JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 19, No. 1, pp. 35-41, Jan. 2018

강화학습 Q-learning 기반 복수 행위 학습 램프 로봇

권 기 현[1] · 이 형 봉[2]*
[1]강원대학교 정보통신공학과
[2]강릉원주대학교 컴퓨터공학과

Multi Behavior Learning of Lamp Robot based on Q-learning

Ki-Hyeon Kwon[1] · Hyung-Bong Lee[2*]
[1]Department of Information & Communication Engineering, Kangwon National University, Samcheok 25913, Korea
[2*]Department of Computer Science & Engineering, Gangneung-Wonju National University, Wonju 25457, Korea

[요 약]

강화학습기반 Q-learning 알고리즘은 이산적인 상태와 액션의 조합을 사용하여, 한 번에 하나의 행위에 대한 목표를 학습하는데 유용하다. 여러 액션을 학습하기 위해서는 행위 기반 아키텍처를 적용하고 적절한 행위 조절 방법을 사용하면 로봇으로 하여금 빠르고 신뢰성 있는 액션을 가능하게 할 수 있다. Q-learning은 인기 있는 강화학습 방법으로 단순하고, 수렴성이 있고 사전 훈련 환경에 영향을 덜 받는 특성(off-policy)으로 인해 로봇 학습에 많이 사용되고 있다. 본 논문에서는 Q-learning 알고리즘을 램프 로봇에 적용하여 복수 행위(사람인식, 책상의 물체 인식)를 학습시키는데 사용하였다. Q-learning의 학습속도(learning rate)는 복수 행위 학습 단계의 로봇 성능에 영향을 줄 수 있으므로 학습속도 변경을 통해 최적의 복수 행위 학습 모델을 제시한다.

[Abstract]

The Q-learning algorithm based on reinforcement learning is useful for learning the goal for one behavior at a time, using a combination of discrete states and actions. In order to learn multiple actions, applying a behavior-based architecture and using an appropriate behavior adjustment method can make a robot perform fast and reliable actions. Q-learning is a popular reinforcement learning method, and is used much for robot learning for its characteristics which are simple, convergent and little affected by the training environment (off-policy). In this paper, Q-learning algorithm is applied to a lamp robot to learn multiple behaviors (human recognition, desk object recognition). As the learning rate of Q-learning may affect the performance of the robot at the learning stage of multiple behaviors, we present the optimal multiple behaviors learning model by changing learning rate.

색인어 : 강화 학습, 행위 조절, 램프 로봇, Q-러닝

Key word : Reinforcement Learning, Behavior Coordination, Lamp Robot, Physical Robot, Q-learning

http://dx.doi.org/10.9728/dcs.2018.19.1.35

Received 07 December 2017 ; **Revised** 23 January 2018
Accepted 29 January 2018

***Corresponding Author; Hyung-Bong Lee**

Tel: +82-33-760-8668
E-mail: hblee@gwnu.ac.kr

Ⅰ. Introduction

A behavior-based architecture is an important concept for implementing a fast and reliable robot. A behavior-based robot does not require modeling of a specific surrounding environment in order to achieve a given task. Any environment model that is necessary to operate the robot is sufficient. Another advantage of this structure is that all actions can be executed in parallel, simultaneously, and asynchronously [1-2].

In the behavior-based structure, the robot needs a behavior coordinator. The subsumption architecture, one method of behavior-based structure may be classified as the competitive method, and it is possible to make the robot learn only one behavior (operation) at a time. This method produces high performance and quick results using a simple method, but in the robot motion, the action is not smooth and accuracy may be somewhat reduced.

To anticipate uncertain circumstances, the robot must have a learning mechanism. While in the supervised learning, a robot learns depending on external knowledge, in the mechanism of unsupervised learning, a robot learns the surrounding environment for itself. The reinforcement learning trains the given situation through the unsupervised learning, and the robot can learn by merely obtaining a reward in the given environment [3-4].

There are several methods that can solve the reinforcement learning problem, but one of the most popular methods is the Temporal Difference Algorithm, especially, the Q-learning algorithm[5-7]. The advantages of Q learning are that the dependence on the given environment is low (off-policy), and algorithms are simple, and converge to optimal policy. However, it can only be used in the state/action of discrete state and if the size of the Q table is large, the algorithm may take a long time to learn[8-10].

Learning algorithms usually use a lot of memory in the control part of the robot, and as the complexity of the given environment increases, the program complexity also increases.

In this paper, we apply the Q-learning algorithm to a lamp robot and present the implementation of algorithms for learning multiple behaviors (human recognition, desk object recognition). A single behavior is learned by a Q-learning algorithm and multiple behaviors are implemented in a form coordinated by subsumption architecture. In Chapter 2, we summarize the Behavior Coordination Architecture and the Q-learning algorithm, and Chapter 3, we explain about the implementation of the lamp robot. And in Chapter 4, we present experimental results and come to conclusion.

Ⅱ. Behaviors Coordination & Q-learning

2-1 Behaviors Coordination Method

It is important to define the behavior of a robot through the method of distinguishing and coordinating the behavior of a robot, and to design the robot so that appropriate actions come out in the real world.

There are competitive and cooperative methods for this. In the cooperative method, only one behavior can be applied to a robot at a time, and the representative of it is subsumption architecture [1]. This method is a type that disassembles the behaviors into several levels, and the behaviors of high levels have high priority, and the behaviors of high levels subsume the behaviors of low levels. A hierarchical behavior control system with subsumption architecture is shown in Fig 1.

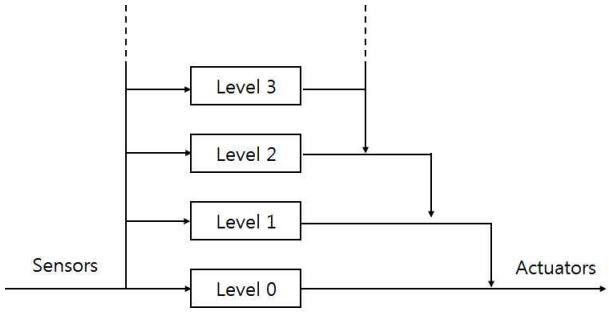

그림 1. 계층형 행위 제어 시스템
Fig. 1. Layered control system [1]

Behaviors for the actions of the lamp robot are 1. waiting, 2. situation whether a person is approaching 3. situation whether the lamp must be lit up on the desk, or 4. stops.

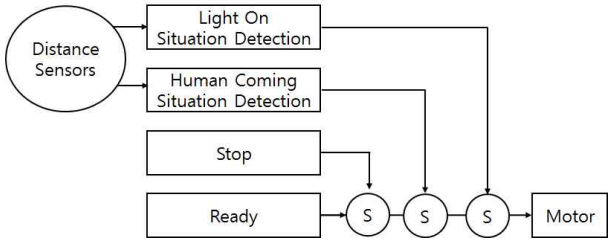

그림 2. 램프 로봇 행위 포섭 구조
Fig. 2. Subsumption Architecture for Lamp Robot

These behaviors must be well coordinated and enable the robot actions synchronously. Behaviors coordination method used in this paper is subsumption architecture, and it is shown as Fig 2.

In Fig 2, Ready is the behavior at the lowest level, and when there is a behavior of higher level, the behaviors of the low levels

are inactive, and it becomes a criterion to judge whether a person is approaching (it reacts from 2m and approaches near 50cm), and whether the lamp must be lit up on the desk according to the learning of the distance sensor value (near 30cm or less).

2-2 Q-learning

Reinforcement learning is a form of unsupervised learning, a method to go on learning from the robot's environment. While learning from the given environment, the robot goes on learning by updating the Q table through the value of the Q table which is the content learned so far and the reward which is obtained when the goal is reached.

Q-learning is used much for robot learning for its characteristics which are simple, convergent and little affected by the prior training environment (off-policy) as a reinforcement learning method. Fig 3 shows the basic structure of reinforcement learning.

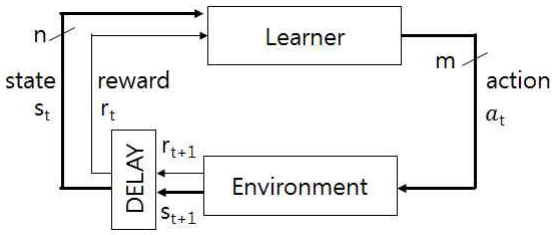

그림 3. 강화 학습 기본 구조
Fig. 3. Reinforcement learning basic scheme [3]

Algorithm1 One-Step Q-learning Algorithm

Initialize $Q(\boldsymbol{s}, a)$ arbitrarily

Repeat(for each episode):

1. Perceive a state $s \in \boldsymbol{s}$, and then select an action $a \in A$ using policy derived from $Q(\boldsymbol{s}, a)$
2. Repeat(for each episode)
3. Take action a, observer next state $\boldsymbol{s}'$ and next reward r.
4. Update the action-value function by $(\boldsymbol{s}, a, \boldsymbol{s}', r)$.
5. $\boldsymbol{s} \leftarrow \boldsymbol{s}'$
6. until $\boldsymbol{s}$ is terminal

Q-learning is widely used TD (temporal difference) and unsupervised learning (model-free RL) [11-13]. During the learning phase, the agent (robot) selects an action $a \in A(s)$ in the state $s \in S$ according to the given policy. After selecting the action, it relays to the next state s', and receives the reward r. And according to the value of reward r, an action may be strengthened or weakened

In state s and action a, the Q table is updated as shown in the following formula (1) [14].

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r + \gamma \max Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \quad (1)$$

$\alpha \in (0,1]$ is the learning rate and $\gamma \in (0,1]$ is the discount rate of the learned value obtained based on the current Q table learning. After repeated learning, the value of Q table converges to Q^* [3]. The Q-learning algorithm is presented in Algorithm 1.

It is important to well define the state, action and reward in Q-learning to improve learning performance.

In the lamp robot, the values of the ultrasonic sensor to obtain the values of the distance to the motor corresponding to the joint of the lamp are used as the state. The motors are used at least two (shoulder, head), and in the state that the rotation angle of the servomotor is limited to the maximum 180^o, it is divided into 10 equal parts, so each motor has 10 states for each motor, and within the range of ultrasonic sensor values (3cm ~ 3m), it is divided into 10 equal parts, so it is made to have 10 states according to distance (Table 1).

In the lamp robot, “up” and “down” of the servomotor, and “on” and “off” as the lighting of the light are used as the action.

표 1. 램프 로봇의 상태(state) 정의
Table. 1. States of Lamp Robot

State	Value
Servo Motor 1(Shoulder)	$a \times (180/10), where\ \ a = \angle$
Servo Motor 2(Head)	$a \times (180/10), where\ \ a = \angle$
Distance Sensor(SRF04)	$d \times (300/10), where: d = distance$

표 2. 램프 로봇의 리워드(reward) 정의
Table. 2. Reward Value of Lamp Robot(s: scale, 1~10)

State	Reward
Human Coming(Shoulder)	$((\boldsymbol{s}/2) - abs(s)) - \boldsymbol{s}/2$
Human Coming(Head)	$s/2 - \boldsymbol{s}/4$
Light On(Shoulder)	$((\boldsymbol{s}/2) - abs(s)) - \boldsymbol{s}/2$
Light On(Head)	$((10/2) - s)/2$

For the reward used in the lamp robot, servo motor1 (shoulder), servo motor2 (head), reward are used as shown in Table 2 depending on whether it is a situation in which a person is approaching or a situation in which the lamp must be lit on the desk.

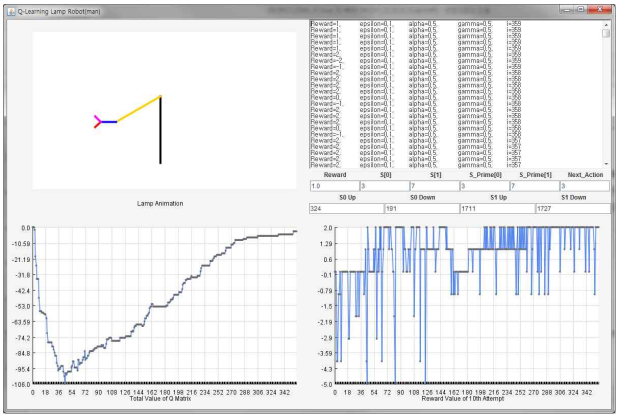

그림 4. 램프 로봇 시뮬레이션(사람 인식)
Fig. 4. Lamp Robot Simulation for Human Recognition(learning rate: 0.5, discount rate: 0.5, random action ratio: 0.1)

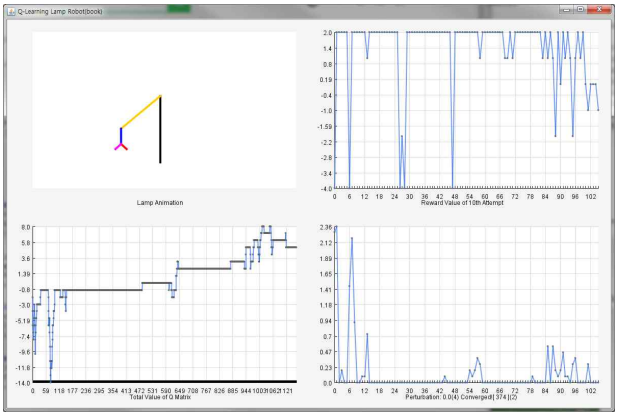

그림 5. 램프 로봇 시뮬레이션(책상의 물체 인식)
Fig. 5. Lamp Robot Simulation for Lamp Light(learning rate: 0.5, discount rate: 0.5, random action ratio: 0.1)

Ⅲ. Lamp Robot Implementation

For implementation of the lamp robot, we examined whether the states, actions and reward given through the simulation work as predicted, and we made a physical robot, worked porting and checked the motion.

3-1 Lamp Robot Simulation

To teach multiple behaviors by applying the Q-learning algorithm to the lamp robot, we performed simulation by classifying the case of human recognition and the object recognition of the desk

For the default value, we set the learning rate as 0.5, the discount rate as 0.5 and epsilon as 0.1 to randomly determine when deciding on the next action, and we performed.

Using the accumulated value of the Q table, we judged whether Q was converging, and every time the learning scenario repeated ten times, we got the change in reward and checked whether the reward is converging to a larger value. Also, we showed in graph the perturbation of the Q table value accumulated while learning and converging, and when the number of repetition in which perturbation stays within the threshold is continuously 20 or more, we judged as convergence, and finished learning

Fig 4 shows that it learned the situation where a person is approaching, and the shape of the lamp faces the direction of a person, and Fig 5 shows that it learned the situation of turning on the lamp light on the desk, and the shape of the lamp faces the direction of a desk.

3-2 Physical Lamp Robot

For the controller of the physical lamp robot, we used Arduino UNO, and implemented using the distance sensor and the servomotor. The speculation of used sensors and hardware are shown in Table 3.

표 3. 하드웨어 스펙
Table. 3. Hardware Specification

Item	Value
Distance Sensor	· SRF04 · Dual Type UltraSonic Sensor · Range 3cm ~ 3m
Controller Spec.	· Arduino UNO R3 · Microcontroller ATmega328 · SRAM 2 KB (ATmega328) · Clock Speed 16 MHz
Servo Motor	· RBM-606MG Metal Servo Motor · Toque(Kg.cm) : 10Kg↓

그림 6. 램프 로봇 제작
Fig. 6. Physical Lamp Robot

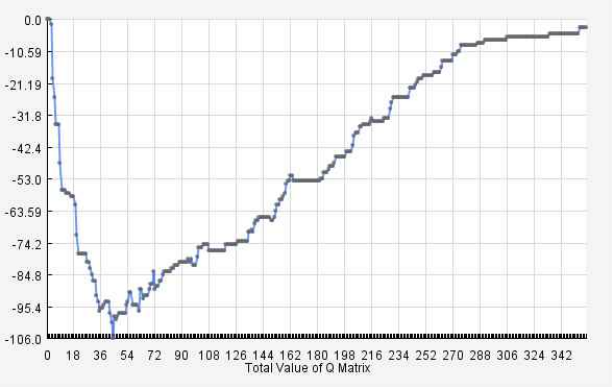

그림 7. 누적 Q 테이블
Fig. 7. Accumulated Q Table

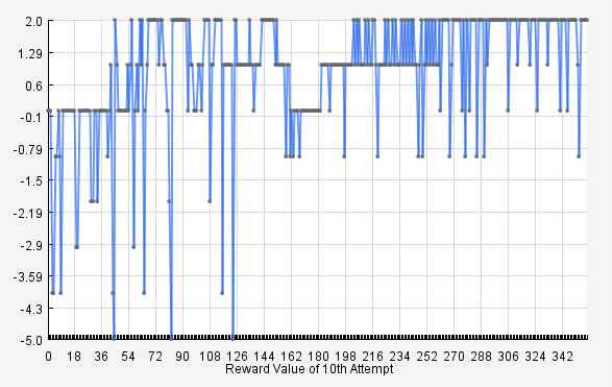

그림 8. 리워드 변화
Fig. 8. Reward Value Converge while Learning

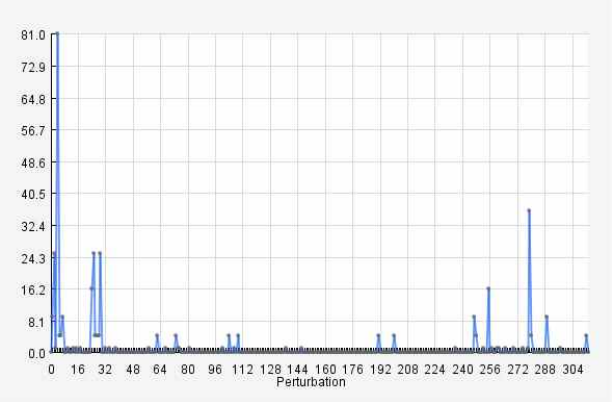

그림 9. 수렴된 Q 테이블의 값의 차이
Fig. 9. Perturbation Converge while Learning

Ⅳ. Experimental Result

The execution time and the number of repetition of the convergence scenario are different each time it is executed, not only in simulation but also in the case of a physical lamp robot. This is the result of intelligent processing of Q-learning, and it does not show the same result each time, but shows functionally similar results.

Fig 7 shows the value of the cumulative Q table that can be seen by the learned knowledge of the robot, that is, the brain, at the stage of 300 times. It can be seen that the increase value of the Q table becomes gentle and stabilized while learning.

Fig 8 shows the value of the reward acquired while the robot is learning, at the stage of 300 times. We can see that the magnitude of reward gradually increases while learning and that as it goes to the latter half, the reward becomes higher value of 2.0.

Fig 9 shows the difference in the values of the Q table at the stage of 300 times, and we can see that the perturbation of the Q table values accumulated while learning and converging is greatly reduced, and that some fluctuations occur in the course of determining the next action at random.

Figure 10 shows the number of scenarios converged through the perturbation of accumulated Q table values by changing the learning rate to 0.25, 0.5, 0.75 and 1.0, and that it converges most rapidly when the learning rate is 0.5. We can see that it is very important to set an appropriate learning rate according to the environment given to the robot because it takes more time to converge when the learning rate is 0.25 or even more than 0.5.

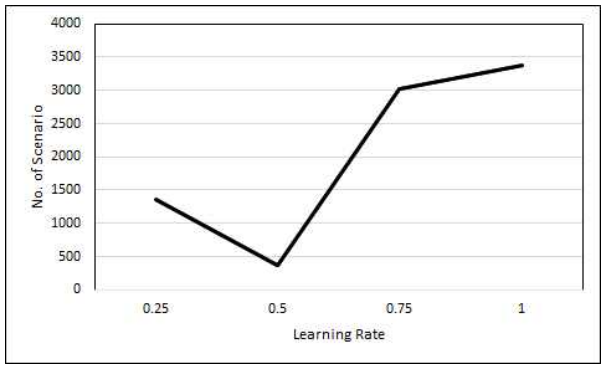

그림 10. 학습속도 변경에 따른 Q learning
Fig. 10. Q learning Behavior with Varying Learning Rate

Ⅴ. Conclusions

Q-learning algorithm based on reinforcement learning is applied to the lamp robot and used to teach multiple behaviors (human recognition, object recognition of the desk), and it shows that it works well through simulation and a physical robot.

It shows that the reward obtained from the Q-learning is continuously accumulated as a positive value and it works well when the perturbation according to the variation of the learning number of times of the Q table cumulative values becomes smaller than the threshold value.

As the learning rate of Q-learning may affect the performance of a robot at the stage of multiple behavior learning, it is important to obtain the optimal learning rate through the coordination of learning rate.

Acknowledgement

This study was supported by 2017 Research Grant from Kangwon National University(No. 620170051)

References

[1] R. Brooks, "A Robust Layered Control System For a Mobile Robot," *IEEE Journal of Robotics and Automation*, Vol. 2, No. 1, pp. 14 – 23, 1986.

[2] R. Hafner, and M. Riedmiller, "Reinforcement Learning on a Omnidirectional Mobile Robot," in *Proceeding of 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 1, Las Vegas, pp. 418 – 423, 2003.

[3] R.S. Sutton, and A.G. Barto, "Reinforcement Learning, an Introduction," *MIT Press*, Massachusets, 1998.

[4] H. Wicaksono, Prihastono, K. Anam, S. Kuswadi, R. Effendie, A. Jazidie, I. A. Sulistijono, M. Sampei, "Modified Fuzzy Behavior Coordination for Autonomous Mobile Robot Navigation System," in *Proceeding of ICCAS-SICE*, 2009.

[5] C. Watkins and P. Dayan, "Q-learning, Technical Note," *Machine Learning*, Vol 8, pp. 279-292, 1992.

[6] Y. G. Seo, “LoRa Network based Parking Dispatching System : Queuing Theory and Q-learning Approach,” *The Journal of Digital Contents Society*, Vol. 18, No. 7, pp. 1443-1450, June 2017.

[7] K. Anam, S. Kuswadi, "Behavior Based Control and Fuzzy Q-Learning For Autonomous Mobile Robot Navigation," in *Proceeding of The 4th International Conference on Information & Communication Technology and Systems (ICTS)*, 2008.

[8] S. M. Rho, “LoRa Network based Parking Dispatching System : Queuing Theory and Q-learning Approach,” *The Journal of Digital Contents Society*, Vol. 18, No. 7, pp. 1443-1450, June 2017.

[9] M.C. Perez, A Proposal of Behavior Based Control Architecture with Reinforcement Learning for an Autonomous Underwater Robot, Ph.D. Dissertation, University of Girona, Girona, 2003.

[10] L. Khriji, F. Touati, K. Benhmed, A.A. Yahmedi, "Q-Learning Based Mobile robot behaviors Coordination," in *Proceeding of International Renewable Energy Congress (IREC*), 2010.

[11] C. J. C. H.Watkins, Learning from delayed rewards, Ph.D. dissertation, Dept. Psychol., Univ. Cambridge, Cambridge, U.K., 1989.

[12] H.Wicaksono, "Q Learning Behavior on Autonomous Navigation of Physical Robot," *The 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI 2011)*, in Songdo Convention, Incheon, Korea, Nov. 23-26, 2011.

[13] C. F. Touzet, "Q-learning for robot," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed. Cambridge, MA, USA: MIT Press, pp. 934-937, 2003.

[14] J.L LIN, K.S. HWANG, W.C. JIANG, and Y.J. CHEN, "Gait Balance and Acceleration of a Biped Robot Based on Q-Learning," IEEE Access, Vol. 4, pp. 2439-2449, 2016.

Ki-Hyeon Kwon

1993 : Kangwon National University of Computer Science(B.S Degree).
1995 : Kangwon National University of Computer Science(M.S Degree).

2000 : Kangwon National University of Computer Science(Ph.D Degree).
2002~now : Professor in Kangwon National University, Samcheok, Korea.
※Research Interests : Pattern recognition, Image processing and IoT

Hyung-Bong Lee

1984 : Seoul National University of Computer Science(B.S. Degree).
1986 : Seoul National University of Computer Science(M.S. Degree).

2002 : Kangwon National University of Computer Science(Ph.D Degree).
2004~now : Professor in Gangneung-Wonju National University, Wonju, Korea.
※Research Interests : Embedded Systems and Sensor Networks.