback control. In *Proceedings of 2nd International Symposium on Adaptive Motion of Animals and Machines*, 2003.

[2] Takashi Takuma, Seigo Nakajima, Koh Hosoda, and Minoru Asada. Design of self-contained biped walker with pneumatic actuators. In *SICE Annual Conference*, 2004.

[3] M. Wisse and J. Frankenhuyzen. Design and construction of mike; a 2d autonomous biped based on passive dynamic walking. In *2nd International Symposium on Adaptive Motion of Animals and Machines*, 2003.

[4] T. Hitomi, K. Shibata, Y. Nakamura, and S. Ishii. Reinforcement learning for quasi-passive dynamic walking of an unstable biped robot. *Robotics and Autonomous Systems*, 54:982–988, 2006.

[5] Tsuyoshi UENO, Yutaka NAKAMURA, Takashi TAKUMA, Tomohiro SHIBATA, Koh HOSODA, and Shin ISHII. Fast and stable learning of quasi-passive dynamic walking by an unstable biped robot based on off-policy natural actor-critic. *Intel ligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 5226–5231, 2006.

[6] M.Vukobratovic and J.Stepanenko. On the stability of anthropomorpihc systems. *Mathematical Biosciences*, 15:1–37, 1972.

[7] G.Taga, Y. Yamaguchi, and H. Shimizu. Self-organized control of bipedal loco-motion by neural oscillators in unpredictable environment. *Biological Cybernetics Biological Cybernetics*, 65:62–82, 1991.

[8] T. Matsubara, J. Morimoto, J. Nakanishi, M. Sato, and Kenji Doya. Learning cpg-based biped locomotion with a policy gradient method. *Robotics and Autonomous Systems*, 54:911–920, 2006.

[9] Y Nakamura, T Mori, Y Tokita, T. Shibata, and S. Ishii. Off-policy natural policy gradient method for a biped walking using a cpg controller. *J Robot Mechatron*, 17(16):636–644, 2005.

[10] S. Collins, A. Ruina, R. Tedrake, and M. Wisse. Efficient bipedal robots based on passive-dynamic walkers. *Science*, 307:1082–1085, 2005.

[11] T. McGeer. Passive dynamics walking. *The International Journal of Robotics Research*, 9(2):62–82, 1990.

[12] Tsuyoshi UENO, Yutaka NAKAMURA, Takashi TAKUMA, Tomohiro SHIBATA, Koh HOSODA, and Shin ISHII. Fast and stable learning of quasi-passive dynamic walking by an unstable biped robot based on off-policy natural actor-critic. *Intel ligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 5226–5231, 2006.

[13] S. Grillner and P. Wallen. Central pattern generators for locomotion, with special reference to vertebrates. *Annual Review of Neuroscience*, 8(1):233–261, 1985.

[14] Open Dynamic Simulator. http://www.ode.org.

[15] Steve Collins, Andy Ruina, Russ Tedrake, and Martijn Wisse. Walkers efficient bipedal robots based on passive-dynamic. *Science*, 307:1082–1084, 2005.

[16] A. Cohen, P. Holmes, and R. Rand. The nature of the coupling between segmental oscillators of the lamprey spinal generator for locomotion: A mathematical model. *Journal of Mathematical Biology*, 13:345–369, 1982.

[17] L. Weaver and N. Tao. The optimal reward baseline for gradient-based reinforcement learning. *In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 17:538–545, 2001.

[18] R. Tedrake, TW Zhang, and HS Seung. Stochastic policy gradient reinforcement learning on a simple 3D biped. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings*, volume 3.

[19] T. McGeer. Passive dynamic walking. *The International Journal of Robotics Research*, 9:62–82, 1990.