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

Consideration on Robotic Giant-swing Motion Generated by Reinforcement Learning

M. Hara, *Member, IEEE*, N. Kawabe, N. Sakai, J. Huang, *Senior Member, IEEE*,
Hannes Bleuler, *Member, IEEE*, and T. Yabuta, *Senior Member, IEEE*

Abstract—**This study attempts to make a compact humanoid robot acquire a giant-swing motion without any robotic models by using reinforcement learning; only the interaction with environment is available. Generally, it is widely said that this type of learning method is not appropriated to obtain dynamic motions because Markov property is not necessarily guaranteed during the dynamic task. However, in this study, we try to avoid this problem by embedding the dynamic information in the robotic state space; the applicability of the proposed method is considered using both the real robot and dynamic simulator. This paper, in particular, discusses how the robot with 5-DOF, in which the Q-Learning algorithm is implemented, acquires a giant-swing motion. Further, we describe the reward effects on the Q-Learning. Finally, this paper demonstrates that the application of the Q-Learning enable the robot to perform a very attractive giant-swing motion.**

I. Introduction

ACQUISITION of robotic motions by applying learning methods is a very attractive theme in robotics. In previous studies, several control algorithms with learning methods have been proposed, such as adaptive control and neural-network control; most of them are categorized into a supervised learning. However, few studies have reported how the robot acquires the optimized motion form, which comprises several action patterns, in the learning process. For example, Doya debated the acquisition of robotic walking due to reinforcement learning [1]. The reinforcement learning is one of unsupervised learning methods where the interaction between the robot and the environment is used instead of teacher signal. As one of the reinforcement learning, Q-Learning is widely employed for the acquisition of robotic actions [2]. In this method, the learning proceeds as an agent interacts with the environment like an evolutionary process of primitive creature. The most significant point in the unsupervised learning is that any preliminary knowledge is not required. Hence, its application to the real robot has a tremendous potential to produce very attractive robotic motions beyond our expectations; the supervised learning cannot generate such unexpected motion. Asada et al. attempted to propose its applications through the RoboCup [3]. Further, Kimura et al. demonstrated that the application of reinforcement learning enabled the robot to cause the advancement actions [4]. In our previous studies, we also have studied on the acquisition of various motions in mobile robots, such as a caterpillar-shaped and a starfish-shaped robots as shown in Fig. 1, by using the Q-Learning [5]. As the other trial, we also attempt the learning of the gait pattern with a gecko-shaped robot from a state of ignorance. These works examined the effect of the environmental variation and reward combination on the acquired motion forms.

As an extension of our previous studies, this study attempts the acquisition of dynamic motions by using the Q-Leaning method although the Q-Learning is generally unsuitable for learning dynamic tasks because Markov property is not necessarily guaranteed. Especially, this study focuses on a giant-swing motion by a compact humanoid robot, which has multi-degree of freedom. With regard to the robotic swing control, many researchers have reported on the Acrobot, a two-link robot with a single actuator. For example, Spong proposed the swing-up algorithm based on zero dynamics for Acrobot [6]. Michitsuji et al. discussed the control of a gymnast-like Acrobat-robot with three links on a horizontal bar [7]. As for the application of learning algorithm, Boone showed that the learning speed could improve by explicitly leaning the system equation [8]. Nishimura et al. achieved a swing-up control of a real Acrobot due to the switching rules of multiple controllers obtained by the reinforcement learning [9]. As a unique study, Fukuda and Hasegawa et al. attempted to realize the robotic bracheation by multi controllers based on learning. However, in these studies, the swing motion could be realized by using the robotic model or several

Fig. 1. Mobile robots for the Q-Learning in the previous studies.

Manuscript received March 1, 2009. This work was supported in part by the Grant-in-Aid for Scientific Research B2 No. 20300075.

M. Hara and H. Bleuler are with Ecole Polytechnique Fédérale de Lausanne, Lausanne, 1015 Switzerland (phone: +41 21 693 59 47, e-mail: masayuki.hara@epfl.ch, hannes.bleuler@epfl.ch).

N. Kawabe is with Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, 113-8656 Japan (e-mail: kawabe@ynl.t.u-tokyo.ac.jp).

N. Sakai and T. Yabuta are with Dept. of Mechanical Engineering, Yokohama National University, Yokohama, 240-8501 Japan (e-mail: sakai@yabsv.jks.ynu.ac.jp, yabuta@ynu.ac.jp).

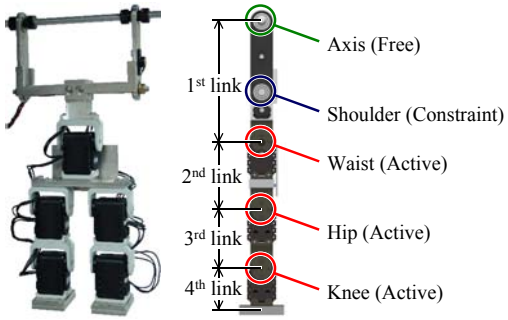
J. Huang is with Dept. of Intelligent Mechanical Engineering, School of Engineering, Kinki University, Higashi-Hiroshima, 739-2116 Japan (e-mail: huang@hiro.kindai.ac.jp).

controllers. Very few studies attempted to control the swing motion of the real robot with multi-degree of freedom without preliminary knowledge. Thus, in particular, we attempt the learning of the giant-swing motion by only the interaction with the environment, in which the preliminary knowledge is removed as far as possible. Further, to achieve the giant-swing motion is very difficult in comparison to previous studies because the employed robot has more degrees of freedom than the conventional Acrobot. This paper demonstrates very attractive giant-swing motions of a real robot generated by the Q-Learning. Further, we discuss how to give the reward to obtain the giant-swing motion.

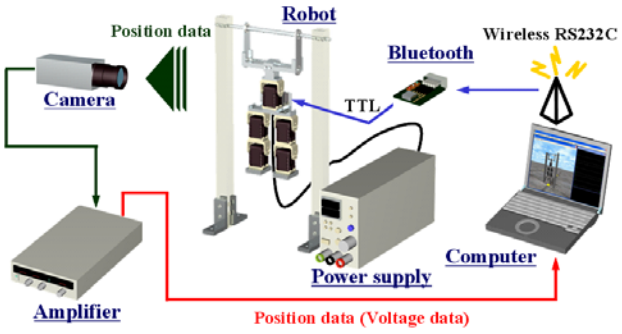
II. Experimental System

A. Giant-swing Robot

In this study, we fabricated and employed a compact humanoid robot with 5-DOF, as shown in Fig. 2; the main parameters are listed in Table I. The robot mainly comprises five actuators (model: Dynamixel AX-12+, ROBOTIS) and they are all enabled as the position control mode for achieving the giant-swing motion. These actuators are allocated on the robot so as to approximate human degree-of-freedom. There exist two bearings between the horizontal bar and robotic arm components, which enable the robot to freely swing around



(a) Giant-swing robot and its link allocation



(b) Experimental system

Fig. 2 Schematic diagram of experimental system.

TABLE I
ROBOTIC PARAMETERS

Size m	1st link	0.141
	2nd link	0.074
	3rd link	0.068
	4th link	0.046
Weight kg		0.801

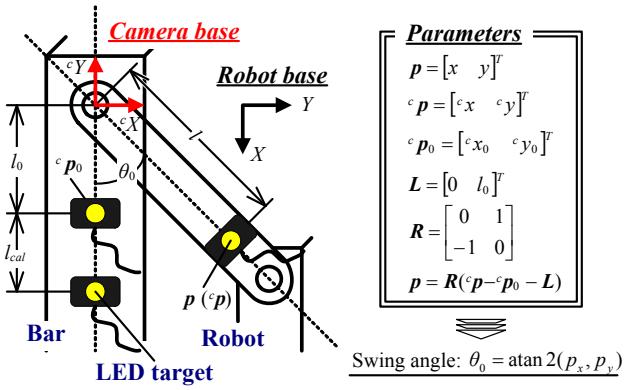


Fig. 3. Calculation method of the robotic swing angle.

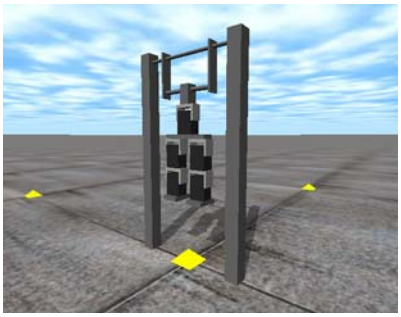


Fig. 4. Dynamic simulator of the giant-swing robot by ODE.

the bar. Similarly, the robotic shoulder comprises free joints, but in this study, two constraint components are attached to restrain the rotation at the shoulder in order to simplify the learning. On the back of the robot, a communication module with Bluetooth is attached as shown in Fig. 2 (b). By using this module, motor commands are transmitted via a radio communication based on RS-232C. As for the measurement system, a PSD-sensor system (model: C5949, Hamamatsu Photonics) is employed in order to obtain the robotic position. Using measured position data, the robotic swing angle can be calculated, as shown in Fig. 3. In this experimental system, the sampling rate is set to 250 ms in order to completely drive the motors to the desired position within a sampling step.

B. Dynamics Simulator

This study also created a dynamics simulator of the developed compact humanoid robot by using Open Dynamics Engine, a free physical-calculation simulator. Fig. 4 shows a three-dimensional graphics of the robot in the developed simulator. In this simulator, the robotic parameters, such as the motor characteristics and damping effect generated by the friction between the horizontal bar and arm components, are adjusted to those of the real robot as far as possible.

III. Reinforcement Learning

A. Q-Learning Algorithm

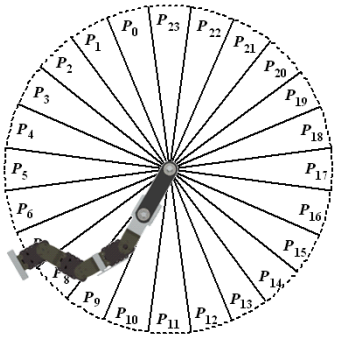
This study applied the Q-Learning algorithm, which is one of the reinforcement learning methods, to the giant-swing robot. The basic equation is given as follow:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_t + \gamma \max Q(s_{t+1}, a) - Q(s_t, a_t)\right] \quad (1)$$

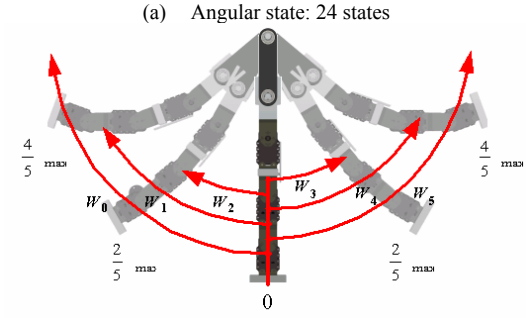
where Q is the action-value function. s_t and a_t are the state and action of the agent at the time t, respectively. α and γ denote the learning rate and discount rate; these two parameters determine the learning responsiveness and convergence, respectively. Finally, r_t indicates the reward that is allocated for all the actions in each state. In this study, an action in each state was selected by using ε-greedy method; in this method, the explorative action is selected with the probability ε ($0 \leq \varepsilon \leq 1$), whereas the agent acts greedily by the probability $1-\varepsilon$. In our learning, ε is simply decreased due to the learning step.

B. Application of the Q-Learning Algorithm

As shown in equation (1), it is necessary to define the state space of the robot. In this study, the robotic state based on the swing angle is employed, as shown in Fig. 4 (a). In addition, its angular velocity state, as shown in Fig. 4 (b), is also considered in order to include the dynamic effect. Finally, we define the state space with 144 states (24 states in the swing angle × 6 states in the angular velocity). In the learning process, the robotic transition state at each step is observed by using the PSD-sensor system. In the learning, the state s_{t+1} at the time t+1, which is obtained as the result of the action a_t, cannot be predicted at the time t because the robotic model is not available. Thus, this study uses a method with the delayed reward, where the action-value function $Q(s_t, a_t)$ is renewed at the time t+1. Also, the renewed action-value function fluctuates due to the reward even if the robot experiences the same transition; the reward includes small fluctuation because the robotic state space is roughly defined to avoid the explosion in the state number. Hence, this study uses the averaged action-value function over a series of renewals.



(a) Angular state: 24 states



(b) Angular velocity state: 6 states

Fig. 5. State space of the giant-swing robot.

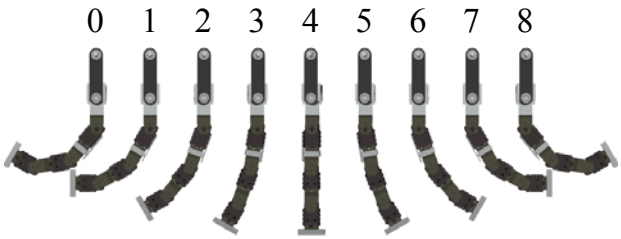


Fig. 6. Robotic action patterns in the preliminary experiment.

IV. Preliminary Experiment

A. Experimental Condition and Method

In order to verify whether the developed robot can achieve the giant-swing motion or not, a preliminary experiment was conducted by actually using the real robot; the dynamic simulator was not employed in this preliminary test. In this preliminary experiment, we allowed each enabled actuator to drive by 9 action patterns—0, ±10, ±20, ±30, and ±40 deg—as shown in Fig. 6. Further, the reward based on the decrease in the height of a LED target attached on the arm was allocated for all the actions in each state; we supposed that the decrease in the swing height includes the potential energy, which may be effective for enhancing the swing motion. As for the ε-greedy method, ε was decreased by 0.2 every 10000 learning steps from 1.0, i.e. it means that the frequency of greedy actions increases every 10000 learning steps. Hence, the leaning was finished when ε became 0.

B. Experimental Results and Discussion

Fig. 7 shows the transition in the swing angle for 60 s when using the action-value functions renewed in 20000 learning steps. In this graph, the swing angle over ±180 deg means that the robot could rotate around the horizontal bar. As shown in Fig. 7, it should be noted that the robot could not perform the giant-swing motion. Fig. 8 shows a distribution of the action-value functions obtained in this learning, where the action-value functions are categorized by five levels. The result shows that the action-value functions in the bottom of the horizontal bar are relatively low and have no remarkable difference in the neighborhood of state 72. In most cases, the robot was easy to fall into a motion loop with few action patterns in the early stage of the greedy action. This motion loop caused the stagnation of the swing at the bottom of the horizontal bar. The stagnant state prevented the robot from shifting to a new state for increasing the swing angle. When the robot accidentally shifted to the new state from the stagnant states, there were cases where the robot was able to achieve the giant-swing motion. Actually, the robot could jump out of the stagnant state if we applied small forced oscillation due to the external force in the early stage. Fig. 9 shows the transition in the swing angle when applying the random actions for 10 s instead of the forced oscillation. In

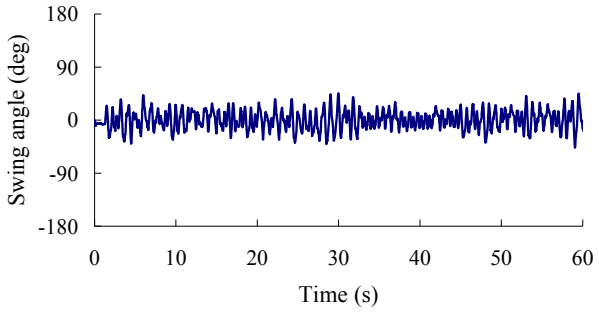


Fig. 7. Transition of the swing angle during the greedy actions.

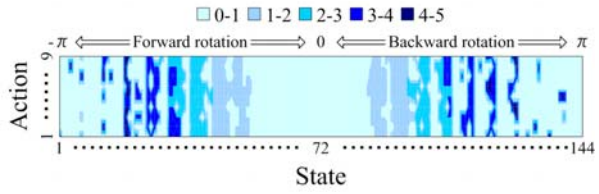


Fig. 8. Distribution of the action value functions.

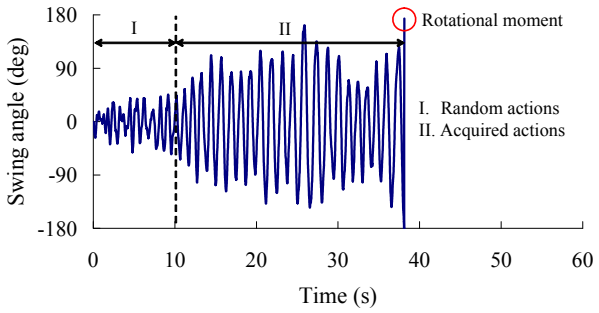


Fig. 9. Transition of the swing angle when applying the random actions for 10 s in the early stage.

this experiment, the robot could select the best actions based on the learning result, which was the same with that in Fig. 7, after the robot performed the random actions for 10 s. The results demonstrate that the robot could rotate around the horizontal bar with some repeatability, as shown in Fig. 9; in this method, the giant-swing motion was almost regularly performed between 30 s and 70 s. Fig. 10 graphically shows the highlight. These results imply that the reward based on the decrease in the swing height was not so effective for generating the giant-swing motion; how to give the reward may break the problem related to the repeatability. However, it could be confirmed that the developed giant-swing robot has the ability to perform the giant-swing motion.

V. Acquisition of Human-like Giant-swing Motion

A. Experimental Condition and Method

We attempted to avoid the stagnant state at the bottom of the horizontal bar by applying various types of rewards. In this experiment, the rewards based on the robotic physical quantities—decrease in swing height, swing angle, tip angle, and mechanical energy—as listed in Table II were given to the robotic actions in the Q-Learning process. The tip angle

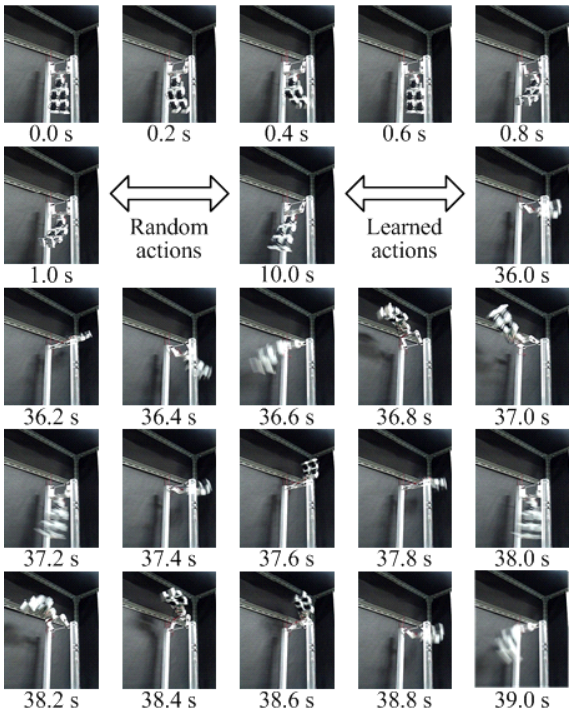


Fig. 10. Highlight of the acquired giant-swing motion.

TABLE II Reward Types in the Q-Learning Simulation

Swing-height type	$\Delta h(t) = l(\cos\theta_0(t) - \cos\theta_0(t-1))$
Swing-angle type	$\lvert\theta_0(t)\rvert$
Tip-angle type	$\lvert\theta_t(t)\rvert$
Energy type	$\Delta E(t) = \frac{1}{2} I \Delta\dot{\theta}_0(t) + mg\Delta h(t)$

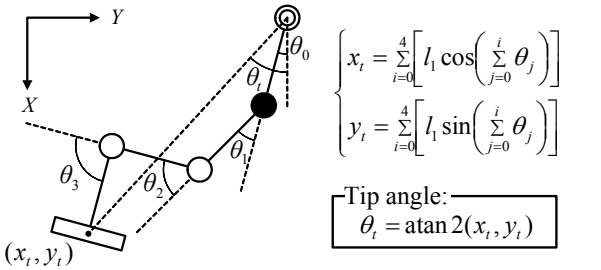


Fig. 11. Measurement of the tip position and angle of the giant-swing robot.

θ_t was calculated by means of the kinematic information of the robot, as shown in Fig. 11. In addition to this condition, the movable ranges of enabled motors were constrained as shown in Fig. 12 in order to imitate those of human beings; we expected that the robot may acquire a human-like motion. Similar to the experiment in chapter IV, the Q-Learning with ε-greedy method was applied. Then, ε was reduced with the time transition at the rate of 2.0×10^{-6} per learning step. First, we executed the Q-Learning for each reward by using the ODE-based dynamic simulator of the giant-swing robot in order to reduce the learning time and to avoid the fatigue breakdown. Subsequently, we attempted to pick up the

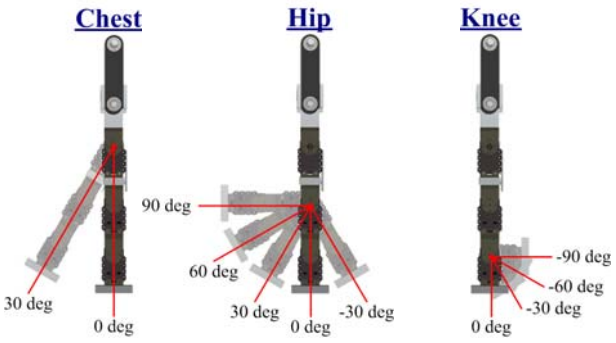
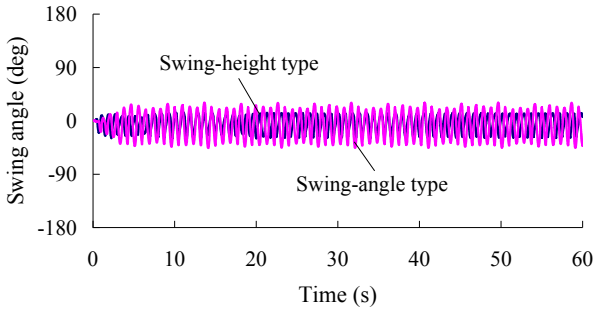
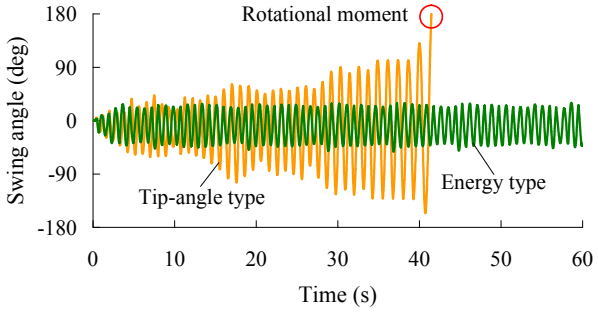


Fig. 12. Robotic action patterns.



(a) Reward: Swing angle and height



(b) Reward: Tip angle and mechanical energy

Fig. 13. Transition of the swing angle in the dynamic simulation.

effective learning results and actually implemented them in the real robot so as to examine the performance and applicability of the learning results obtained in the simulation.

B. Simulation Results and Its Application

Fig. 13 shows the results when the four types of rewards, as listed in Table II, were applied to the robot in the dynamic simulator. As shown in the blue line, the robot could not intensify the swing motion when applying the learning results for the swing-height-based reward. This result corresponds with the result in chapter IV. In addition, the robot could not also make the motion when the swing angle and the change in mechanical energy were given as the rewards, as shown in the red and green lines. On the other hand, as shown in the orange line, only the reward based on the tip angle enables the robot to achieve the giant-swing motion in the simulation. These results imply that the kinematic information of the robot might be significant to perform the giant-swing motion. With regard to the rewards except the tip-angle-based reward, they all use the swing angle of the arm; these rewards have only the information from the horizontal bar to the robotic arm. Hence, the robot was not able to know its own posture beyond the arms in the learning process. This condition might hinder jumping out of the stagnant action loop at the bottom of the horizontal bar. On the other hand, the tip-angle-based reward includes the information of robotic posture, as shown in Fig. 11. Thus, it is implied that the posture information enabled the comprehensive exploration in the learning process and might result in achieving the giant-swing motion.

As the next step, we practically implemented the learning result obtained by the application of the tip-angle-based reward in the real robot. The result demonstrated that the real robot could make a revolution around the horizontal bar. However, the robot frequently stopped the movement due to the motor spec; the employed motor has the characteristic that automatically stops the motor drive if the intensive actions are taken many times. In the most serious case, the cog in the motor chipped due to the intensive actions. This is because in the obtained learning result, the robot moves very vigorously at the bottom of the horizontal bar so as to jump out of the stagnant action loop; the motors were exposed to heavy loads around the limitation. Thus, the perfect repeatability could not be verified due to the marginal performance in the real robot. However, these results imply the possibility of the obtained learning result for achieving the giant-swing motion.

C. Manipulation of Learning Results

When the gymnasts start to perform the giant-swing motion, they usually bend their elbows and lift their body in order to promote the swing angle; they do not start the giant-swing motion while stretching out their arms like our robot. Hence, it may be necessary to change the learning strategy at the bottom of the horizontal bar to smoothly perform the giant-swing motion. In this study, we attempted to divide the giant-swing motion into two stages as shown in Fig. 14 and employed the individual learning results obtained in each stage. In the first stage, by using the dynamic simulator again, we had the robot learn the optimized angular frequency for releasing from the stagnant action loop by using a sinusoidal action pattern. According to the result, the best parameter obtained in this simulation was 6.2 rad/s; the robot could not perform the giant-swing motion even if the sinusoidal action with this parameter was applied over all the states. Once the robot jumped out of the stagnant state, the learning result in the first stage would not be used anymore. Instead, the learning results based on the tip-angle-based reward in the previous simulation were employed in the second stage.

Fig. 15 shows a transition in the swing angle when applying these two learning results to the real robot in each stage; Fig. 16 demonstrates the highlight. These results indicate that the real robot was able to rotate around the horizontal bar quickly. To confirm the repeatability, we also tried the giant-swing motion ten times with the same results. The robot could regularly rotate around the horizontal bar for 24.7 s on average. Hence, the reliability of learning results in section *B*

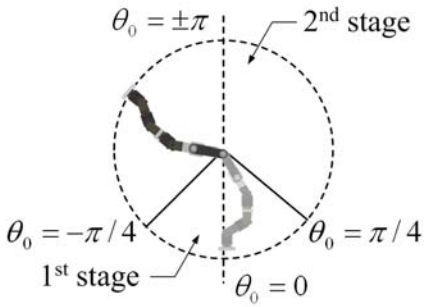


Fig. 14. Division of the learning stage..

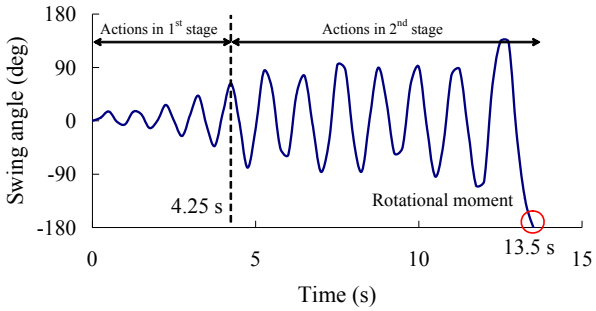


Fig. 15. Transition of the swing angle when applying the two learning results to the real robot.

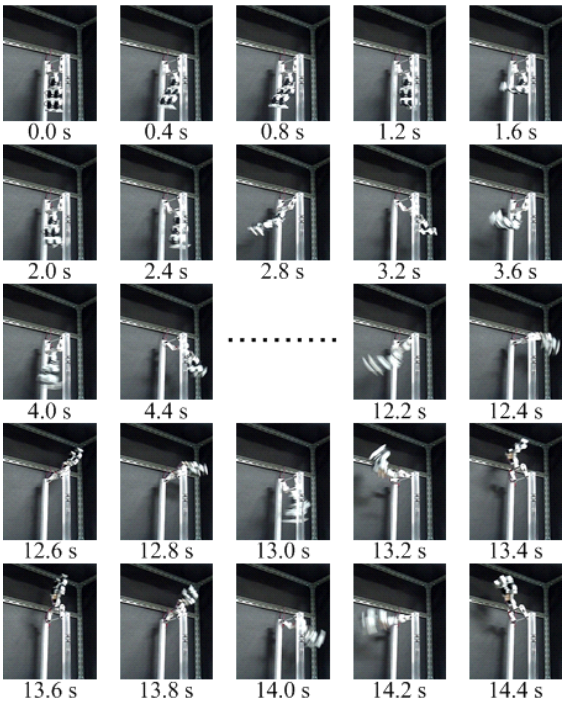


Fig. 16. Highlight of the giant-swing motion generated by the integrated learning results.

can be verified. These results imply that the action patterns for jumping out of the stagnant action loop in the early stage is required to enhance the swing motion and perform the giant-swing motion smoothly.

VI. Conclusion

In this study, we attempted to have a compact humanoid robot rotate around the horizontal bar from a state of ignorance by applying the Q-Learning method. Generally, the Q-Learning method is not suitable for learning dynamic motions. However, this study attempted to avoid this problem by considering the dynamic information in the robotic state space. First, we preliminarily confirmed that the developed robot had the ability to perform the giant-swing motion. After that, the effects of several rewards on acquiring the motion were examined in the dynamic simulator. The simulation results implied that the reward with the robotic kinematic information is effective for learning the giant-swing motion. Further, we practically applied an effective learning result to the real robot and examined the performance. The result showed the possibility of the learning result, but did not exhibit the repeatability because of the marginal performance in the motors. To solve this problem, the other learning result was applied in the early stage. The integration of two learning results enabled the robot to smoothly perform the giant-swing motion with the repeatability.

This paper implied that the Q-Learning has a possibility to enable the robot to acquire the dynamic task like the giant-swing motion by slightly giving knowledge in the early stage. In our future work, the robotic degree-of–freedom will be increased by removing the hampers at the shoulder and we try to examine how the robot with two free joints acquires the giant-swing motion in the Q-Learning.

References

[1] K. Doya, "Reinforcement learning in animals and robots," *International Workshop on Brainware*, 1996, pp. 69-71.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, A Bradford Book, MIT Press, 1998.

[3] M. Asada, T. Nakamura, and K. Hosoda, "Vision-Based Robot Reinforcement Learning for Purposive Behavior Acquisition," *Proc. of ICRA'95*, 1995, pp. 146-153.

[4] H. Kimura and S. Kobayashi, "Reinforcement learning using stochastic gradient algorithm and its application to robots," *IEE Japan Trans. on Electronics*, *Information and Systems*, 119-C, vol. 119, no. 8, 1999, pp. 931-934. (in Japanese)

[5] M. Hara, M. Inoue, H. Motoyama, J. Huang and T. Yabuta, "Study on Motion Forms of Mobile Robots Generated by Q-Learning Process Based on Reward Databases," *Proc. of SMC'06s, Man and Cybernetics*, 2006, pp. 5112-5117.

[6] M. W. Spong, "The swing up control problem for the Acrobot," *IEEE Control Magazine*, 1995, vol. 15, no. 1, pp. 49-55.

[7] Y. Michitsuji, H. Sato, and M. Yamakita, "Giant swing via forward upward circling of the Acrobat-robot," *Proc. of 2001 ACC*, 2001, pp. 3262-3267.

[8] G. Boone, "Efficient reinforcement learning: Model-based Acrobot control," *Proc. of ICRA'97*, 1997, pp. 229-234.

[9] M. Nishimura, J. Yoshimoto, Y. Tokita, Y, Nakamura, and S. Ishii, "Control of Real Acrobot by Learning the Switching Rule of Multiple Controllers," *Trans. of the IEICE. A*, vol. J88-A, no. 5, 2005, pp. 646-657. (in Japanese)

[10] T. Fukuda, and Y. Hasegawa, "Learning Method for Multiple-Control for Robot Behavior," *Trans. of the JSME. C*, vol. 63, no. 610, 1997, pp. 2043-2051. (in Japanese)

[11] Y. Hasegawa, Y. Ito, and T. Fukuda, "A Study of the Brachiation Type of Mobile Robot (7th Report, Behavior Learning for Hierarchical Behavior-based Controller)," *Trans. of the JSME. C*, vol. 67, no. 662, 2001, pp. 3204-3211. (in Japanese)