View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by IntechOpen

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

10

Reinforcement Learning of Stable Trajectory for Quasi-Passive Dynamic Walking of an Unstable Biped Robot

Tomohiro Shibata[1], Kentarou Hitomoi[3], Yutaka Nakamura[2] and Shin Ishii[1]
[1]Nara Institute of Science and Technology, [2]Osaka University, [3]DENSO CORPORATION
Japan

1. Introduction

Biped walking is one of the major research targets in recent humanoid robotics, and many researchers are now interested in Passive Dynamic Walking (PDW) [McGeer (1990)] rather than that by the conventional Zero Moment Point (ZMP) criterion [Vukobratovic (1972)]. The ZMP criterion is usually used for planning a desired trajectory to be tracked by a feedback controller, but the continuous control to maintain the trajectory consumes a large amount of energy [Collins, et al. (2005)]. On the other hand, PDW enables completely unactuated walking on a gentle downslope, but PDW is generally sensitive to the robot's initial posture, speed, and disturbances incurred when a foot touches the ground. To overcome this sensitivity problem, ``Quasi-PDW'' [Wisse & Frankenhuyzen (2003); Sugimoto & Osuka (2003); Takuma, et al. (2004)] methods, in which some actuators are activated supplementarily to handle disturbances, have been proposed. Because Quasi-PDW is a modification of the PDW, this control method consumes much less power than control methods based on the ZMP criterion. In the previous studies of Quasi-PDW, however, parameters of an actuator had to be tuned based on try-and-error by a designer or on *a priori* knowledge of the robot's dynamics. To act in non-stationary and/or unknown environments, it is necessary for robots that such parameters in a Quasi-PDW controller are adjusted autonomously in each environment.

In this article, we propose a reinforcement learning (RL) method to train a controller designed for Quasi-PDW of a biped robot which has knees. It is more difficult for biped robots with knees to walk stably than for ones with no knees. For example, Biped robots with no knee may not fall down when it is in an open stance, while robots with knees can easily fall down without any control on the knee joints.

There are, however, advantages of biped robots with knees. Because it has closer dynamics to humans, it may help to understand human walking, and to incorporate the advantages of human walking into robotic walking. Another advantage is that knees are necessary to prevent a swing leg from colliding with the ground. In addition, the increased degrees of freedom can add robustness given disturbances such as stumbling.

Open Access Database www.i-techonline.com

Source: Humanoid Robots: Human-like Machines, Book edited by: Matthias Hackel
ISBN 978-3-902613-07-3, pp. 642, Itech, Vienna, Austria, June 2007

Our computer simulation shows that a good controller which realizes a stable Quasi-PDW by such an unstable biped robot can be obtained with as small as 500 learning episodes, whereas the controller before learning has shown poor performance.
In an existing study [Tedrake, et al. (2004)], a stochastic policy gradient RL was successfully applied to a controller for Quasi-PDW, but their robot was stable and relatively easy to control because it had large feet whose curvature radius was almost the same as the robot height, and had no knees. Their robot seems able to sustain its body even with no control. Furthermore, the reward was set according to the ideal trajectory of the walking motion, which had been recorded when the robot realized a PDW. In contrast, our robot model has closer dynamics to humans in the sense that there are smaller feet whose curvature radius is one-fifth of the robot height, and knees. The reward is simply designed so as to produce a stable walking trajectory, without explicitly specifying a desired trajectory. Furthermore, the controller we employ performs for a short period especially when both feet touch the ground, whereas the existing study above employed continuous feedback control. Since one definition for Quasi-PDW is to emit intermittent control signals as being supplementary to the passivity of the target dynamics, a design of such a controller is important.
The rest of the article is organized as follows. Section 2 outlines our approach. Section 3 introduces the details of the algorithm using policy gradient RL as well as simulation setup. Section 4 describes simulation results. We discuss in section 5 with some directions in future work.

2. Approach Overview

Fig. 1 depicts the biped robot model composed of five links connected by three joints: a hip and two knees. The physical parameters of the biped robot model are shown in Table 1. The motions of these links are restricted in the sagittal plane. The angle between a foot and the corresponding shank is fixed. Because we intend to explore an appropriate control strategy based on the passive dynamics of the robot in this study, its physical parameters are set referring to the existing biped robots that produced Quasi-PDW [Wisse & Frankenhuyzen (2003); Takuma, et al. (2004)]. As described in Fig. 1, θ stands for the absolute angle between the two thighs, θ_{knee1} and θ_{knee2} denote the knee angles, and ω denotes the angular velocity of the body around the point at which the stance leg touches the ground. The motion of each knee is restricted within [0, π/4] [rad].

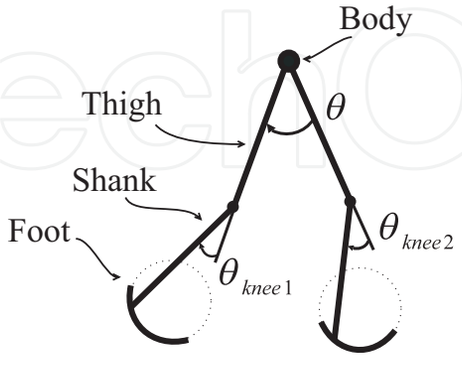


Figure 1. 2D Biped Model

	Length [m]	Mass [kg]
Body	0.0	4.0
Thigh	0.33	0.50
Shank	0.33	0.50
Foot	0.13*	0.020

Table 1. Physical parameters of the robot. ∗ Value of curvature radius

Our approach to achieving adaptive controls consists of the following two stages.
(1) The two knees are locked, and the initial posture which realizes PDW by this restricted system are searched for. The initial posuture is defined by the initial absolute angle between two thighs, θ^s, and the initial angular velocity of the body around the point at which the stance leg touches the ground, ω^s. These values are used for the initial setting of the robot in the next stage.
(2) The two knees are then unlocked, and the robot is controlled by an intermittent controller with adjustable parameters. The parameters are modified by reinforcement learning (RL) so that the robot keeps stable walking.
These two stages are described in detail in the followings.

2.1 Searching for the initial conditions

In the first stage, we searched for an initial posture, denoted by θ^s and ω^s, which realize PDW by the robot with the locked knees, on a downslope with a gradient of $\varepsilon = 0.03$ [rad]. For simplicity, we fixed $\theta^s = \pi/6$ [rad] and searched a region from 0 to π [rad/sec] by $\pi/180$ [rad/sec], for ω^s that maximizes the walking distance. The swing leg of compass-like biped robots which have no knees necessarily collides with the ground, leading to falling down. Thus, in this simulation, the collision between the swing leg and the ground was ignored. We found $\omega^s = 58 \times \pi/180$ [rad/sec] was the best value such to allow the robot to walk for seven steps.

2.2 Design of a Controller

In light of the design of control signals for the existing Quasi-PDW robots, we apply torque inputs of a rectangular shape to each of the three joints (cf. Fig. 2). One rectangular torque input applied during one step is represented by a fourdimensional vector $\boldsymbol{\tau} = \{\tau_{Hip,Amp}, \tau_{Hip,Dur}, \tau_{Kne,Flx}, \tau_{Kne,Ext}\}$. $\tau_{Hip,Amp}$ and $\tau_{Hip,Dur}$ denote the amplitude and the duration of the torque applied to the hip joint, respectively, and $\tau_{Kne,Flx}$ and $\tau_{Kne,Ext}$ are the amplitude of torques that flex and extend the knee joint of the swing leg, respectively. The manipulation of the knees follows the simple scheme described below to avoid the collision of the swinging foot with the ground, so that a swing leg is smoothly changed into a stance leg (cf. Fig. 2). First, the knee of the swing leg is flexed with $\tau_{Kne,Flx}$ [Nm] when the foot of the swing leg is off the ground (Fig. 2(b)). This torque is removed when the foot of the swing leg goes ahead of that of the stance leg (Fig. 2(c)), and, in order to make the leg extended, a torque of $-\tau_{Kne,Ext}$ is applied after the swing leg turns into the swing down phase from the swing up phase according to its passive dynamics (Fig. 2(d)). To keep the knee joint of the stance leg being extended, 1 [Nm] is applied to the knee joint. $\boldsymbol{\tau}$ is assumed to be distributed as a

Gaussian noise vector, while the mean vector $\bar{\tau}$ is modified by the learning, as described in the next section.

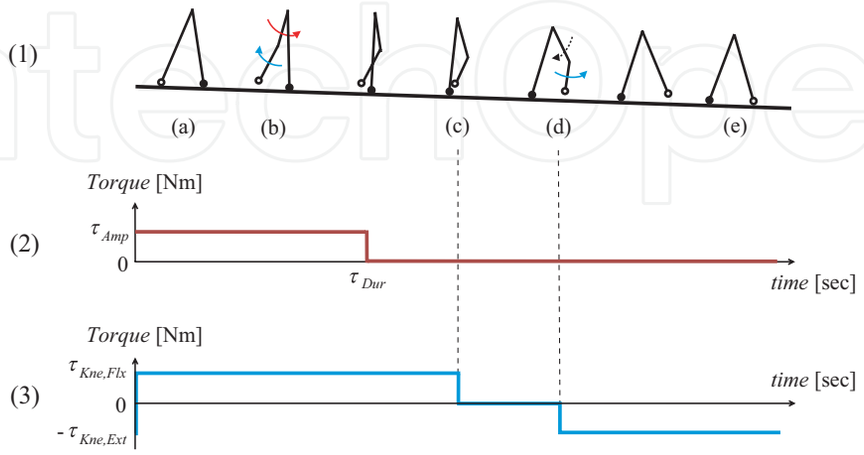


Figure 2. Torque applied to the hip joint and the knee joint. (1) Motions of the swing leg during a single step. (2) Torque applied to the hip joint. (3) Torque applied to the knee joint of the swing leg. (a) A single step starts when both feet touch the ground. (b) After the swing leg is off the ground, the robot begins to bend the knee of the swing leg by applying a torque of $\tau_{Kne,Flx}$ [Nm]. (c) The torque to the knee is removed when the foot of the swing leg goes ahead of that of the stance leg. (d) When the thigh of the swing leg turns into the swing down phase from the swing up phase, a torque of $\tau_{Kne,Ext}$ [Nm] is applied in order to extend the swing leg. (e) The swing leg touches down and becomes the stance leg

3. Learning a Controller

3.1 Policy gradient reinforcement learning

In this study, we employ a stochastic policy gradient method [Kimura & Kobayashi (1998)] in the RL of the controller's parameter $\bar{\tau}$, by considering the requirement that the control policy should output continuous values. The robot is regarded as a discrete dynamical system whose discrete time elapses when either foot touches the ground, i.e., when the robot takes a single step. The state variable of the robot is given by $s_n = (\theta_n, \omega_n)$, where n counts the number of steps, and θ_n and ω_n stand for the absolute angle between two thighs at the n-th step and the angular velocity of the body around the point at which the stance leg touches the ground, respectively.

At the onset of the n-th step, the controller provides a control signal τ_n, which determines the control during the step, according to a probabilistic policy $\pi(\tau \mid \bar{\tau})$. At the end of this step, the controller is assumed to receive a reward signal r_n. Based on these signals, a temporal-difference (TD) error δ is calculated by

$$\delta = \{r_n + \gamma V(s_{n+1})\} - V(s_n), \quad (1)$$

where $\gamma (0 \le \gamma \le 1)$ is the discount rate. V denotes the state value function and is trained by the following TD(0)-learning:

$$V(s_n) = V(s_n) + \alpha\delta \quad (2)$$

$$e_n = \frac{\partial}{\partial \bar{\tau}} \ln\left(\pi\left(\tau|\bar{\tau}\right)\right)\Big|_{\tau=\tau_n,\bar{\tau}=\bar{\tau}_n} \tag{3}$$

$$D_n \leftarrow e_n + \beta D_{n-1} \tag{4}$$

$$\bar{\tau}_{n+1} = \bar{\tau}_n + \alpha_p \delta D, \tag{5}$$

where $\boldsymbol{e}$ is the eligibility and $\boldsymbol{D}$ is the eligibility trace. β ($0 \le \beta \le 1$) is the diffusion rate of the eligibility trace and a_p is the learning rate of the policy parameter. After policy parameter $\bar{\tau}_n$ is updated into $\bar{\tau}_{n+1}$, the controller emits a new control signal according the new policy $\pi(\tau \mid \bar{\tau}_{n+1})$. Such a concurrent on-line learning of the state value function and the policy parameter is executed until the robot tumbles (we call this period an episode), and the RL proceeds by repeating such episodes.

3.2 Simulation setup

In this study, the stochastic policy is defined as a normal distribution:

$$\pi\left(\tau|\bar{\tau}\right) = \frac{1}{\left(2\pi^2\right)|\Sigma|^{1/2}} \times \exp\left\{-\frac{1}{2}\left(\tau-\bar{\tau}\right)^T \Sigma^{-1}\left(\tau-\bar{\tau}\right)\right\} \tag{6}$$

so that the covariance $\boldsymbol{\Sigma}$ is given by

$$\Sigma = \begin{pmatrix} \sigma_{Hip,Amp^2} & 0 & 0 & 0 \\ 0 & \sigma_{Hip,Dur^2} & 0 & 0 \\ 0 & 0 & \sigma_{Kne,Flx^2} & 0 \\ 0 & 0 & 0 & \sigma_{Kne,Ext^2} \end{pmatrix}, \tag{7}$$

where $\sigma_{Hip,Amp}$, $\sigma_{Hip,Dur}$, $\sigma_{Kne,Flx}$ and $\sigma_{Kne,Ext}$ are constant standard deviations of noise, set at 0.3, 0.05, 0.3 and 0.3, respectively. We assume each component of τ is 0 or positive, and if it takes a negative value accidentally it is calculated again, similarly to the previous study [Kimura, et al. (2003)]. The reward function is set up as follows. If a robot walks stably, ω_n and θ_n should repeat similar values over steps. Furthermore, the robot should take no step in the same place, i.e., θ_{n+1} needs to be large enough. To satisfy these requirements, we define the reward function as

$$r_n = \theta_{n+1} \exp\left(-\left|\theta_{n+1} - \theta_n\right|^2\right). \tag{8}$$

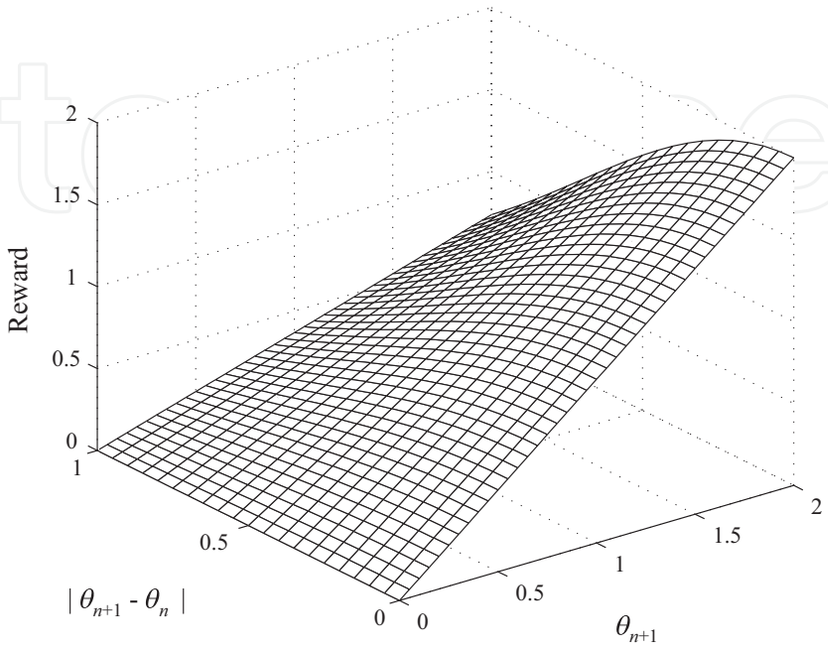


Figure 3. Landscape of the reward function

Figure 3. shows the landscape of this reward function. The value function is represented by a table over grid cells in the state space, and the value for each grid cell is updated by equation (2). In this study, we prepared 10 grid cells; the center of the fifth cell was for θ^s (Fig. 4), and the grid covered the whole state space, by assigning the 0-th cell to the range: $\theta < 0$, and the 9-th cell to the range: $\theta > 2\theta^s$. We used $a = 0.5$, $a_p = 0.01$, and $\beta = \gamma = 0.95$

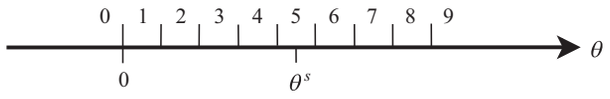


Figure 4. Discretization of the state space

In this study, we used a 3D dynamics simulator, Open Dynamics Engine [ODE]. In simulation experiments, motions of the robot were restricted in the sagittal plane by configuring a symmetric robot model with nine links (Fig. 5). It should be noted this nine-links robot has equivalent dynamics to the five-links model (Fig. 1), under the motion restriction in the sagittal plane; this nine-links model was also adopted by Wisse and Frankenhuyzen (2003) and by Takuma et al. (2004).

4. Simulation Results

Although the physical parameters of our robot were set referring to the existing Quasi-PDW robots, our robot with unlocked knees was not able to produce stable walking by itself. Then, this section describes the way to train the controller according to our RL scheme.

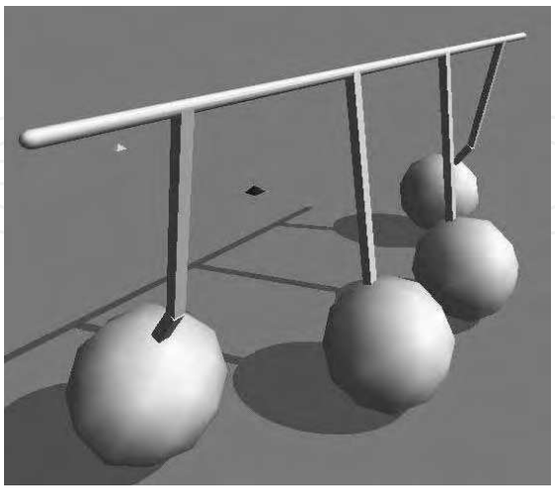


Figure 5. Dynamics simulation of the nine-links model with ODE

4.1 Passive walking without learning

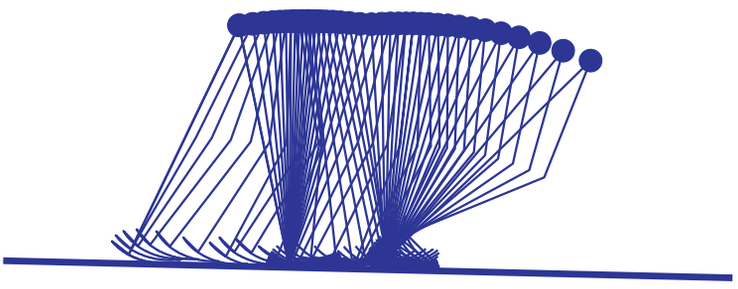


Figure 6. Stick diagram of the passive motion by the robot with knees. Plot intervals are 50 [ms]

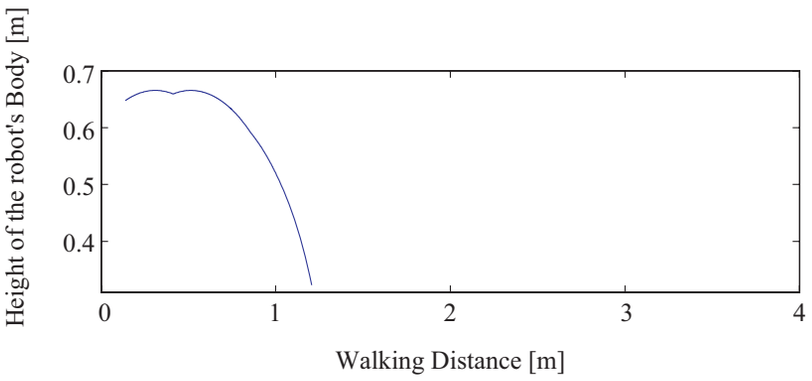


Figure 7. Body's trajectory of the passive robot with knees

First, we examined whether the robot with unlocked knees was able to produce stable walking on a downslope with ε = 0.03 [rad], when it received no controls to the hip joint or the knee joints. The unlocked knees were controlled in the same manner as that described in section 2.2. Initial conditions were set at $\theta_0 = \theta^s$[rad] and $\omega_0 = \omega^s$ [rad/sec], which are the same as those where the knee-locked model performed seven steps walking. As Fig. 7 shows, the robot with unlocked knees walked for 80 cm and then fell down. The robot was not able to walk passively when the knees were unlocked but uncontrolled.

4.2 Learning a controller

The experiment in section 4.1 showed that the robot with unlocked knees was not able to produce stable walking without any control to the hip joint or the knee joints, even when starting from possibly good initial conditions θ^s and ω^s. Then, in this section, we applied on-line RL to the automatic tuning of the parameter $\bar{\tau}$. At the beginning of each episode, the robot's initial conditions were set at $\theta_0 = \theta^s$, $\omega_0 = \omega^s$, and the episode was terminated either when the robot walked for 20 steps or fell down. When the height of the robot's 'Body' became smaller than 80% of its maximum height, it was regarded as a failure episode (falling down). RL was continued by repeating such episodes.

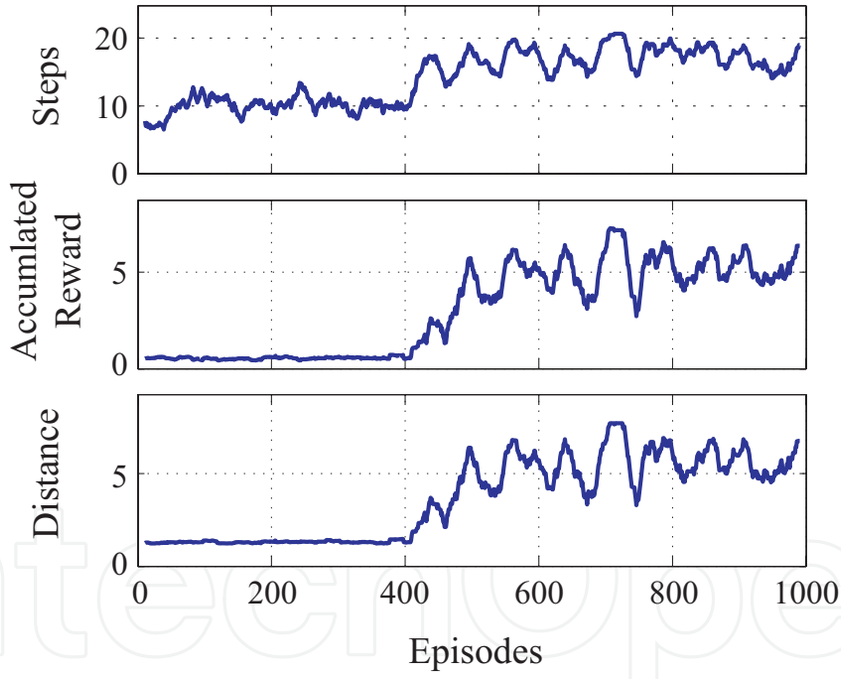


Figure 8. Moving averages of number of steps, cumulative reward, and distance to have walked

Fig. 8 shows the moving averages for ±20 episodes of walking steps (top), cumulative reward (middle), and walking distance (bottom), achieved by the robot. The steps increased after about 400 learning episodes, and went up to nearly 20 steps after about 500 learning episodes. In the early learning stage, the cumulative reward and walked distance were small

though the robot walked for more than 10 steps, indicating the robot was walking stumbling with small strides. Using the deterministic controller with the parameter $\bar{\tau}$ after 500 training episodes, the robot was able to walk for more than 20 steps (Fig. 9). The parameter at this time was $\bar{\tau}$ = (0.70, 0.17, 0.93, 0.51).

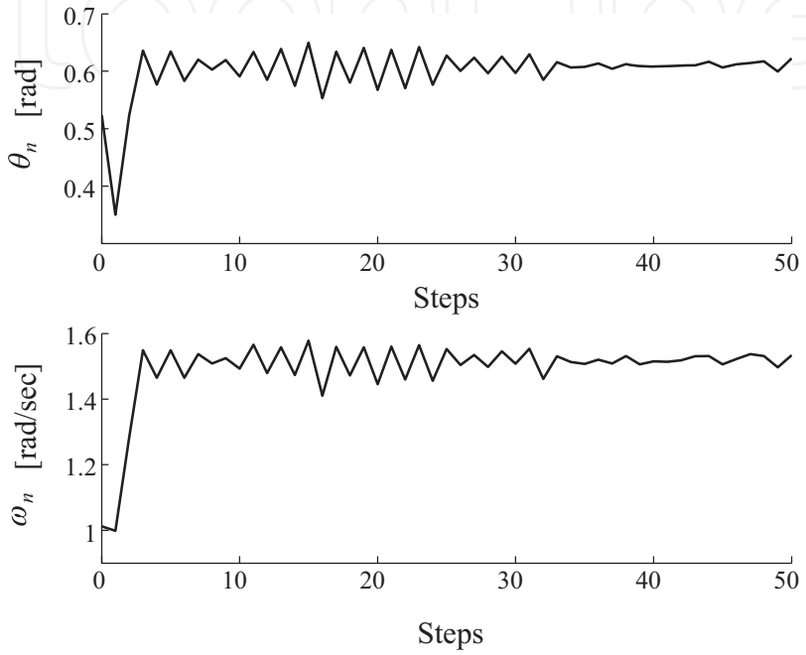


Figure 9. Values of θ_n and ω_n during the walking for 50 steps by the controller after 500 learning episodes

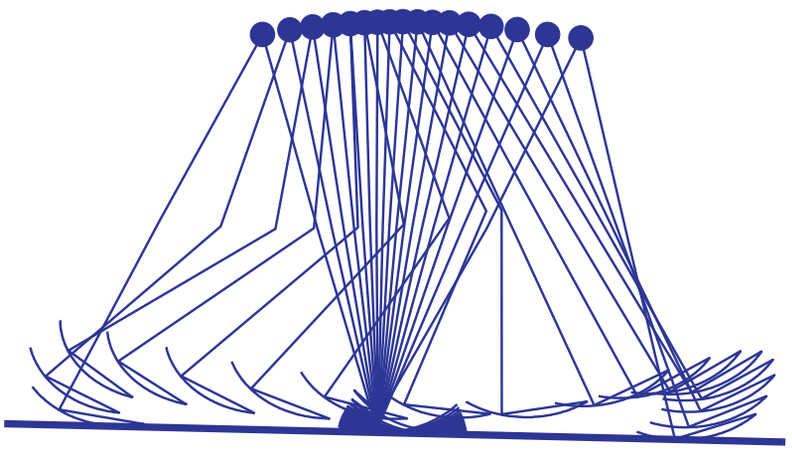


Figure 10. Stick diagram motion by the robot after 500 learning episodes. Plot intervals are 50 [ms]

4.3 Energy efficiency

	This study	Human*	Dynamite*	Asimo*
c_{mt}	0.093	0.05	0.04	1.6

Table 2. Energy efficiency calculated as c_{mt}. *These values are excerpted from literature [12]

Table 2 compares energy efficiency of Quasi-PDW acquired in this study with the others. For this comparison, the dimension-less cost of transport,

$$c_{mt} = (\text{used energy})/(\text{weight} \times \text{distance}) \quad (9)$$

was employed [Collins, et al. (2005)]; c_{mt} is suitable for comparing energy efficiency of simulators with that of real robots, because c_{mt} evaluates the effectiveness of the mechanical design and controller independently of the actuator efficiency. Note that energy from the gravity is included in the calculation of c_{mt} (= 0.093) for our simulated robot. The c_{mt} value achieved by our on-line RL is larger than the one of the PDWcontrolled robot (Dynamite), while it is much smaller than the one of the ZMP criterion (ASIMO).

4.4 Robustness against disturbances

To see the robustness of the acquired Quasi-PDW against possible disturbances from the environment, we conducted two additional experiments.

First, we let the robot with the control parameter after 500 training episodes walk on downslopes with various gradients. Fig. 11 shows the results for ε = 0.02 - 0.05 [rad]. The robot was able to walk for more than 50 steps on downslopes with ε = 0.02 - 0.04 [rad], and 22 steps with 0.05 [rad]; the controller acquired through our on-line RL was robust against the variation (in the gradient) of the environment. Second, we applied impulsive torque inputs to the hip joint during walking. Fig. 12 shows the time-series of θ_n in the same condition as Fig. 9, except that impulsive torque inputs were applied as disturbances at the time points with the arrows. Each disturbance torque was 1 [Nm] and was applied so as to pull the swing leg backward for 0.1 [sec] when 0.4 [sec] elapsed after the swing leg got off the ground. As this figure shows, θ_n recovered to fall into the stable limit cycle within a few steps after disturbances, implying that the attractor of the acquired PDW is fairly robust to noise from the environment.

Additional qualitative analysis by means of return map was performed in order to investigate changes in walking robustness through learning. Fig. 15 plots return maps during walking after disturbances as well as steady state walking at 440 and 500 episodes, respectively. In this figure, the return map is depicted by circles, crosses, and triangles for right after disturbance, next step, and two steps after, respectively (cf. Fig. 14). The maps for steady-state walking (Fig. 15(a),(c)) show the robot was walking stably by keeping θ_n at around 0.6 [rad] in both cases of after 440 and 500 learning episodes. The upper part of Table 3 shows $<\theta_n>$, the average θ_n during steady-state walking, after 440, 480, 500, 700, and 900 learning episodes. $<\theta_n>$ gets large after 500 episodes, which would be induced by the increase in the accumulated reward (Eq. 8). This reward increase was mainly due to the increase in the step length, even after θ_{n+1}-θ_n became almost zero achieved by making periodic walking.

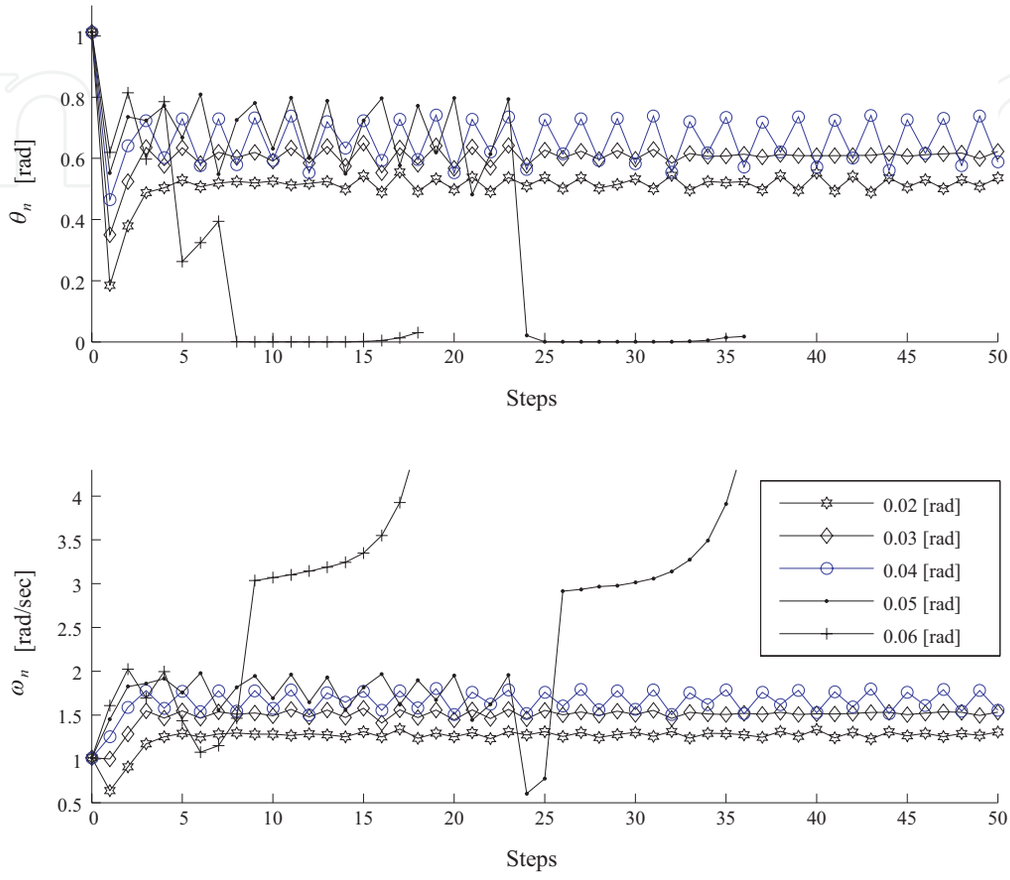


Figure 11. Values of θ_n and ω_n on downslopes with various gradients

The maps after disturbance (Fig. 15(b),(d)) show the disturbed walking recovered quickly the steady-state walking. The lower part of Table 3 shows the average step required for recovery did not decrease in a monotonic fashion, but they are all small enough regardless of the gradual increase in $<\theta_n>$ through learning.

Episodes	440	480	500	700	900
$<\theta_n>$ during steady-state walking	0.5960	0.5957	0.6043	0.6100	0.6120
Average steps for recovery	2.071	2.253	1.182	1.556	2.020

Table 3. Mean values of θ_n during steady walking and of numbers of steps necessary to recover steady walking after the disturbance

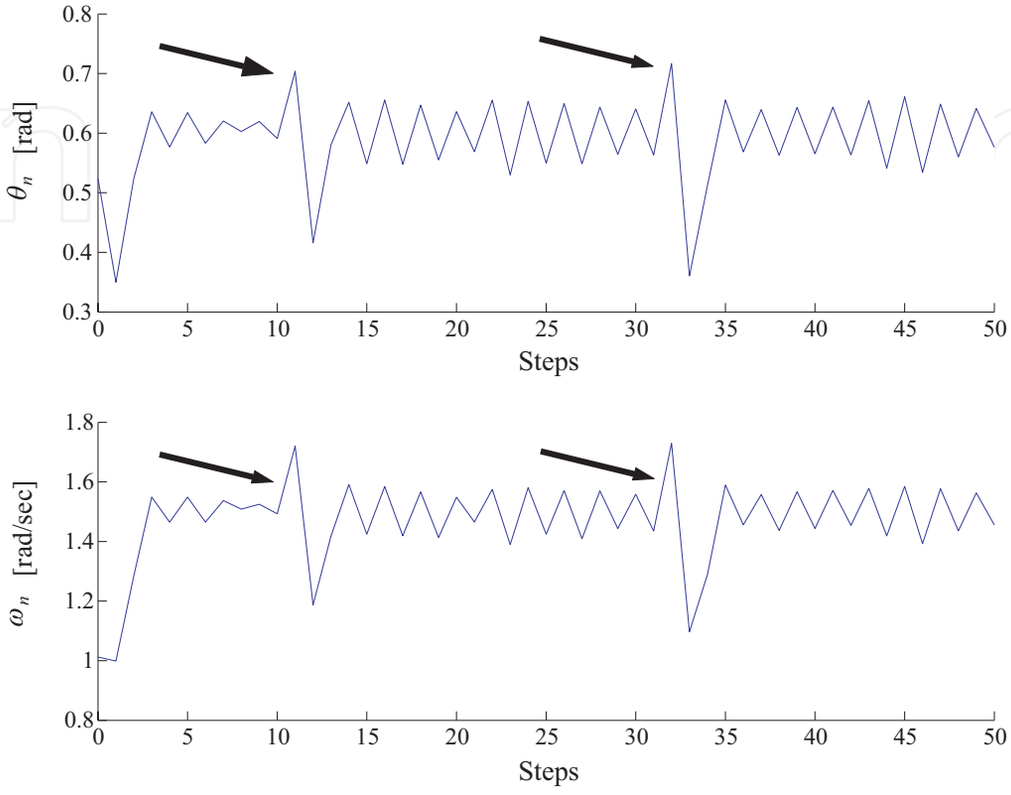


Figure 12. Perturbation of θ against impulsive disturbances (arrowed)

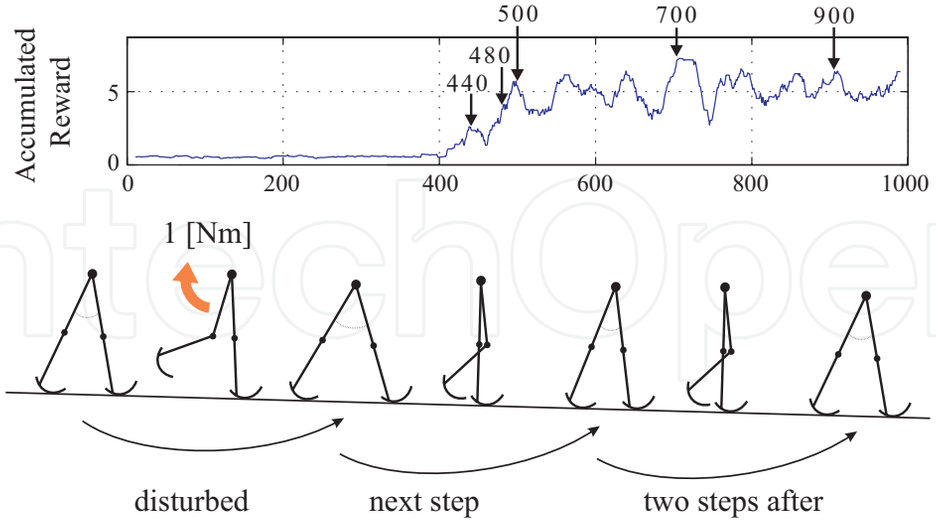


Figure 14. Step counts after disturbance

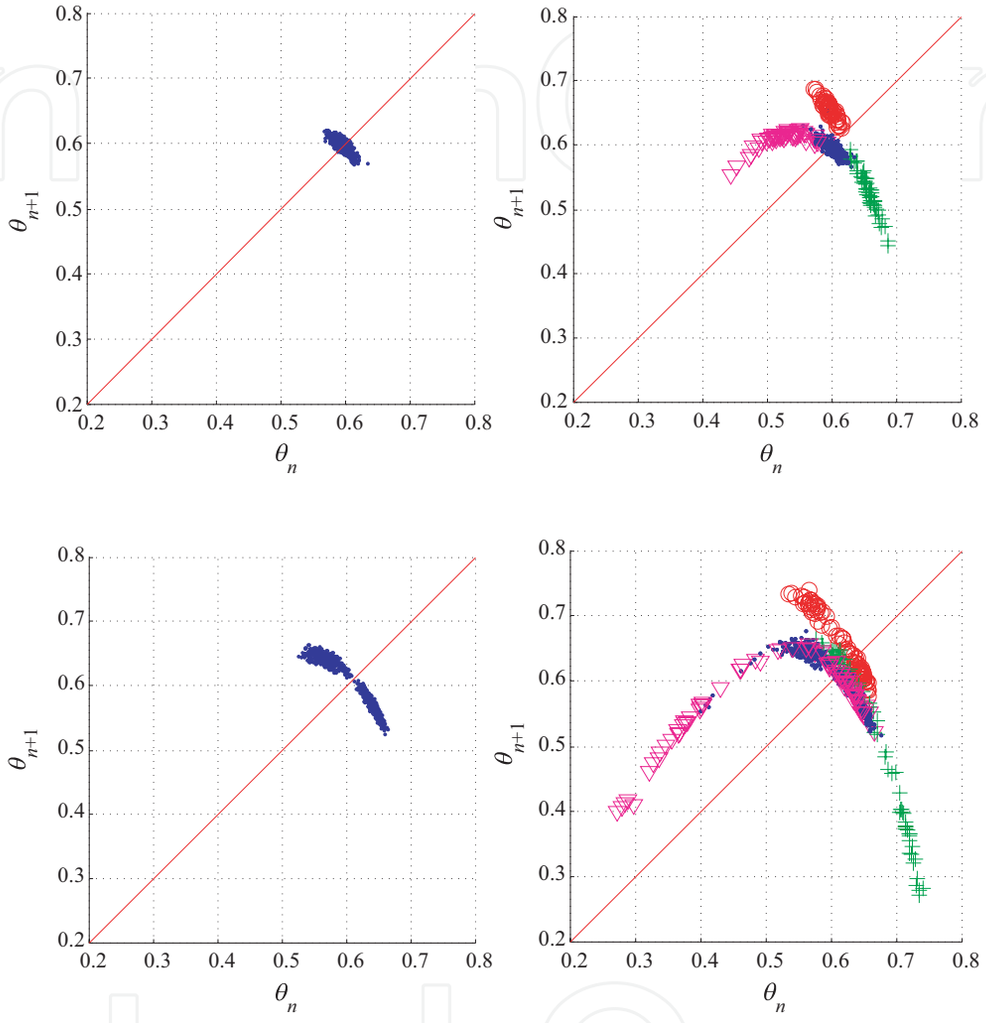


Figure 15. Return maps after 440 or 500 learning episodes

5. Discussion

In this study, we proposed an on-line RL method suitable for Quasi-PDW by a 2D biped robot, whose possession of knees makes the system unstable. Our study is underlaid by the perspective of low energy consumption and good correspondence to human walking. RL was applied not only for the hip joint but also for the knee joints of the robot, and our learning method was successful in making the unstable robot produce stable walking after as small as 500 training episodes, despite of usage of simple intermittent controllers. Although the controller itself was simple, simulation experiments on downslopes with

various gradients and through addition of impulsive disturbances have shown that the stochastic policy gradient method with a reward function that encourages continuing rhythmic walking steps have successfully contributed to making the PDW robust against various noise in the environment.

Our learning method consisted of two stages, as described in section 2. After roughly searching in the first stage for an initial angular velocity with which the robot with locked knees walked for several steps, RL was applied to the robot with unlocked knees, starting from the initial condition obtained in the first stage. This two-stages learning reminds us of a developmental progression found at least in humans [Bernstein (1968); Newell & Vaillancourt (2001)] which increases the degree of freedoms as the learning proceeds; after a primitive control is achieved for a restricted system with a low dimensionality, frozen dimensionality is gradually released to realize more complex and smooth movements by the high-dimensional system. Furthermore, animals seem to employ different controllers in the initiation phase and in the maintenance phase for effect e.g., it has been known that three steps in average are required to initiate stationary walking in humans [Miller & Verstraete (1996)]. We consider the first stage of our approach could correspond to the initiation stage above.

As another reason for our successful results, our adaptive controller was trained by RL as to apply intermittent energy for maintaining stable PDW. This intermittent control was inspired by the studies of the measurement of human EMG [Basmajian (1976)] and robot control based on the idea of Quasi-PDW conducted by Collins et al. (2005) or by Takuma et al. (2004). To develop an energy-efficient control method of robots, considerable care about the passivity of the robot should be taken, as Collins suggested. Furthermore, the dynamics of robots with many degrees of freedom generally constitutes a nonlinear continuous system, and hence controlling such a system is usually very difficult. Our approach successfully realized efficient learning, which required as small as 500 learning episodes even with learning for knee joints, by introducing the policy that emits intermittent control signals and a reward function encouraging stable motions, both of which well utilized the passivity of the robot. Our learning method is not restricted to locomotion, since the computational problem and the importance of passivity are both general, although what kind of controllers should be activated or switched when and how are remained as interesting and significant problems for general applicability. From a theoretical point of view, our results indicate that passivity of the robot together with the two-stages scheme effectively restricted a high-dimensional control space of the robot. Nakamura et al. demonstrated that an RL in which a central pattern generator (CPG) was employed succeeded in training a simulated biped robot which had also five links including knees [Nakamura, et al. (2004)]. In their method, the control space was restricted such that the outputs of the controller were likely rhythmic control signals. Combination of such CPG-constrained learning scheme and the passivity constrained learning scheme would be interesting not only for more robust locomotion but also for control of various types of high-dimensional robots. It should also be noted here that CPG seems to be employed for human locomotion [Dietz, et al. (2002)]. Our approach would be plausible in the perspective of energy efficiency and understanding of human walking [Basmajian (1976)]. Along with this issue, how to incorporate the idea of energy efficiency into the reward function is interesting. Another interesting avenue for future work is to devise a method to produce stable walking on a level ground. In addition, we are conducting experiments with a real

biped robot [Ueno, et al. (2006)], which would enhance the applicability of the current methodological study.

6. Acknowledgement

We thank Dr. Koh Hosoda and Mr. Takuma at Graduate School of Engineering, Osaka University, for giving us information about Passive DynamicWalking and their biped robot. This study is partly supported by Graint-in-Aid for Scientific Research of Japan Society for the Promotion of Science, No. 16680011 and 18300101.

7. References

Basmajian, J. (1976). *The human bicycle: an ultimate biological convenience*. The Orthopedic clinics of North America, 7, 4, 1027–1029.

Bernstein, N. (1968). *The coordination and regulation of movements*. Pergamon.

Miller, C.A. & Verstraete, M.C. (1996). Determination of the step duration of gait initiation using a mechanical energy analysis. *Journal of Biomechanics*, 29, 9, 1195–1199.

Collins, S.H. & Ruina, A. (2005). A bipedal walking robot with efficient and human-like gait. *Proceedings of IEEE International Conference on Robotics and Automation*.

Collins, S; Ruina, A; Tedrake & M.Wisse. (2005). Efficient bipedal robots based on passivedynamic walkers. *Science*, 307, 1082–1085.

Dietz, V.; Muller, R & Colombo, G. (2002). Locomotor activity in spinal man: significance of afferent input from joint and load receptors, *Brain*, 125, 2626-2634.

Kimura, H. & Kobayashi, S. (1998). Reinforcement learning for continuous action using stochastic gradient ascent. *Intelligent Autonomomus Systems*, 288–295.

Kimura, H. & Kobayashi, S. (1998). An analysis of actor/critic algorithms using eligibility traces: Reinforcement learning with imperfect value function. *Proceedings of 15th International Conference on Machine Learning*, 278–286.

Kimura, H.; Aramaki, T. & Kobayashi, S. (2003). A policy representation using weighted multiple normal distribution. *Journal of the Japanese Society for Artificial Intelligence*, 18, 6, 316–324.

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

Newell, K.M. & Vaillancourt, D.E. (2001). Dimensional change in motor learning. *Human Movement Science*, 20, 695–715.

Nakamura, Y.; Mori, T. & Ishii, S. (2004). Natural policy gradient reinforcement learning for a CPG control of a biped robot. *Proceedings of International conference on parallel problem solving from nature*, 972–981.

ODE, http://ode.org/.

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

Takuma, T.; Nakajima, S.; Hosoda, K. & Asada, M. (2004). Design of self-contained biped walker with pneumatic actuators. *Proceedings of SICE Annual Conference*.

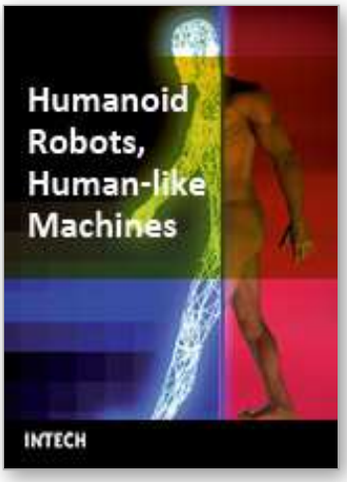
Tedrake, R.; Zhang, T.W.. & Seung,H.S. (2004). Stochastic policy gradient reinforcement learning on a simple 3D biped. *Proceedings of the IEEE International Conference on Intelleigent Robots and Systems*.

Ueno, T.; Nakamura, Y.; Shibata, T.; Hosoda, K. & Ishii, S. (2006). Fast and stable learning of quasi-passive dynamic walking by an unstable biped robot based on off-policy natural actor-critic. *Proceedings of IEEE/RSJ Int Conf Intell Robot Syst*, 5226-5231.

Vukobratovic, M. & Stepanenko, J. (1972).On the stability of anthropomorphic systems. *Mathematical Biosciences*, 15, 1–37.

Wisse, M. & Frankenhuyzen, J. (2003). Design and construction of mike; a 2D autonomous biped based on passive dynamic walking. *Proceedings of 2nd International Symposium on Adaptive Motioon of Animals and Machines*.

IntechOpen



Humanoid Robots, Human-like Machines

Edited by Matthias Hackel

ISBN 978-3-902613-07-3

Hard cover, 642 pages

Publisher I-Tech Education and Publishing

Published online 01, June, 2007

Published in print edition June, 2007

In this book the variety of humanoid robotic research can be obtained. This book is divided in four parts: Hardware Development: Components and Systems, Biped Motion: Walking, Running and Self-orientation, Sensing the Environment: Acquisition, Data Processing and Control and Mind Organisation: Learning and Interaction. The first part of the book deals with remarkable hardware developments, whereby complete humanoid robotic systems are as well described as partial solutions. In the second part diverse results around the biped motion of humanoid robots are presented. The autonomous, efficient and adaptive two-legged walking is one of the main challenge in humanoid robotics. The two-legged walking will enable humanoid robots to enter our environment without rearrangement. Developments in the field of visual sensors, data acquisition, processing and control are to be observed in third part of the book. In the fourth part some "mind building" and communication technologies are presented.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Tomohiro Shibata, Kentarou Hitomoi, Yutaka Nakamura and Shin Ishii (2007). Reinforcement Learning of Stable Trajectory for Quasi-Passive Dynamic Walking of an Unstable Biped Robot, Humanoid Robots, Human-like Machines, Matthias Hackel (Ed.), ISBN: 978-3-902613-07-3, InTech, Available from: http://www.intechopen.com/books/humanoid_robots_human_like_machines/reinforcement_learning_of_stable_trajectory_for_quasi-passive_dynamic_walking_of_an_unstable_biped_r

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2007 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen