12

Characterization of Motion Forms of Mobile Robot Generated in Q-Learning Process

Masayuki HARA[1], Jian HUANG[2] and Testuro Yabuta[3]
[1]*École Polytechnique Fédérale de Lausanne (EPFL)*
[2]*Kinki University*
[3]*Yokohama National University*
[1]*Switzerland*
[2,3]*Japan*

1. Introduction

Acquisition of unique robotic motions by machine learning is a very attractive research theme in the field of robotics. So far, various learning algorithms—e.g., adaptive learning, neural network (NN) system, genetic algorithm (GA), etc.—have been proposed and applied to the robot to achieve a target. It depends on the persons, but the learning method can be classified roughly into supervised and unsupervised learning (Mitchell, 1997). In supervised learning, the ideal output for target task is available as a teacher signal, and the learning basically proceeds to produce a function that gives an optimal output to the input; the abovementioned learning methods belong to supervised learning. Thus, the learning results should be always within the scope of our expectation. While, the teacher signal is not specifically given in unsupervised learning. Since the designers do not need to know the optimal (or desired) solution, there is a possibility that unexpected solution can be found in the learning process. This article especially discusses the application of unsupervised learning to produce robotic motions.

One of the most typical unsupervised learning is reinforcement learning that is a evolutionary computation (Kaelbling et al., 1996; Sutton & Barto, 1998). The concept of this learning method originally comes from the behavioral psychology (Skinner, 1968). As seen in animal evolution, it is expecting that applying this learning method to the robot would have a tremendous potential to find unique robotic motions beyond our expectation. In fact, many reports related to the application of reinforcement learning can be found in the field of robotics (Mahadevan & Conell, 1992; Doya, 1996; Asada et al, 1996; Mataric, 1997; Kalmar et al., 1998; Kimura & Kobayashi, 1999; Kimura et al., 2001, Peters et al., 2003; Nishimura et al., 2005). For example, Doya has succeeded in the acquistion of robotic walking (Doya, 1996). Kimura et al. have demonstrated that reinforcement learning enables the effective advancement motions of mobile robots with several degrees of freedom (Kimura & Kobayashi, 1999; Kimura et al., 2001). As a unique challenge, Nishimura et al. achieved a swing-up control of a real Acrobot—a two-link robot with a single actuator between the links—due to the switching rules of multiple controllers obtained by reinforcement learning (Nishimura et al., 2005). Among these studies, Q-learning, which is a method of

reinforcement learning, is widely used to obtain robotic motions. Our previous studies have also introduced Q-learning to acquire the robotic motions, e.g., advancement motions of a caterpillar-shaped robot and a starfish-shaped robot (Yamashina et al., 2006; Motoyama et al., 2006), gymnast-like giant-swing motion of a humanoid robot (Hara et al., 2009), etc. However, most of the conventional studies have discussed the mathematical aspect such as the learning speed, the convergence of learning, etc. Very few studies have focused on the robotic evolution in the learning process or physical factor underlying the learned motions. The authors believe that to examine these factors is also challenging to reveal how the robots evolve their motions in the learning process.

This article discusses how the mobile robots can acquire optimal primitive motions through Q-learning (Hara et al., 2006; Jung et al., 2006). First, Q-learning is performed to acquire an advancement motion by using a caterpillar-shaped robot. Based on the learning results, motion forms consisting of a few actions, which appeared or disappeared in the learning process, are discussed in order to find the key factor (effective action) for performing the advancement motion. In addition to this, the environmental effect on the learning results is examined so as to reveal how the robot acquires the optimal motion form when the environment is changed. As the second step, the acquisition of a two-dimensional motion by Q-learning is attempted with a starfish-shaped robot. In the planar motion, not only translational motions in X and Y directions but also yawing motion should be included in the reward; in this case, the yawing angle have to be measured by some external sensor. However, this article proposes Q-learning with a simple reward manipulation, in which the yawing angle is included as a factor of translational motions. Through this challenge, the authors demonstrate the advantage of the proposed method and explore the possibility of simple reward manipulation to produce attractive planer motions.

2. Q-learning algorithm

Q-learning is one of reinforcement learning methods and widely used in the field of robotics. In Q-learning, an agent selects an action from all the possible actions in a state following some policy—a mapping of probability selecting action—and causes an interaction with an environment at a certain time. A reward based on the interaction and the target task is allocated to the selected action from the environment as a scalar value. At this time, the agent renews the database due to the given reward. Repeating this process, the action values are renewed and stored in each state. After the learning, an optimal motion for the desired task can be realized by just selecting the actions with the highest action value in each state. In Q-learning, the convergence to the optimal solution is promised as long as the series of learning process follows Markov Decision Process (MDP). The equation is simply expressed as follow:

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

where $Q(s_t, a_t)$ is action-value function when the agent selects an action a_t in a state s_t at time t. α and γ represent learning rate and discount rate, respectively. α ($0 < \alpha < 1$) dominates the learning responsiveness (speed); basically, a value near 1 is selected. On the other hand, γ is related to the convergence of learning. In general, Q-learning can be considered as a

learning method to maximize the expected value of reward that will be received in the future. However, the rewards which will be given from the environment in the future are basically unknown. So, the future rewards should be estimated by using the discount rate γ, as shown in equation (2):

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2}$$

Without the discount rate, the agent gets interested in only the immediate rewards and this causes the myopic action selection. Therefore, the introduction of the discount rate enables Q-learning with a long-term view. Ideally, an eigenvalue near 1 is used for the discount rate, but it is pointed out that the duration until the convergence of learning becomes longer if the value is too close to 1 (Schewartz, 1993; Mahadevan, 1996). Hence, the discount rate is set at 0.8 or 0.9 in this article.

In previous studies, several augmented Q-learning methods have been proposed and discussed in order to improve the learning performance (Konda et al., 2003; Mori et al., 2005; Peter & Shaal, 2008; Juang & Lu., 2009; Rucksties et al., 2010). For example, Mori et al. demonstrated that the application of Actor–Critic using a policy gradient method is effective to the learning of CPG-Actor-Critic model even if a high-order state space is configured (Mori et al., 2005). Peters and Schaal proposed Natural Actor-Critic expanding the idea of Actor–Critic using a policy gradient method (Peter & Shaal, 2008). However, in this article, the simplest Q-learning algorithm is applied to mobile robots in order to achieve robotic primitive motions with as minimum information as possible and to facilitate the discussion of how the robots acquire the primitive motion in such the condition.

3. Experimental system

3.1 Mobile robots

As mentioned above, this article introduces the acquisition of advancement and planar motions generated by Q-learning. In Q-learning for the advancement motion, a simple caterpillar-shaped robot is designed and employed. The caterpillar-shaped robot comprises four actuators (AI-Motor, Megarobotics Co., Ltd.) as shown in Fig. 1. In this robot, two actuators on both the tips are enabled; the others are completely fixed under the position control. On the other hand, a starfish-shaped robot, which has four enabled AI-Motors as shown in Fig. 2, is applied in Q-learning for acquiring the planar motion. In these robots, the motor commands are communicated between a computer and each AI-Motor via RS232C interface.

3.2 Experimental system

A schematic diagram of experimental system is shown in Fig. 3. In the experimental system, a function that provides rewards based on the robotic actions to these robots is required in order to renew the action-value function in each state. To perform Q-learning of the advancement and planar motions, it is necessary to measure the robotic travel distance in each leaning step by using some external sensor such as a motion capture system. In this article, a position sensitive detector (PSD) system (C5949, Hamamatsu Photonics) is

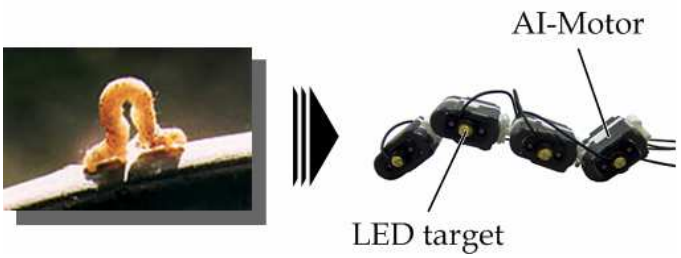


Fig. 1. Caterpillar-shaped robot for Q-learning of advancement motion: only the AI-Motors at both the sides are enabled

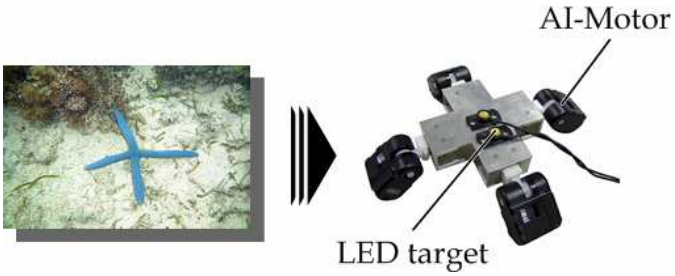


Fig. 2. Starfish-shaped robot for Q-learning of two-dimensional motions: all the AI-Motors (legs) are enabled

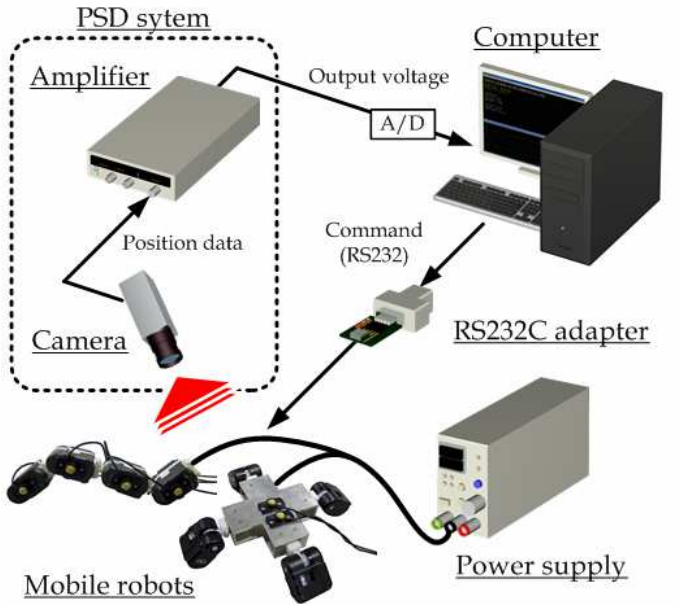


Fig. 3. A schematic diagram of experimental system: PSD system is used for motion-tracking of mobile robots

employed. The PSD system comprises a charge coupled device (CCD) camera, light-emitting diode (LED) targets, and an amplifier. In the PSD system, the CCD camera detects the LED targets which have individual IDs and the amplifier sends the two-dimensional position of each LED target to the computer as a voltage value; maximum 7 LED targets can be detected as analogue data at the same time. In Q-learning with the caterpillar-shaped robot, the CCD camera is fixed on the ground to enable tracking the LED targets attached on the side of the robot. Whereas, in Q-learning with the starfish-shaped robot, the CCD camera is attached on the ceiling to take a panoramic view of robotic two-dimensional motions. Two LED targets are attached on the top of the starfish-shaped robot; one is for measuring the robotic center position, and the other that is a bit shifted from the center of robot is for calculating the yawing angle.

3.3 Off-line Q-learning simulator based on reward database

In Q-learning, a considerable number of learning steps is required to reach an optimal solution. The long-term learning often causes the fatigue breakdown and the performance degradation in the real robot. In addition, the possibility that the mobile robots jump out of the motion-tracking-enabled area is quite high in the long-term learning; once the mobile robot gets out of the area, Q-learning has to be stopped immediately, and resumed after resetting the mobile robot within the motion-tracking-enabled area. So, the use of off-line learning is desired to facilitate Q-learning and to shorten the learning time. In general, robotic simulator is used instead of real robot to shorten the learning time. However, the robotic simulator has a technical issue related to the model error. The model error can be decreased by precisely configuring the robotic parameter in the simulator, but it causes the increase in the computational time (simulation time). Hence, this article proposes an off-line Q-learning simulator based on reward databases, which involving the real information of interaction between the robot and the environment. Here, the concept of reward-database-based Q-learning is introduced.

A flow chart of off-line Q-learning simulator is shown in Fig. 4. First, as for the reward database, ID numbers are assigned to all the action patterns and all the state transitions are performed among all the IDs several times by actually using robots. In parallel, some physical quantity related to the target motion, such as a travel distance and a yawing angle, all over the transition states are measured several times. The measured physical quantities are averaged by the number of times that the robot took the same state transition, and the averaged values are stored into a database as a representative reward data; the reward data is normalized at this time. In Q-learning with the real robots, the interaction between the robot and the environment must be simulated to allocate a reward to a selected action in each state to renew the action-value function. However, in the proposed method, once the reward database is established, the robot is not needed anymore because the reward database includes all the real interactions and related physical quantities. Hence, the proposed method can omit the computation of the interaction. In Q-learning with the reward database, a reward is just referred from the database depending on the state transition, and uses the selected reward to renew the action-value function. This is an advantage of the proposed method for the conventional methods with the robotic simulator although the preliminary experiment is needed; the conventional methods require the computation of the interaction every learning step.

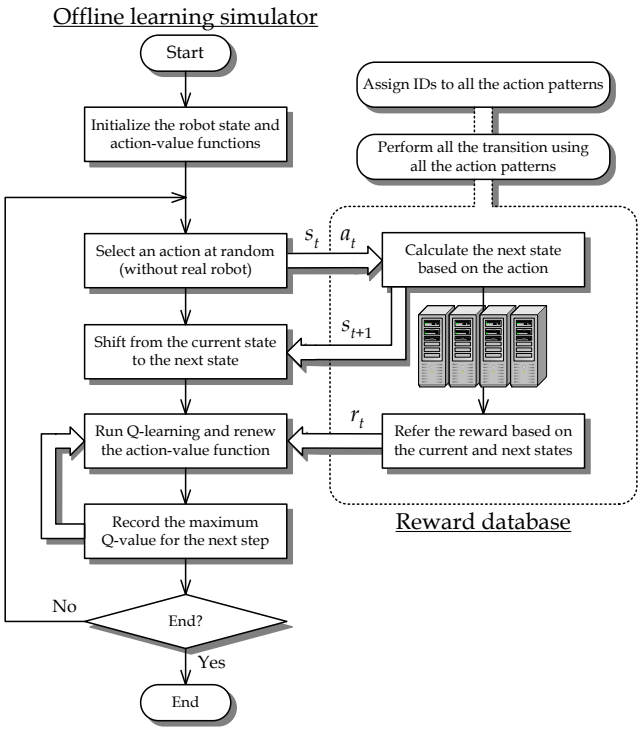


Fig. 4. A flow chart of off-line Q-learning using the reward database

4. Analysis of robotic advancement motions generated in Q-learning

4.1 Acquisition of advancement motion with the caterpillar-shaped robot

Q-learning is applied to acquire advancement motion of caterpillar-shaped robot. In this Q-learning travel distance, which is the representative data averaged by 10000 step-actions, from a state to the next state is given as reward. The action patterns of the caterpillar-shaped robot are shown in Fig. 5; the two-enabled motors at both the sides are controlled at 5 positions (0, ±26, and ±52 deg), respectively. The caterpillar-shaped robot randomly selects one of 25 actions in a learning step—random action policy. Totally, 625 ($5^2 \times 5^2$) state transitions can be selected in the learning. The learning rate and discount rate are configured at 0.9 and 0.8, respectively. Under these experimental conditions, Q-learning is performed in the proposed off-line simulator.

Fig. 6 shows the transitions of travel distances per a leaning step when only the highest Q-values are selected, i.e., when the caterpillar-shaped robot takes the greedy action; the blue, red, and green lines—in this experiment, Q-learning was performed three times—respectively indicate the relationships between the learning steps and the averaged distance traveled by the caterpillar-shaped robot in a step. From Fig. 6, it should be noted that the three trials finally reach the same performance (about 4.3 mm travel distance in a step) with the similar profile. This result also implies the good learning convergence and repeatability; all Q-learning are converged at around the 5000 learning steps in this case.

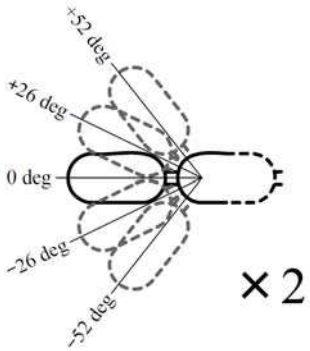


Fig. 5. Action patterns of caterpillar-shaped robot: 5^2 action patterns in each side

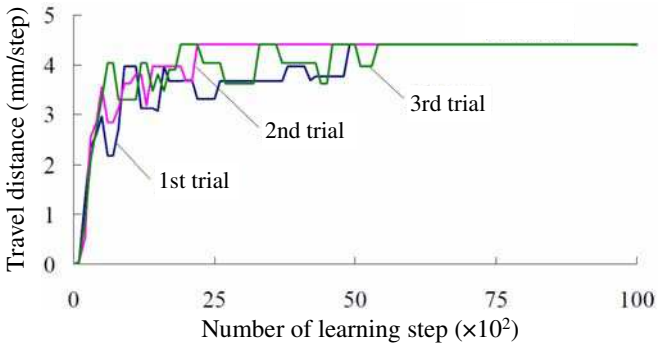


Fig. 6. Relationships between the number of learning step and averaged distance traveled by the caterpillar-shaped robot in a step under Q-learning conditions a = 0.9 and γ = 0.8

4.2 Transition of motion forms during Q-learning

When the caterpillar-shaped robot takes the greedy actions after Q-learning, a series of robotic actions defined by the ID numbers appear as an optimal motion. This article defines this series of robotic actions as "motion form". The motion forms consisting of a loop of a few actions appear with different patterns during Q-learning. Here, the transition of motion forms is analyzed to reveal how the caterpillar-shaped robot acquires an optimal advancement motion through the interaction with the environment. To achieve this, the motion forms are observed by extracting the learning results every 100 step until 5000 steps. Through the observation, it is found that four representative motion forms, as shown in Fig. 7, appear and disappear until the caterpillar-shaped robot reaches an optimal motion form. The number written over the robotic figure is the ID number that is allocated to the states in the database. Each motion form comprises a few actions and these actions are periodically repeated in the advancement motion. Note that these motion forms except the optimal motion form are not necessarily observed at the same timing when performing Q-learning several times because the random action policy is applied to the robot in this experiment; different environments and different learning parameters would cause other motion forms. However, since these four motion forms are frequently taken in Q-learning, this article discusses the transition of these motion forms.

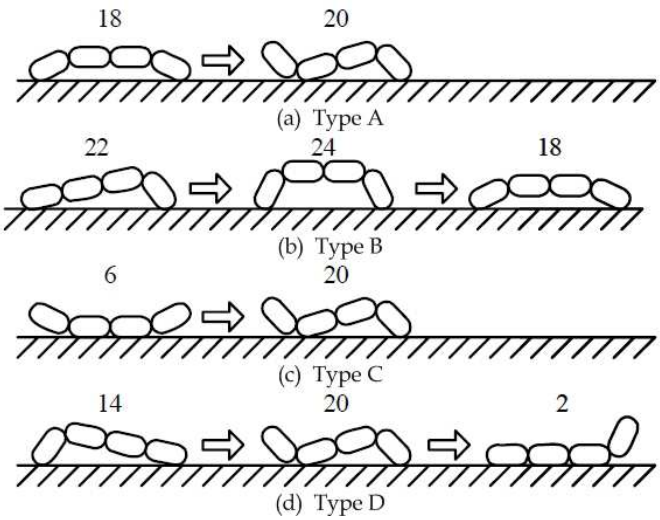


Fig. 7. Representative motion forms frequently observed in Q-learning under the conditions of a = 0.9 and γ = 0.8: Type D is the optimal motion form which brings the most effective advancement motion to the caterpillar-shaped robot

Fig. 8 shows a representative transition of motion forms until the caterpillar-shaped robot acquires an optimal motion form. In this learning, the optimal motion forms for the advancement motion is Type D consisting of three states (ID: 2, 14, and 20). In the early process of Q-learning, Type A, B, and C repeatedly appear and disappear until 800 steps; sometimes these motion forms are combined each other. From 800 steps to 4800 steps, major change in the motion form is not observed as shown in Fig. 8. In this phase, the states 14 and 20 are almost fixed and the subsequent state was changed variously. Through the several transitions, finally, the motion form is converged to the optimal motion form—Type D. Focusing on the transition, Q-learning might be divided into two phases based on 800 steps— early and later learning phases. In the early stage, the caterpillar-shaped robot attempts to establish some rough frameworks of motion forms for effectively performing the advancement motion. On the other hand, it seems that the robot selects a possible candidate from several key motion forms and performs the fine adjustment in the later phase. This implies the evolutionary feature of Q-learning.

Here is the discussion about the transition of motion forms. In general, the rewards possess the following relationships:

$$r_1 > r_2 > r_3 > \cdots > r_n > r_{n-1} \tag{3}$$

In a finite state, Q-learning is considered as a problem that finds out a group of actions that maximize the expected value of discount return R_t. Ideally, only the actions that have the highest Q-value should be selected in each state to maximize the expected value of R_t. However, the robot cannot take only the actions with the highest Q-value because of the finite space. So, the robot also has to select the actions with lower Q-value in some states to maximize R_t. Under this constraint, the robot attempts to find out a group of actions—motion form—with a maximum expected value of R_t. This is a big feature of unsupervised

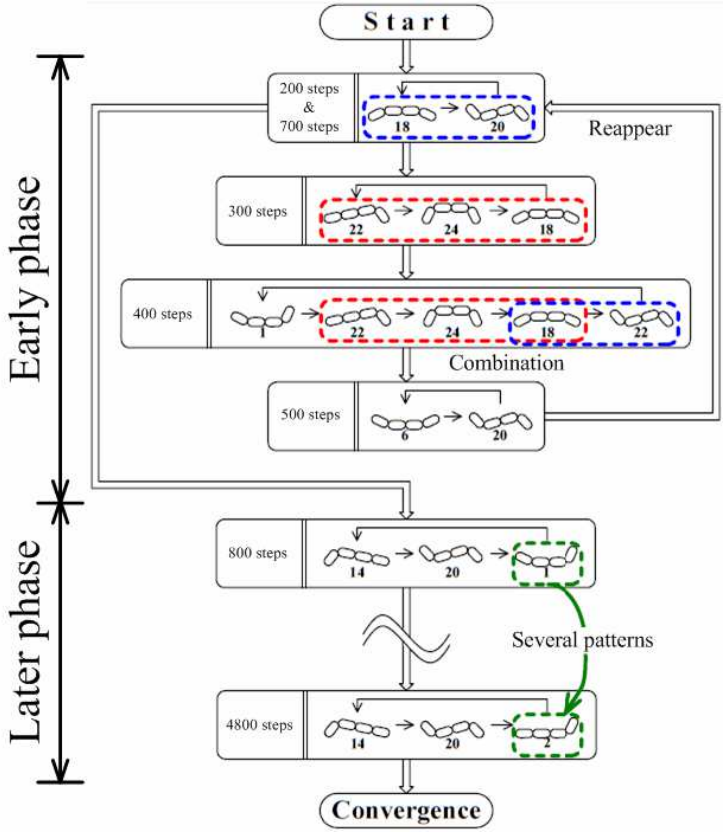


Fig. 8. A representative transition of motion forms during Q-learning process

learning; such the process cannot be found in supervised learning. Here, from equations (1) and (2), it should be noted that the discount rate γ significantly affects the transition of motion forms. In fact, our previous studies demonstrated that it is easier to produce the optimal motion form with a very few actions when γ is configured at a large value; vice versa, an inverse result is observed when γ is a smaller value (Yamashina et al., 2006; Motoyama et al., 2006). Hence, it is assumed that the discount rate is a significant factor to generate the series of motion forms in the learning process.

4.3 Environmental effect on the optimal motion form

As the next step, the environmental effect on the optimal motion form is investigated to know how Q-learning adapts to the environmental change. It is expected that Q-learning is performed involving the environmental change in the interaction and generates a motion form optimized to the given environment. In this article, ascending and descending tasks are tried by changing the inclination of floor, as shown in Fig. 9. The inclination is adjusted at ±5 deg in each task. A carpet made of the same fiber, which is used on the flat floor in the experiment of section 4.1, is attached on the slope so as to make the friction property between the caterpillar-shaped robot and the slope the same. Under this condition,

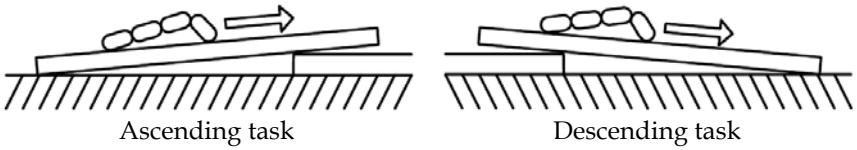


Fig. 9. Environmental changes in Q-learning: ascending and descending ±5 deg slopes

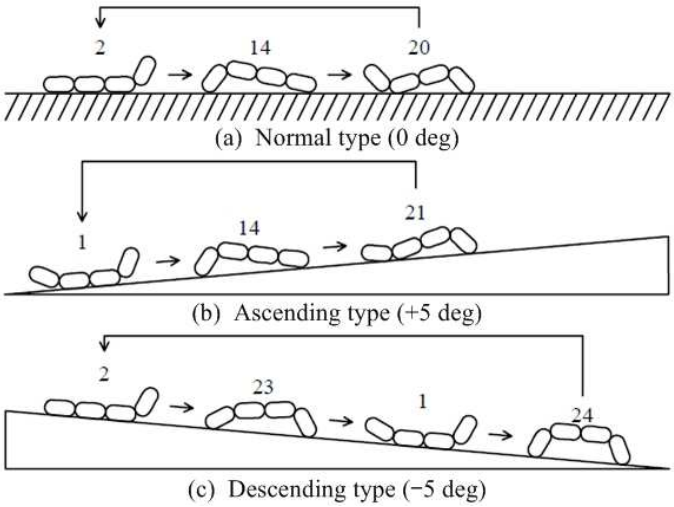
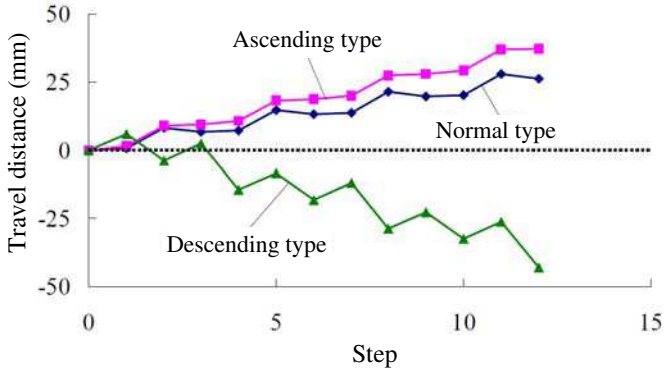


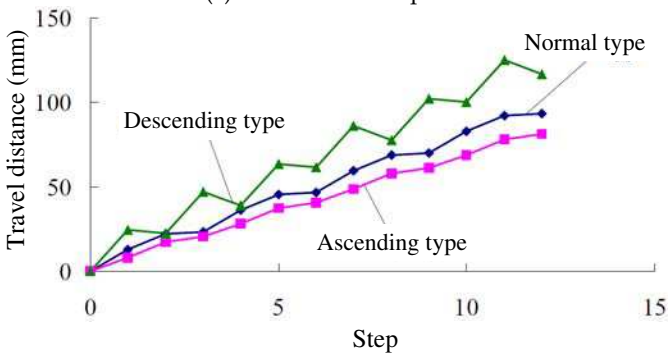
Fig. 10. Optimal motion forms in the three environments

Q-learning with the same learning parameters in section 4.1 was performed. Fig. 10 compares the optimal motion forms generated in the three environments—flat floor, uphill, and downhill. Here, let's define these optimal motion forms as normal-type, ascending-type, and descending-type motion forms, respectively. As expected, the caterpillar-shaped robot acquires different optimal motin forms. This implies that the caterpillar-shaped robot has learned the effective advancement motion in the individual environments.

The performance of each motion form is examined by comparing the travel distance in each result. The cumulative travel distances over 20 steps are shown in Fig. 11. Figs. 11 (a) and (b) show the results on the uphill and the downhill, respectively. In the case of uphill, the caterpillar-shaped robot could advance when applying the normal-type and ascending-type motion forms, whereas the robot slipped on the slope toward the opposite direction during the descending-type motion form; in this case, the ascending-type motion form demonstrated the best performance. Here, these optimal motion forms are analyzed to reveal the key factor for ascending the uphill. In the ascending task, it is considered that generating the force against the gravity and keeping the friction force not to slip on the slope would be very important. Focusing on the normal-type and ascending-type motion forms, it should be noted that the rear part pushes the caterpillar-shaped robot when the state is shifted from 14 to the next state—20 in the normal-type motion form and 21 in the ascending-type motion form. Such the state transition cannot be found during the descending-type motion form. As for the advancement motion on the uphill, this pushing action might be needed to produce the propulsion in the caterpillar-shaped robot. In adition,



(a) Environment: Uphill



(b) Environment: Downhill

Fig. 11. Performance of optimal motion forms in several environmental conditions

the comparison of the normal-type and ascending-type motion forms tells us that the contact area between the robot and the slope is a bit larger in the ascending-type motion form after the pushing action. So, it results in larger friction force during the ascendig-type motion form and it would enable the caterpillar-shaped robot to advacne without slipping on the slope. This difference might produce the different performances of advancement motion in the normal-type and ascending-type motion forms, as shown in the blue and red lines in Fig. 11 (a). Hence, these results imply that the pushing action and large contact area after the pushing action are necessary to make the robot effectively ascend the slope. On the other hand, in the case of downhill, the robot can take advantage of slip to make a large step. In this case, it is considered that the dynamic motion and less friction force would be effective to descend the slope. The descending-type motion form shows the best performance among the three types as expected. In this motion form, the shape like a bridge is formed (23 and 24) and it is broken at the next state (1 and 2); this series of actions could be considered as a jumping. This jumping-like motion could produce the dynamic advancement motion with less friction and lead to a good performance, as shown in Fig. 11 (b).

Thus, this section demonstrated that Q-learning could find out the most optimal motion form that is peculiar to the environment. In addition, the analysis of the motion forms implies that the learned motion form is reasonable from a viewpoint of robotic kinematics.

5. Acquisition of robotic planar motion by Q-learning

5.1 Advancement motions of the starfish-shaped robot in *X* and *Y* directions

The starfish-shaped robot can basically select two actions in the horizontal and vertical directions (*X* and *Y* directions). Here, the performances of advancement motions on the flat floor in the two directions are introduced. Regarding Q-learning, four enabled motors are controlled at 3 positions (0 and ±52 deg), as shown in Fig. 12. Similar to Q-learning in the caterpillar-shaped robot, the random action policy is taken; in a learning step, the starfish-shaped robot randomly selects one of 81 actions. Totally, 6561 ($3^4 \times 3^4$) state transitions are selectable. Under the conditions $\alpha = 0.9$ and $\gamma = 0.9$, Q-learning, whose reward databases are based on the travel distances averaged by 10000 step-actions in each direction, is performed in the proposed off-line simulator. Fig. 13 shows the optimal motion form in the *X* direction; in the *Y* direction, the optimal motion form becomes the same as that in the *X* direction that rotated by 90 deg. The transitions of travel distances in a learning step and the robotic trajectories within 20 steps are shown in Figs. 14 and 15, respectively.

As shown in Fig. 14, the performances in both the directions are almost the same. Here, the most noteworthy point is the magnitude of distance traveled in one step. The travel distance by the starfish-shaped robot (about 9.0 mm) is twice as long as that of caterpillar-shaped robot (about 4.3 mm) although each motor takes only the three positions.

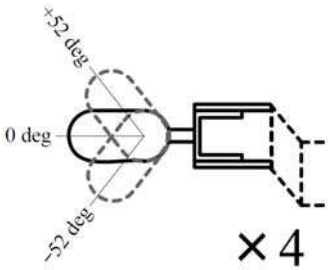


Fig. 12. Action patterns of starfish-shaped robot: 3^4 action patterns in each leg

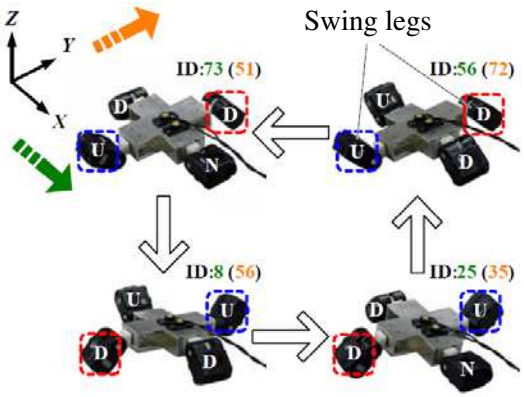


Fig. 13. Optimal motion form for the advancement motion in the *X* direction

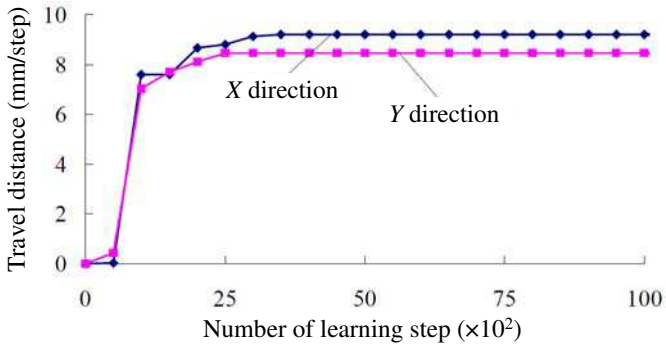


Fig. 14. Relationships between the number of learning step and averaged distance traveled by the starfish-shaped robot per a step under Q-learning conditions a = 0.9 and γ = 0.9

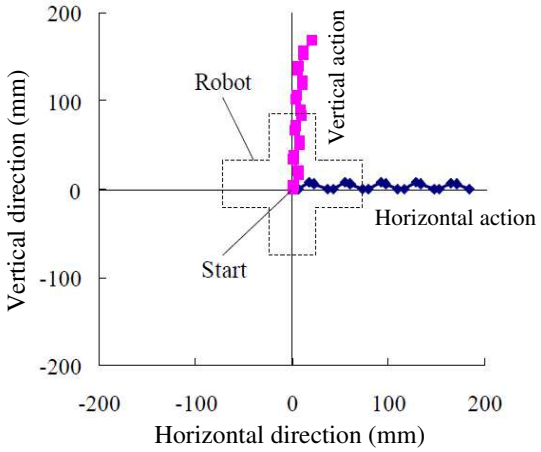


Fig. 15. Trajectories of the starfish-shaped robot in the horizontal and vertical directions

This difference would come from the availability of side legs (right and left motors); without the side legs, it can be considered that Q-learning of the starfish-shaped robot is almost the same as that of the caterpillar-shaped robot. Focusing on the motion form, it should be noted that the front and rear motors are driven for the advancement motion, whereas the right and left motors are used for helping the advancement motion. That is, it is thought that the motions of side legs prevent the starfish-shaped robot from slipping on the flat floor or moving backward. In fact, if the optimal motion form is performed without side legs, the travel distance in one step becomes significantly short. Therefore, these results imply that the starfish-shaped robot skillfully employed the advantage of the swing legs.

5.2 Planar motion by a reward manipulation

To achieve planar motion, the rewards should be configured at least including the horizontal and vertical positions and yawing angle of the starfish-shaped robot. The use of these parameters would make Q-learning complicated and it is not intuitive anymore. In

this article, the possibility of producing a planar motion by a reward manipulation is discussed. Concretely, the advancement motion in an oblique direction is challenged by simply manipulating the normalized reward databases for the horizontal and vertical motions, i.e., $\boldsymbol{r}_x$ and $\boldsymbol{r}_y$ obtained in the experiment of section 5.1. This challenge can be realized only in Q-learning based on the reward database; Q-learning with the robotic simulator cannot allow this. In the reward manipulation, the following equation is employed to make a new reward database $\boldsymbol{r}_{\text{new}}$:

$$\boldsymbol{r}_{\text{new}} = w\boldsymbol{r}_x \pm \text{sgn}(w)(1-|w|)\cdot \boldsymbol{r}_y \tag{4}$$

where w ($-1 \le w \le 1$) is a weight parameter that determines the priority of the two rewards. sgn(w) represents the sign of the weight parameter. In this experiment, w is set at ±0.5 in order to achieve the advancement motions in the directions of 45 deg and 225 deg with respect to the horizontal direction. Based on $\boldsymbol{r}_{\text{new}}$, Q-learning is performed in each condition by means of the proposed off-line simulator. Fig. 16 shows the trajectories of the starfish-shaped robot traveled within 20 steps. The results show that the starfish-shaped robot could approximately advance in the directions that the authors aimed at although the directions were not able to be completely corresponding to the requested directions. Also, this demonstrates the possibility of the proposed reward manipulation to generate several planar motions. In general, the acquired Q-values should be completely renewed due to the coherence of the rewards when the agent learns new tasks, i.e., the agent cannot acquire multiple actions at a time due to the oblivion of the knowledge. Therefore, the proposed method might bring a breakthrough in generating multiple and novel motions through Q-learning.

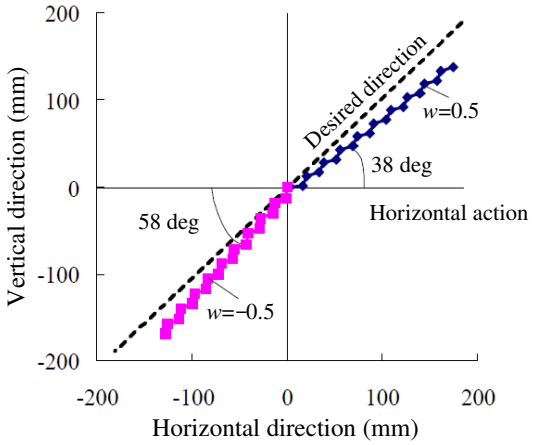


Fig. 16. Trajectories of starfish-shaped robot in the direction of 45 deg and 225 deg after Q-learning based on new reward databases manipulated by $\boldsymbol{r}_x$ and $\boldsymbol{r}_y$

6. Conclusion

In this article, the authors have focused on the key factors of robotic motions generated in Q-learning process. First, an off-line learning simulator based on the reward databases was proposed to facilitate Q-learning. Then, Q-learning was performed in the caterpillar-shaped

robot to generate the advancement motion. The observation of learning process implied that some key motion forms appear or disappear in the early learning phase and Q-learning adjusts them to an optimal motion form in the later learning phase. In addition, the effect of environmental changes on the optimal motion form was discussed by using an uphill condition and a downhill condition. Even if the environment was changed, Q-learning resulted in the motion forms which are optimized for the individual environment. As the next step, the planar motion by the starfish-shaped robot was tried. The results in the horizontal and vertical actions demonstrated that the starfish-shaped robot skillfully used their advantage (side legs) to enable longer travel distance. In addition, a reward manipulation with multiple reward databases was proposed to produce the planar motions. The application implied that there is potential to yield unique robotic motions.

7. Future works

This article discussed the motion forms yielded during Q-learning by using a caterpillar-shaped robot and a starfish-shaped robot. In our future work, the authors will examine the acquisition process of a gymnast-like giant-swing motion by a compact humanoid robot and explore the key factor. Through these attempts, the authors aim at having a much better understanding of evolutionary aspect of Q-learning.

8. References

Asada, M.; Noda, S.; Tawaratsumida, S. & Hosoda, K. (1996). Purposive behaviour acquisition for a real robot by vision-based reinforcement learning. *Machine learning*, Vol. 23, pp. 279-303

Doya, K. (1996). Reinforcement learning in animals and robots. *Proceedings of International Workshop on Brainware*, pp. 69-71,

Hara, M.; Inoue, M.; Motoyama, H.; Huang, J. & Yabuta, T. (2006). Study on Motion Forms of Mobile Robots Generated by Q-Learning Process Based on Reward Database. *Proceedings of 2006 IEEE International Conference on Systems, Man and Cybernetics*, pp. 5112-5117

Hara, M.; Kawabe, N.; Sakai, N.; Huang, J. & Yabuta, T. (2009). Consideration on Robotic Giant-Swing Motion Generated by Reinforcement Learning," *Proceedings of 2009 IEEE International Conference on Intelligent Robots and Systems*, pp. 4206–4211

Juang. C. & Lu, C. (2009). Ant colony optimization incorporated with fuzzy Q-learning for reinforcement fuzzy control. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, Vol. 39, No. 3, pp. 597–608

Jung, Y.; Inoue, M.; Hara, M.; Huang, J. & Yabuta, T. (2006). Study on Motion Forms of a Two-dimensional Mobile Robot by Using Reinforcement Learning. *Proceedings of SICE-ICCAS International Joint Conference 2006*, pp. 4240-4245

Kaelbling, L. P.; Littman, M. L. & Moore, A. W. (1996). Reinforcement learning; A survey. *Journal of Artificial Intelligence Research*, Vol. 4, pp. 237-285,

Kalmar, Z.; Szepesvari, C & Lorincz, A. (1998). Module-Based Reinforcement Learning: Experiments with a Real Robot. *Autonomous Robots*, Vol. 5, pp. 273–295,

Kimura, H. & Kobayashi, S. (1999). Reinforcement learning using stochastic gradient algorithm and its application to robots. *IEE Japan Transaction on Electronics, Information and Systems*, Vol. 119-C, No. 8, pp. 931–934 (in Japanese).

Kimura, H.; Yamashita, T. & Kobayashi, S. (2001). Reinforcement Learning of Walking Behavior for a Four-Legged Robot. *Proceedings of CDC2001*, pp. 411–416

Konda, V. R. & Tsitsiklis, J. N. (2003). On Actor-Critic Algorithms. *Society for Industrial and Applied Mathematics*, Vol. 42, No. 4, pp. 1143–1166

Mahadevan, S. & Connell, J. (1992). Automatic programming of behaviour-based robots using reinforcement learning. *Artificial Intelligence*, Vol. 55, pp. 311-365

Mahadevan, S. (1996). Average Reward Reinforcement Learning: Foundations, Algorithms, and Empirical Results. *Machine Learning*, Vol. 22, pp. 159-196.

Mataric, M. J. (1997). Reinforcement learning in the multi-robot domain. *Autonomous Robots*, Vol. 4, pp. 73-83

Mitchell, T. (1997). Machine Learning, MacGraw-Hill Science

Mori, T; Nakamura, Y. & Ishii, S. (2005). Reinforcement Learning Based on a Policy Gradient Method for a Biped Locomotion. *Transactions of the IEICE*, Vol. J88-D-II, No. 6, pp. 1080–1089

Motoyama, H.; Yamashina, R.; Hara, M.; Huang, J. & Yabuta, T. (2006). Study on Obtained Advance Motion Forms of a Caterpillar Robot by using Reinforcement Learning. *Transaction of JSME*, Vol. 72, No. 723, pp. 3525-3532 (in Japanese)

Nishimura, M.; Yoshimoto, J.; Tokita, Y.; Nakamura, Y. & Ishii, S. (2005). Control of Real Acrobot by Learning the Switching Rule of Multiple Controllers. Transactions of the IEICE, Vol. J88-A, No. 5, pp. 646-657 (in Japanese)

Peters, J.; Vijayakumar, S. & Shcaal, S. (2003). Reinforcement Learning for Humanoid Robotics. *Proceedings of Humanoids 2003*

Peters. J. & Schaal, S. (2008). Natural Actor-Critic. *Neurocomputing*, Vol. 71, pp. 1180–1190

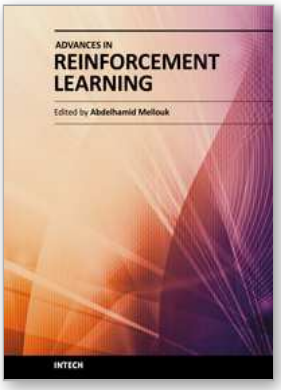
Rucksties, T.; Sehnke, F.; Schaul, T.; Wierstra, J.; Sun, Y. & Schmidhuber, J. (2010). Exploring Parameter Space in Reinforcement Learning. *Journal of Behavioral Robotics*, Vol.1, No. 1, pp. 14–24

Schwartz, A. (1993). A Reinforcement Learning Method for Maximizing Undiscounted Rewards. *Proceedings of the 10th International Conference, on Machine Learning*, pp. 298-305

Skinner, B. F. (1968). The technology of teaching. Prentice Hall College Div

Sutton, R. S. & Barto, A. G. (1998). Reinforcement Learning: An Introduction. MIT Press

Yamashina, R.; Maehara, S.; Urakawa, M.; Huang, J. & Yabuta, T. (2006). Advance Motion Acquisition of a Real Robot by Q-learning Using Reward Change. *Transaction of JSME*, Vol. 72, No. 717, pp. 1574-1581 (in Japanese)



Advances in Reinforcement Learning

Edited by Prof. Abdelhamid Mellouk

ISBN 978-953-307-369-9
Hard cover, 470 pages
Publisher InTech
Published online 14, January, 2011
Published in print edition January, 2011

Reinforcement Learning (RL) is a very dynamic area in terms of theory and application. This book brings together many different aspects of the current research on several fields associated to RL which has been growing rapidly, producing a wide variety of learning algorithms for different applications. Based on 24 Chapters, it covers a very broad variety of topics in RL and their application in autonomous systems. A set of chapters in this book provide a general overview of RL while other chapters focus mostly on the applications of RL paradigms: Game Theory, Multi-Agent Theory, Robotic, Networking Technologies, Vehicular Navigation, Medicine and Industrial Logistic.

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