Artif Life Robotics (2016) 21:11–17
DOI 10.1007/s10015-015-0258-1

ORIGINAL ARTICLE

Reinforcement learning in dynamic environment: abstraction of state-action space utilizing properties of the robot body and environment

Kazuyuki Ito[1] · **Yutaka Takeuchi**[2]

Received: 2 December 2014 / Accepted: 24 November 2015 / Published online: 18 January 2016

Abstract In this paper, we address the autonomous control of a 3D snake-like robot through the use of reinforcement learning, and we apply it in a dynamic environment. In general, snake-like robots have high mobility that is realized by many degrees of freedom, and they can move over dynamically shifting environments such as rubble. However, this freedom and flexibility leads to a state explosion problem, and the complexity of the dynamic environment leads to incomplete learning by the robot. To solve these problems, we focus on the properties of the actual operating environment and the dynamics of a mechanical body. We design the body of the robot so that it can abstract small, but necessary state-action space by utilizing these properties, and we make it possible to apply reinforcement learning. To demonstrate the effectiveness of the proposed snake-like robot, we conduct experiments; from the experimental results we conclude that learning is completed within a reasonable time, and that effective behaviors for the robot to adapt itself to an unknown 3D dynamic environment were realized.

Keywords Dynamic environment · Reinforcement learning · Crawler robot · Snake-like robot

1 Introduction

Recently, robots that have many degrees of freedom have drawn considerable attention. Snake-like robots or robots of serially connected crawlers are typical examples. In general, these robots have high mobility that is realized due to the many degrees of freedom while in motion. So, they can be operated in complex and quick-shifting environments like rubble, and they are expected to be used for practical applications such as rescue operations [1–8].

On the other hand, the application of reinforcement learning for controlling real robots also attracts considerable attention because robots can learn effective behavior by trial-and-error [9–11]. Therefore, by applying reinforcement learning to a robot with many degrees of freedom, an autonomous robot with high mobility can be realized; such a robot will be very useful for various tasks such as rescue operations. However, conventional reinforcement learning has a serious problem when applied to robots with many degrees of freedom that operate in complex environments. These are known as the state explosion problem and the lack of generalization ability. The size of the state-action space increases exponentially with an increase in the robot's degrees of freedom as well as the complexity of the environment. As a result of this increase, learning cannot be completed within a reasonable time limit [10]. In addition, if some part of environment is changed, additional learning is required. So, it is highly difficult to apply reinforcement learning to robots operating in a dynamic environment.

In contrast, animals can learn by trial-and-error in spite of the many degrees of freedom afforded by their bodies and despite the fact that the daily environment that they live in, and contend with, is very complex and dynamic. How animals are able to cope with such a challenging environment is still an open question. But in embodied cognitive science

This work was presented in part at the 19th International Symposium on Artificial Life and Robotics, Beppu, Oita, January 22–24, 2014.

✉ Kazuyuki Ito
ito@hosei.ac.jp

[1] Hosei University, Tokyo, Japan

[2] Murata Manufacturing Co., Ltd., Nagaokakyo, Japan

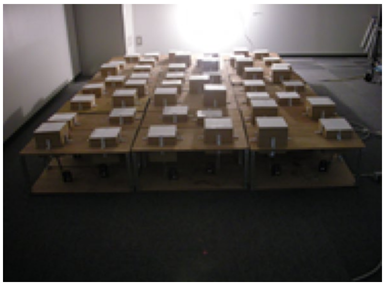


Fig. 1 Example of a dynamic operating environment

[12], or ecological psychology [13], it is thought that the body plays an important role in realizing adaptive behaviors.

In this paper, we have designed the mechanical body of a robot such that it can abstract state-action space, on the basis of our previous work [14, 15]; this design is expected to solve the problems outlined above. Using the proposed mechanism, the size of the state-action space can be reduced drastically, so that it is possible for the robot to learn in real time. Moreover, the robot's obtained policy is generalized and is applicable without additional learning even if the environment is changed dynamically.

To demonstrate the effectiveness of the proposed mechanism, we conduct experiments. The task of the robot is to learn the effective behavior for moving towards a light source in a dynamic environment. There are many unknown obstacles in the environment, and the obstacles move up and down constantly. The learning process is performed in both the static and dynamic environments, and the obtained policy is applied to other environments as well, where the obstacles are placed in different positions.

2 Task and environment

Figure 1 shows a dynamic environment. The environment is composed of 24 modules, and each module has two obstacles that move vertically. The range of movement of the obstacles ranges from a height of 5–12 cm. The dimensions of the environment are 240 cm (length) and 180 cm (width). We employ five different modules, as shown in Fig. 2, all of which are placed randomly. There is one light source that the robot must move towards, and the aim of the task is to learn effective behavior for homing in, on the light source within this 3D dynamic environment.

3 Proposed method

3.1 General framework

Figure 3 shows the general framework for the abstraction of the state-action space that we proposed in past studies [14, 15]. The robot consists of the generalization module and the learning module. A remarkable concept central to the proposed framework is that the generalization module is realized by the body of the robot itself instead of being implemented in a remote computer. The body is designed such that the necessary calculations for generalization are performed by utilizing the physical properties of the real world. The abstracted information is passed to the learning module in a computer, where the learning process is actually completed, with the selected action being fed back to the generalization module. The generalization module embodies the abstracted action that enables the complex movement of the robot. In the following subsection, we discuss the design the body of a snake-like robot on the basis of this framework.

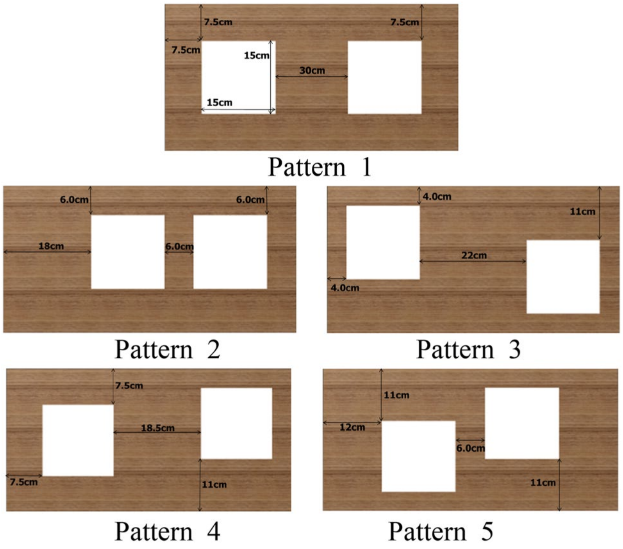


Fig. 2 Layout pattern of the obstacles

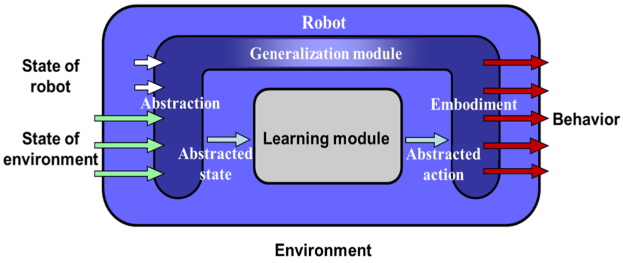


Fig. 3 General framework

3.2 Hardware design of body

In this study, we develop a 3D snake-like robot by improving our previous robot [14]. Figure 4 shows the mechanism of the robot. The robot is composed of three links that have crawlers. The joints are realized by rubber poles and are

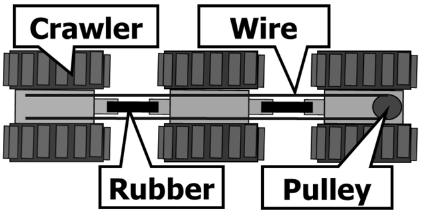


Fig. 4 Outline of robot

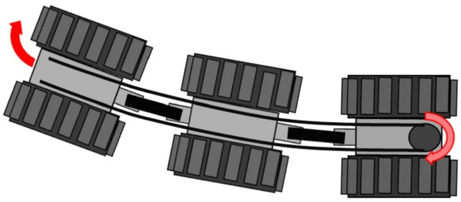
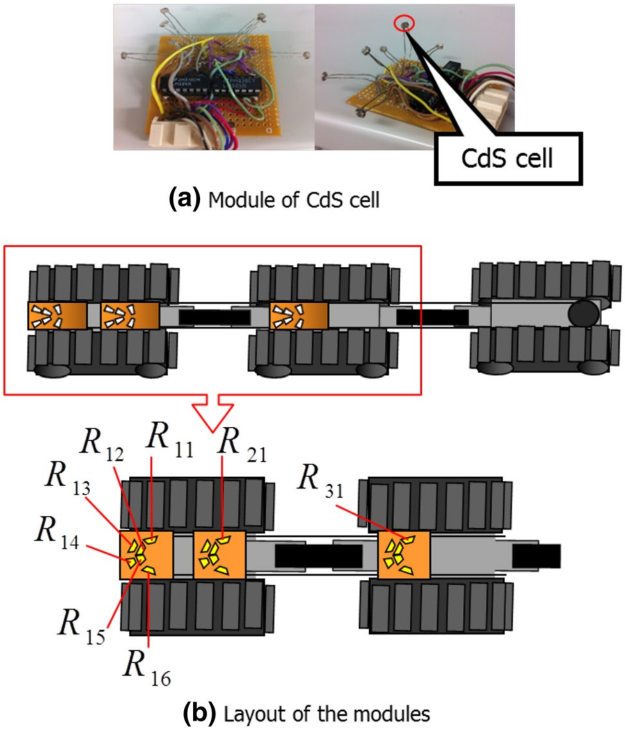


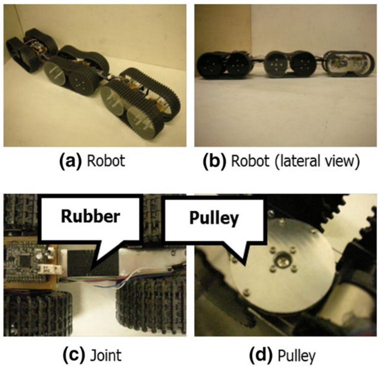
Fig. 5 Mechanism for pulling wires



(a) Module of CdS cell

(b) Layout of the modules

Fig. 6 Sensing system using CdS cells



(a) Robot **(b)** Robot (lateral view)

(c) Joint **(d)** Pulley

Fig. 7 Developed robot

moved passively. Two wires are installed on both the sides, and their length can be adjusted by an active pulley that is mounted on the rear end. Upon turning the active pulley, one wire is rolled up while the other wire is loosened. This causes the body of the robot to turn, as shown in Fig. 5.

Due to the passive nature of the joints and the constraining influence of the wires, the robot can adapt to a bumpy and unstable environment. The robot's direction of movement can be controlled by just actuating the active pulley.

3.3 Hardware design for sensing

We employ 18 CdS cells for detecting the direction that the light is coming from, as shown in Fig. 6. Each module is composed of 6 CdS cells (Fig. 6a), and three such modules are embedded on the robot as shown in Fig. 6b). The CdS cells have directional characteristics and their layout is hemispherical. Therefore, the light direction can be obtained by Eq. (1):

$$x = \frac{3\sum_{i=1}^{3}(1/R_{i1}) + 2\left\{\sum_{i=1}^{3}(1/R_{i2}) + \sum_{i=1}^{3}(1/R_{i3})\right\} + \sum_{i=1}^{3}(1/R_{i4}) + \sum_{i=1}^{3}(1/R_{i5})}{\sum_{j=1}^{6}\sum_{i=1}^{3}\left(1/R_{ij}\right)} \tag{1}$$

where R_{ij} denotes the electrical resistance of the j-th CdS cell in the i-th module. x represents the center of gravity of the light intensity and is equivalent to the light direction. For more details, please refer to [14].

3.4 Developed robot

Figure 7 and Table 1 show the developed robot and its specification. The passive joints are realized by using rubber poles. To obtain adequate elasticity, the dimensions of the rubber poles are very important. However, it is very difficult to determine the dimensions theoretically. So, in this paper, these sizes were tuned by preliminary experiments.

Figure 8 shows the result of the preliminary experiment performed for examining the capability of the sensor module. From the results, we find that the horizontal direction of the light can be obtained from the output voltage.

Table 1 Specification of the developed robot

Length	85 cm
Width	13 cm
Height	11 cm
Weight	2 kg

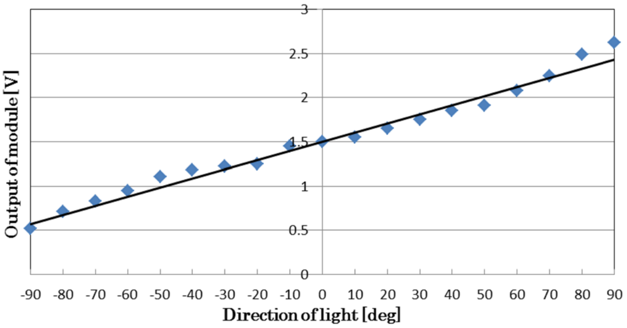


Fig. 8 Output of the sensor module

Table 2 States of the direction of light

State	Voltage [V]
0	(0.50, 1.00)
1	(1.00, 1.40)
2	(1.40, 1.60)
3	(1.60, 2.00)
4	(2.00, 2.50)

4 Experiment

4.1 Setting of reinforcement learning

We employ typical Q-learning [9] as reinforcement learning. Equation (2) represents Q-learning [9].

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\{r(s,a) + \gamma \max_{a'} Q(s',a')\} \tag{2}$$

where s is the state; a, the action; r, the reward; α, the learning rate; and γ, the discount rate.

We set α as 0.5 and γ as 0.9. The action is selected by using the ε–greedy method, and the probability of random selection of the action is 0.1.

One trial was conducted for a duration of 24 s, and calculations were performed every action is conducted by using Eq. (2).

Table 2 and Fig. 9 show the state of the light direction. The values are the outputs of Eq. (1). Table 3 shows the state of the body. These values are the angle of the active pulley. Table 4 shows the actions. While the robot is moving towards the light source (the state of the light direction is 2 and the value of the angle of the active pulley is 2), a reward of 100 is assigned for any action. When the robot loses the light source (light source goes out of the perceivable range), a negative reward −100 is assigned for any action. The trial then ends, and the next trial starts from initial position.

Table 3 States of the body

State	0	1	2	3	4
Degree [°]	−50	−25	0	25	50
State of robot	Left bend		Straight	Right bend	

Table 4 Action

Action	Motion
0	Turn tail motor by −25° (Robot will turn left)
1	Hold the tail motor (Maintain state of the body)
2	Turn tail motor by +25 (Robot will turn right)

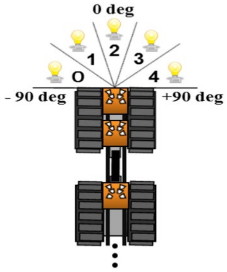


Fig. 9 States of the direction of light

4.2 Experiment

The obstacle is placed randomly as shown in Fig. 10, and we conduct the learning process by using a real robot. We consider two cases. One is a static environment where the obstacles do not move. The other is a dynamic environment in which objects move vertically. The learning process is conducted separately for each case.

After the learning process is completed, we apply the policy that is obtained in the static environment to the dynamic environment without additional learning, and we confirm the generalization capabilities of the proposed framework.

Figure 11 shows the learning curves, and Figs. 12 and 13 show the behavior after learning. In both cases, the value of the obtained reward has converged during the 40th trial.

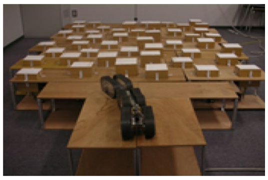


Fig. 10 Environment

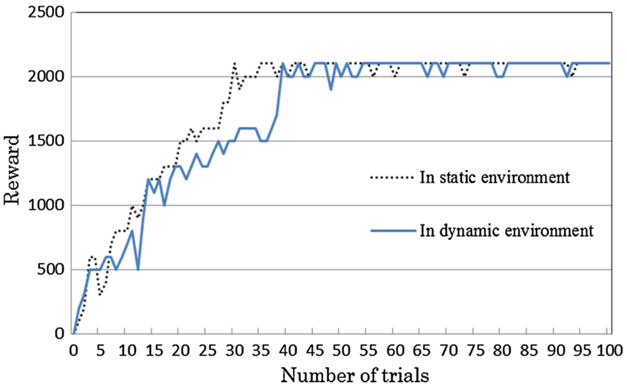


Fig. 11 Learning curve

This implies that the learning can be completed in a reasonable time. From Figs. 12 and 13, we can confirm that though the moving direction of the robot is changed by the obstacles, the robot can always recover and can move to the light source.

From the result, we conclude that the problem of real-time learning is solved and effective policy is obtained.

Next we consider the generalization capability of the proposed framework. Figure 14 shows an example of behavior realized by applying the policy obtained in the static environment to the dynamical environment. In this case, there was no additional learning. Nevertheless, the robot could move to the light source. This implies that the obtained policy can be generalized and is applicable to unknown but similar environments without requiring re-learning.

To confirm the generality, we compare state transitions. Figure 14 shows an example of a state transition graph in the static environment, and Fig. 15 shows one for the dynamic environment. From these figures, we found that similar policy was obtained. The policy is very simple, it is “If the light is on the left (right), then turn the motor to the left (right)”. The differences of the environment is abstracted by the body. So, the robot can behave effectively by the simple obtained policy even in such complex environment. We can confirm that effective generalized policy was obtained.

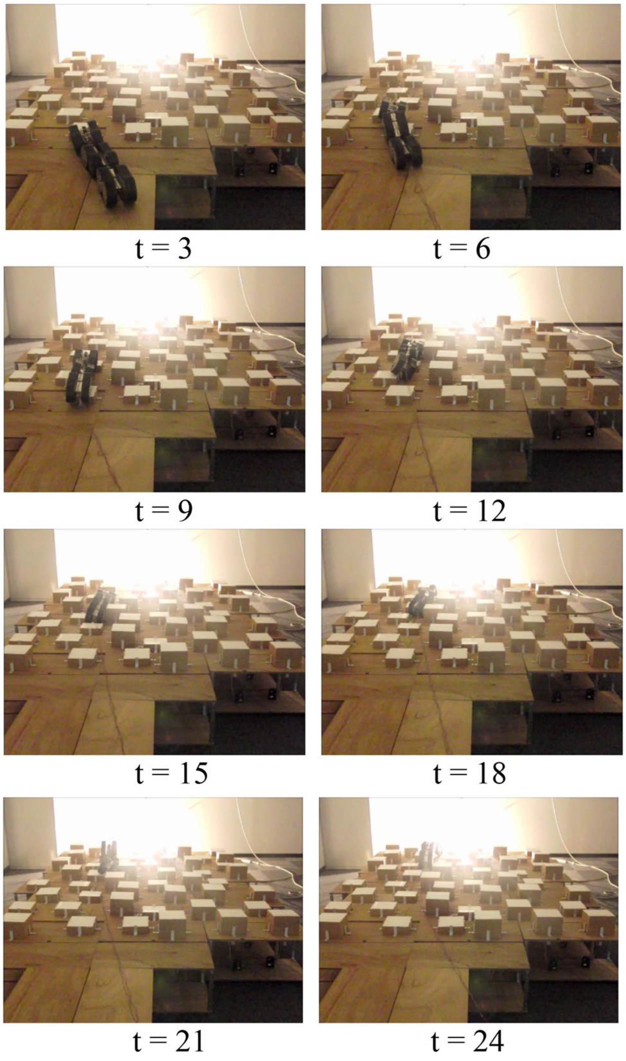
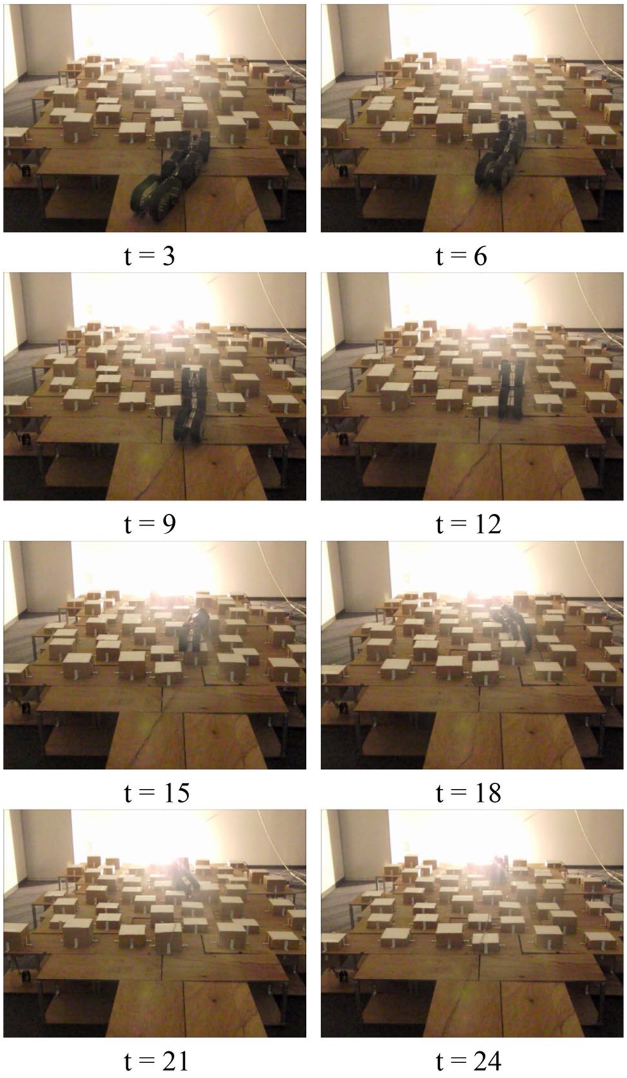


Fig. 12 Behavior after learning (static environment)

5 Conclusion

In this study, we examined autonomous control of a multi-crawler robot in a dynamic environment through the use of reinforcement learning. To solve the problems inherent to conventional approaches, which pertain to real-time learning and a lack of generality, we redesigned the mechanical body to abstract state-action space by utilizing the physical properties of the body. To demonstrate the effectiveness of our proposed approach, a prototype robot was developed and experiments were conducted. As a result, the behavior for moving towards a light source was learned within a reasonable time limit and the obtained policy could be generalized.


t = 3

t = 6

t = 9

t = 12

t = 15

t = 18

t = 21

t = 24

Fig. 13 Behavior after learning (dynamic environment)

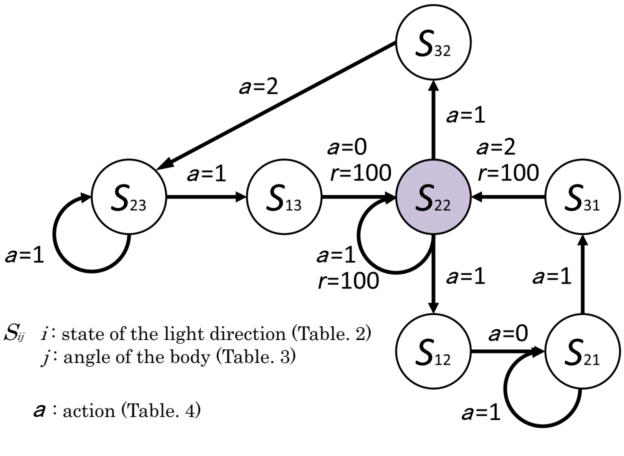


Fig. 14 State transitions in the static environment

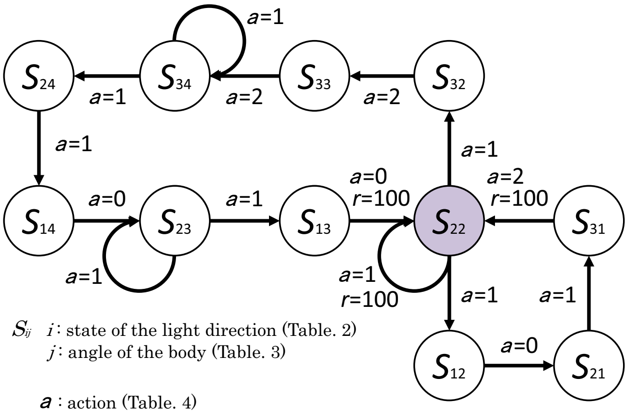


Fig. 15 State transitions in the dynamical environment

We conclude that the proposed framework is effective for developing autonomous robots that operate in dynamic environments.

Acknowledgments This study was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (Grant-in-Aid for Young Scientists (B), 22700156, 2011).

References

1. Arai M, Takayama T, Hirose S (2004), Development of Souryu-III: Connected crawler vehicle for inspection inside narrow and winding spaces, Proc. Int. Conf. Intelligent Robots and Systems, p 52–57
2. Paap KL, Christaller T, Kirchner F (2000) A robot snake to inspect broken buildings, Proc. Int. Conf. Intelligent Robots and Systems, p 2079–2082
3. Wolf A, Brown HB, Casciola R et al (2003) A mobile hyper redundant mechanism for search and rescue tasks, Proc. Int. Conf. Intelligent Robots and Systems, p 2889–2895
4. Kamegawa T, Yamasaki T, Igarashi H et al (2004) Development of the snake-like rescue robot "KOHGA," Proc. 2004 IEEE Int. Conf. on Robotics and Automation, p 5081–5086
5. Yamada H, Mori M, Hirose S (2007) Stabilization of the head of an undulating snake-like robot, Proc. Int. Conf. Intelligent Robots and Systems, p 3566–3571
6. Ito K, Fukumori Y (2006) Autonomous control of a snake-like robot utilizing passive mechanism, Proc. 2006 IEEE Int. Conf. Robotics and Automation, p 381–386
7. Ito K, Kamegawa T, Matsuno F (2003) Extended QDSEGA for controlling real robots -Acquisition of locomotion patterns for snake-like robot-, Proc. 2003 IEEE Int. Conf. Robotics and Automation, Sep 14–19, p 791–796
8. Murai R, Ito K, Matsuno F (2006) An intuitive human-robot interface for rescue operation of a 3D snake robot, Proc. 12th IASTED Int. Conf. Robotics and Applications p138–143
9. Watkins CJ, Dayan P (1992) Q-learning. Mach Learn 8:279–292
10. Kober J et al (2013) Reinforcement learning in robotics: a survey. Int J Robot Res 32(11):1238–1274
11. Kimura H, Yamashita T, Kobaysahi S (2001) Reinforcement learning of walking behavior for a four-legged robot, Proc. 40th IEEE Conf. Decision and Control, p 411–416

12. Pfeifer R (2001) "Understand Intelligence," The MIT Press, New edition
13. Gibson JJ (1987) The ecological approach to visual perception. Hillsdale, NJ, Lawrence Erlbaum Associates
14. Ito K, Takayama A, Kobayashi T (2009) "Hardware design of autonomous snake-like robot for reinforcement learning based on environment -Discussion of versatility on different tasks-," The 2009 IEEE/RSJ Int. Conf. Intelligent Robots and Systems, p 2622–2627
15. Ito K, Fukumori Y, Takayama A (2007) Autonomous control of real snake-like robot using reinforcement learning -abstraction of state-action space using properties of real world-, Proc. Int. Conf. Intelligent Sensors, Sensor Networks and Information Processing, p 389–394