Journal of Korea Robotics Society (2019) 14(2):081-086
https://doi.org/10.7746/jkros.2019.14.2.081 ISSN: 1975-6291 / eISSN: 2287-3961

보틀플리핑의 로봇 강화학습을 위한 효과적인 보상 함수의 설계

Designing an Efficient Reward Function for Robot Reinforcement Learning of The Water Bottle Flipping Task

양 영 하[1] · 이 상 혁[1] · 이 철 수†
Young-Ha Yang[1], Sang-Hyeok Lee[1], Cheol-Soo Lee†

Abstract: Robots are used in various industrial sites, but traditional methods of operating a robot are limited at some kind of tasks. In order for a robot to accomplish a task, it is needed to find and solve accurate formula between a robot and environment and that is complicated work. Accordingly, reinforcement learning of robots is actively studied to overcome this difficulties. This study describes the process and results of learning and solving which applied reinforcement learning. The mission that the robot is going to learn is bottle flipping. Bottle flipping is an activity that involves throwing a plastic bottle in an attempt to land it upright on its bottom. Complexity of movement of liquid in the bottle when it thrown in the air, makes this task difficult to solve in traditional ways. Reinforcement learning process makes it easier. After 3-DOF robotic arm being instructed how to throwing the bottle, the robot find the better motion that make successful with the task. Two reward functions are designed and compared the result of learning. Finite difference method is used to obtain policy gradient. This paper focuses on the process of designing an efficient reward function to improve bottle flipping motion.

Keywords: Robotic Arm, Reinforcement Learning, Motion Tracking, Bottle Flipping

1. 서 론

로봇은 각종 산업 현장과 서비스 분야에서 활발하게 사용되고 있으며, 그 활용도는 점점 증가하는 추세이다. 하지만 아직까지 대부분의 로봇은 고정된 환경에서 반복적인 수행을 할 수 밖에 없는데 그 원인은 로봇의 제어 및 작동 방식에 기인한다. 로봇 작동의 전통적 방식은 환경 및 사물의 변화와, 복잡한 문제를 해결하기에 적합하지 않다. 그것은 로봇이 처한 환경과 다루어야 할 사물의 위치, 무게, 모양 등을 측정하기 힘들거나, 로봇과 사물의 상호 작용을 수식화 하기 어렵다면, 로봇의 동작 계획이 불가능 하기 때문이다.

로봇이 처한 환경과 사물의 역학적 측정 및 계산이 힘든 상황에서는 강화 학습을 적용하여 문제를 해결할 수 있다. 일반적으로 로봇에서의 강화 학습은 로봇이 처한 환경, 로봇과 사물의 상호 작용 등을 측정 및 계산하지 않는다. 강화 학습은 로봇이 동작을 수행하는 도중 또는 수행한 후 그 동작을 점수화 하여, 더 높은 점수를 도출할 확률이 높은 방향으로 다음번 동작의 변수를 개선해가는 방법이다. 이때 동작을 점수화 하는 보상 함수는 일반적으로 주어진 목적의 달성에 가까울수록 높은 점수를 도출한다. 그래서 강화 학습은 로봇이 처한 모든 상황과 로봇이 다루어야 할 사물의 거동을 정확하게 측정 및 계산할 수 없는 경우에도 로봇이 주어진 목적을 달성할 수 있다.

강화 학습은 학습이 시작될 때의 로봇의 초기 동작이 필요하다. 초기 동작은 학습의 수렴 속도에 큰 영향을 준다[1]. 로봇 강화 학습은 현실세계의 실제 로봇이 학습을 수행하는 것이므로, 컴퓨터 시뮬레이션과는 달리 시간, 로봇과 환경의 내구성 등의 영향을 받기 때문에 수행 횟수의 제약을 받는다[2]. 따라서 학습의 시작이 되는 로봇의 초기 동작은 학습 수렴 속도를 높이기 위해 최종적으로 달성하고자 하는 목적에 타당해야 한다.

초기 동작을 생성하는 방법은 다양하다. 사람이 직접 로봇을 잡고 움직인 경로를 모터 엔코더로 기록하는 방법[3,4], 가속도센서, 자이로스코프, 동작 센서 등을 사용자의 몸에 부착하

Received : Dec. 7. 2018; Revised : Jan. 11. 2019; Accepted : Jan. 15. 2019
1. MS Student, Mechanical Engineering , Sogang University, Seoul, Korea (dakmuk, jywhyuck@sogang.ac.kr)
† Professor, Corresponding author: Mechanical Engineering, Sogang University, Seoul, Korea (cscam@sogang.ac.kr)

여 동작을 수행한 후 센서 값을 로봇의 변수에 맞게 변환하여 로봇에게 동작을 지령하는 방법[5], 로봇 시뮬레이터 또는 물리 엔진을 이용하여 가상의 공간에 로봇과 환경을 구현한 후 원하는 초기 동작을 생성하는 방법이 있다[6]. 본 논문에서는 사람의 동작을 이미지 추적하여 초기 동작을 생성하였다.

공 던지기, 다트 던지기 등 다양한 주제에 초기 동작 생성, 보상값(Reward) 평가 방식, 경사법(Gradient method) 등에 대한 다양한 연구가 이루어지고 있다[7-11].

본 논문에서 강화 학습을 이용하여 로봇에게 학습시킬 주제는 보틀 플리핑이다. 이는 액체가 든 병을 던져 공중에서 일회전 시킨 후 바닥에 수직으로 착지시키는 놀이이다. 학습 알고리즘은 초기 동작과 평가 과정만을 중점적으로 다루기 위해 경사법 중 가장 간단한 FDM (Finite Difference Method)[11]알고리즘을 사용한다.

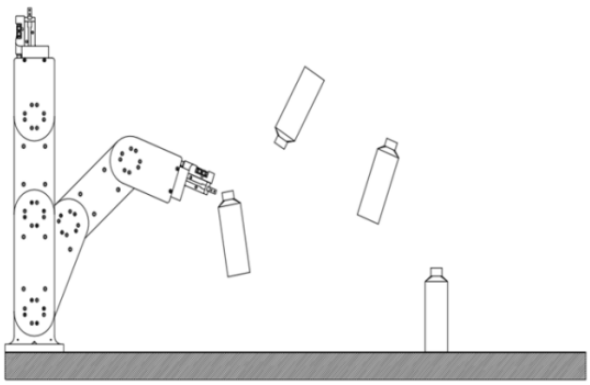
로봇이 수행할 놀이의 정의와 실험에 사용된 시스템은 2장에서 기술한다. 3장에서는 초기 동작 생성 및 학습 방법을 기술한다. 4장은 다양한 보상 함수를 통해 로봇 강화 학습을 위한 효과적인 보상 함수 설계 과정을 기술한다.

2. 강화 학습 임무 및 시스템 설계

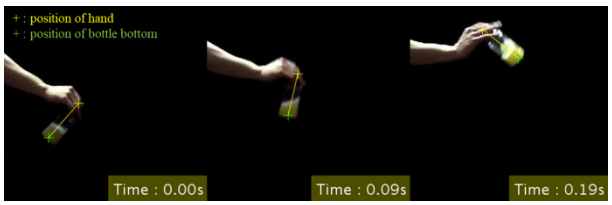
2.1 임무

본 논문에서 로봇에게 학습시킬 임무는 [Fig. 1]과 같이로봇이 물병을 던져 공중에서 일회전 시킨 후 바닥에 수직으로 착지시키는 놀이인 보틀 플리핑이다. 본 강화 학습은 로봇과 사물의 상호 관계를 찾지 않고 목적의 달성만을 고려하는 방식이다. 따라서, 로봇 각 관절의 길이 모터의 입력 펄스 대비 회전각 등의 기구학 및 병의 무게, 액체의 양 등의 환경 조건을 계산하지 않는다.

학습의 과정은 크게 두 부분으로 이루어 진다. [Fig. 2]와 같이 사람이 병을 던지는 동작을 촬영 후 로봇이 이미지 상의 손을 추적하여 초기 동작을 생성한다. 이를 통해 사람이 로봇을 직접 손으로 잡고 움직일 수 없거나, 빠른 동작으로 인해 구현하기 어려운 동작을 생성할 수 있다. 또 하나의 중점 내용은 보상 함수다. 보상 함수는 사용자의 관찰에 따라 주관적으로 만들어지므로 같은 임무에 대해서도 사람에 따라 다르게 설계할 수 있고, 실험을 통해 보정되는 상수들이 덧붙여진다. 본 논문의 4장에 한가지 임무에 대하여 다양한 보상 함수를 적용했을 때 학습 속도에 대한 영향을 기술하였다.



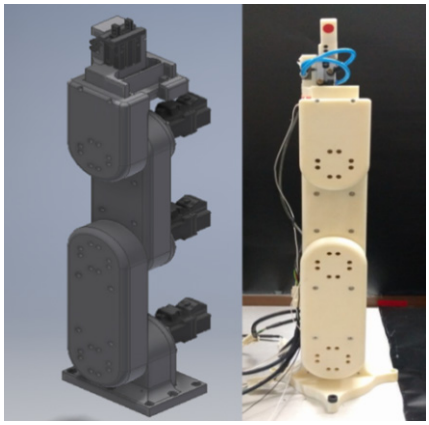
[Fig. 1] Sketch of the bottle flipping task



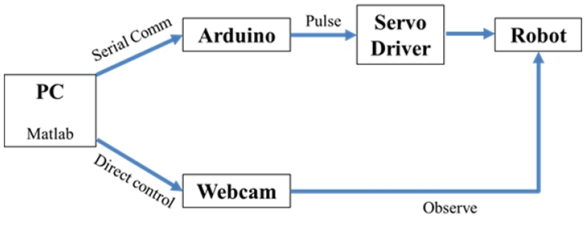
[Fig. 2] Demonstrated motion

2.2 시스템 설계

본 논문의 로봇 강화 학습에 사용되는 로봇은 [Fig. 3]와 같이 AC서보 모터와 공압 그리퍼를 이용한 3축 로봇이다. 로봇의 몸체는 3D 프린팅으로 제작되었으며 앞서 기술한 것과 같이 로봇 각 축의 길이와 말단부의 그리퍼 길이 등은 실험에 고려되지 않는다. 카메라로 위치를 측정하기 때문에 엔코더는 사용하지 않는다.



[Fig. 3] 3DOF Robot Arm



[Fig. 4] System flowchart

로봇을 구동하는 시스템은 PC와 Arduino로 구성되어 있다. [Fig. 4]과 같이 PC로 부터 Arduino로 펄스 수, 간격이 전달되고 Arduino는 그 값에 따라 모터에 펄스를 입력한다. 외부 카메라가 100 FPS의 간격으로 이미지를 캡쳐하여 PC에 전달한다. 이미지로 부터 추출한 픽셀 좌표로 보상 값을 계산한다[12].

3. 초기 동작 생성 및 학습 방법

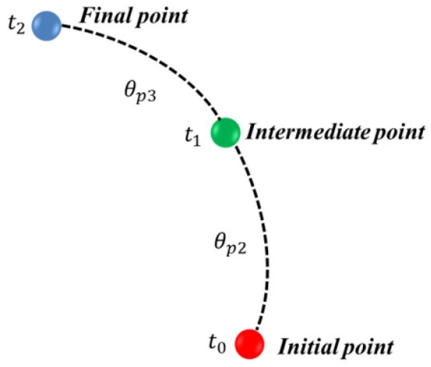
3.1 초기 동작 생성

사람이 물병을 던지는 것을 촬영하여 손과 물병의 바닥 부분의 점을 얻는다. [Fig. 2]의 시작, 중간, 마지막 세 부분의 프레임을 통해 얻은 세 쌍의 점을 순서대로 보간하여 [Fig. 5]와 같이 동작을 완성한다.

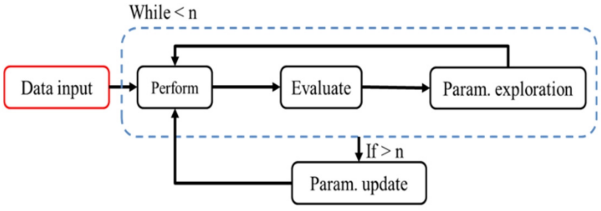
3.2 학습의 과정

학습은 [Fig. 6]의 프로세스로 진행된다. 입력값(Data input), θ_k는 각 모터와 그리퍼에 입력되는 변수다. 입력값을 받은 로봇이 동작을 실행(Perform)하고 보상 함수를 통해 보상값을 구해 동작을 평가한다.

$$\theta_{batch} = \begin{bmatrix} p_{m1_{batch}} \\ \vdots \\ p_{mi_{batch}} \end{bmatrix} = \begin{bmatrix} p_{m1} \\ \vdots \\ p_{mi} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \end{bmatrix} \quad (1)$$



[Fig. 5] An example of motion generation



[Fig. 6] General reinforcement learning process

[Fig. 6]의 Param. exploration은 식 (1)으로 입력값에 난수를 추가하는 과정이다. 난수를 생성하는 범위에 따라 결과의 수렴 속도가 차이날 수 있다. 반복하여 각각 Batch iteration이라 정의한다. 본 논문에서는 3회의 반복횟수로 실행했다. 다음 각각의 변수에 대해 동작을 수행하여 결과를 평가한다.

$$\theta_{k+1} = \theta_k + \alpha \nabla_\Theta J_\Theta \quad (2)$$

[Fig. 6]의 Param. Update는 Batch iteration의 변수와 평가값을 가지고 식 (2)를 이용하여 입력 값을 업데이트 한다. 이때 $\nabla_\Theta J_\Theta$는 FDM으로 계산한 모터 변수와 그 점수의 기울기(Gradient)이다. 여기에 상수α를 반영하여 더한다.

$$\nabla_\Theta J_\Theta = (\Delta\boldsymbol{\Theta}^T \Delta\boldsymbol{\Theta})^{-1} \Delta\boldsymbol{\Theta}^T \Delta\hat{\mathbf{J}} \quad (3)$$

식 (3)의 $\Delta\boldsymbol{\Theta}$는 Batch iteration을 생성할 때 입력값에 추가한 난수, $\Delta\theta_n \in \mathbb{R}^i$ 벡터를 모아놓은 행렬로 $\Delta\boldsymbol{\Theta} = [\Delta\theta_1, \cdots, \Delta\theta_i]^T$으로 나타낼 수 있다. $\Delta\hat{\mathbf{J}}$는 모든 Batch iteration에 대하여 로봇의 임무 수행 후 평가되는 보상 값을 난수를 추가하기 전의 보상값과의 차이, $\Delta\hat{J}_n$를 계산하여 모아놓은 행렬이다. 이는 $\Delta\hat{\boldsymbol{J}} = [\Delta\hat{J}_1, \cdots, \Delta\hat{J}_i]^T$로 나타낼 수 있다. 식 (3)으로 구한 기울기 값은 보상 값이 가장 크게 증가할 방향을 나타내며, 식 (2)에 대입하여 모터에 입력될 변수를 갱신하는데 이용된다.

4. 보틀 플리핑을 위한 보상 함수

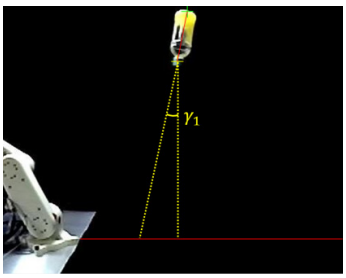
4.1 최고점과 착지 순간의 보상 함수

$$R = R_m + R_l \quad (4)$$

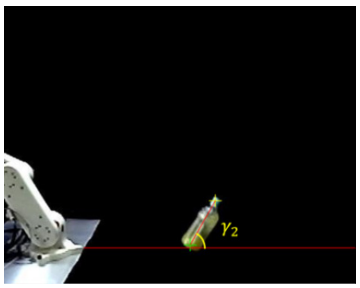
식 (4)의 보상 함수는 물병이 던져진 후 최고점에 도달했을 때의 보상 값, R_m과 물병이 바닥에 착지 하는 순간의 보상 값, R_l을 합산한다.

$$R_m = (\alpha_1 - |\gamma_1|) + \alpha_2 (z_g - z_0) \quad (5)$$

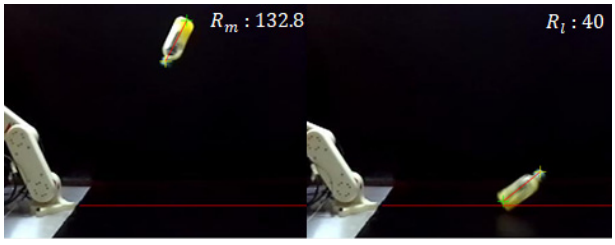
식 (5)는 물병이 최고점에 도달했을 때의 보상 함수다. 병 뚜껑의 위치 좌표, z_g과 바닥의 픽셀 위치 좌표, z_0로 부터 물병이 도달한 높이를 구하고 그때 지면과 이루는 각도, γ_1를 통해 보상 값을 구한다. γ_1는 [Fig. 7]과 같이 병이 지면에 대해 거꾸로 수직하게 서있는 상태를 기준으로 회전한 각도다. 이 보상 함수는 γ_1가 0에 가까울 수록 병뚜껑이 도달한 최고 높이가



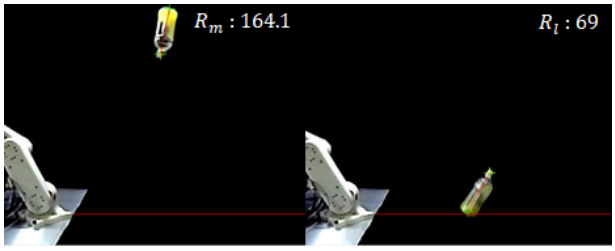
[Fig. 7] A reward function for the maximum height



[Fig. 8] Reward function for the landing moment



(a)



(b)

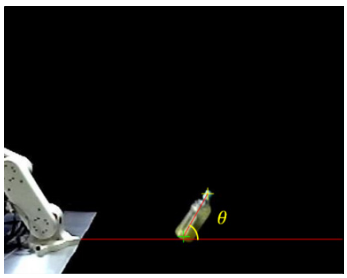
[Fig. 9] Results of the first reward function

높을수록 높은 보상 값을 갖도록 주관적인 판단에 의거하여 세운 식이다. α_1와 α_2는 보상 값을 보정하기 위한 상수다.

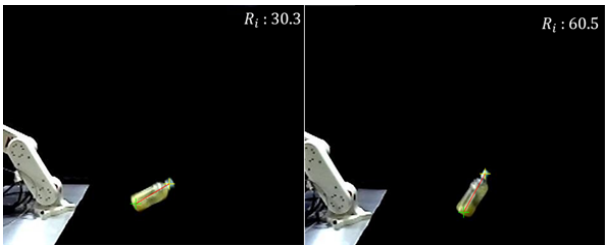
$$R_l = \alpha_3(90 - |90 - \gamma_2|) \tag{6}$$

식 (6)은 물병이 바닥에 착지하는 순간의 보상 함수다. 이때 [Fig. 8]과 같이 이미지로부터 추출하여 병이 지면과 이루는 각도, γ_2를 얻는다. 이 보상 함수는 물병이 지면과 이루는 각도가 수직에 가까울수록 최종 목표 달성의 가능성이 높다는 전제조건 하에 설계되었고 마찬가지로 관찰을 통한 주관적인 판단이 개입되었다.

앞서 기술한 보상 함수의 예시는 [Fig. 9]와 같다. [Fig. 9]의



[Fig. 10] Reward function consider only landing moment



[Fig. 11] Reward consider only landing moment

(a)의 합산된 보상 값은 233.1로 (b)의 172.8보다 높고 사진에서 확인할 수 있는 것과 같이 (b)보다 (a)의 최종 성공 가능성이 더 크다. 하지만 이러한 보상 함수는 주관적인 판단과 관찰이 크게 반영되었기 때문에, 물병이 최고점에 도달했을때의 보상 값과 최종 성공에 대한 상관 관계를 논리적으로 설명하기 어렵다.

4.2 착지 순간의 보상 함수

$$R = \beta_1 - |90 - \theta| \tag{7}$$

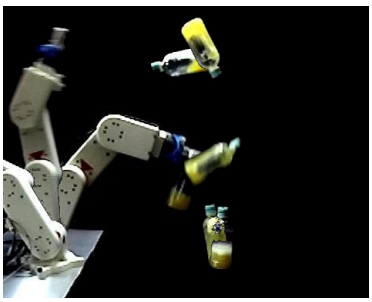
착지 순간만을 고려하는 보상 함수는 최종 목표인 물병을 지면에 수직하게 세운다는 것이 물병이 지면에 착지한 순간의 각도가 수직에 가까울수록 성공 가능성이 높다는 판단에 의해 설계되었다. 따라서 물병의 중간 거동을 평가하지 않고, [Fig. 10]과 같이 충돌시의 지면과 물병의 각도, θ만을 추출하여 식 (7)을 통해 보상 값을 도출한다. [Fig. 11]은 위와 같은 방법으로 구한 보상 값의 예시이다.

이러한 보상 함수는 지면과 병의 각도 자체를 보상 값으로 도출하여 간단하고 직관적이다. 하지만 미소한 점수차로 성공 여부가 갈릴 수 있어서 학습에 어떤 악영향을 미칠 수 있다.

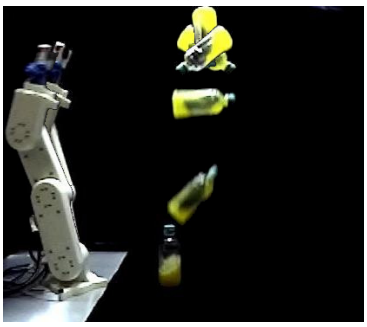
5. 강화 학습의 결과

5.1 강화 학습의 결과

이 절에선 4.1절과 4.2절에서 제시한 보상 함수에 대한 강화 학습 수행 결과를 기술한다.



(a)



(b)

[Fig. 12] Results of the bottle flipping task learning

[Fig. 12]의 (a)는 4.1절의 보상 함수를 이용한 학습 결과이다. 62번째에 물병을 지면에 세울 수 있었다. [Fig. 12]의 (b)는 4.2절의 보상 함수를 이용한 학습 결과로서 30번째에 목적을 달성했다.

보상 함수는 학습 성공을 어느정도 잘 표현할 수 있으나 주관적인 판단이 비교적 크게 반영되었기 때문에 보정 상수 등의 변화에 따라 매우 다른 학습 속도를 보일 수 있다. 본 논문에서는 $\alpha_1 = 100$, $\alpha_2 = 0.3$, $\alpha_3 = 1$, $\beta = 500$을 적용하였다.

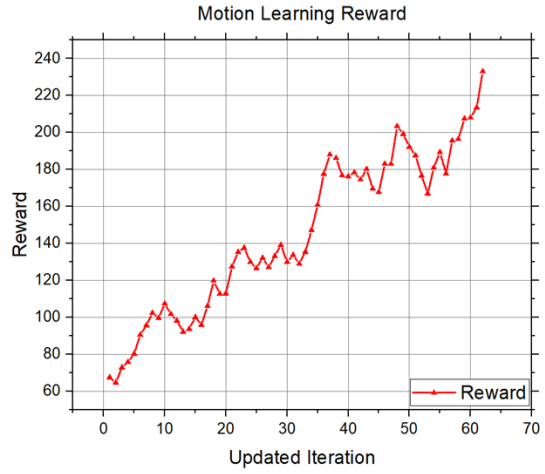
4.1절의 보상 함수의 중간 지점에 대한 평가가 복잡하다. 그래서 목표 성공과는 거리가 먼 어떠한 모션에 대해 높은 평가 점수를 도출할 수 있다. 그 결과 [Fig. 13]의 (a)와 같이 더 많은 시행횟수에도 보상 값이 수렴하지 못하였다.

4.2절의 보상 함수는 앞선 것보다 보상 값의 평가가 더욱 직관적이었다. 그 결과 좀더 빠르게 학습이 수렴되었다. 이는 목표 성공과 거리가 먼 어떠한 모션에도 낮은 점수를 도출하고 성공 가능성이 큰 경우 항상 높은 점수를 도출하기 때문이다.

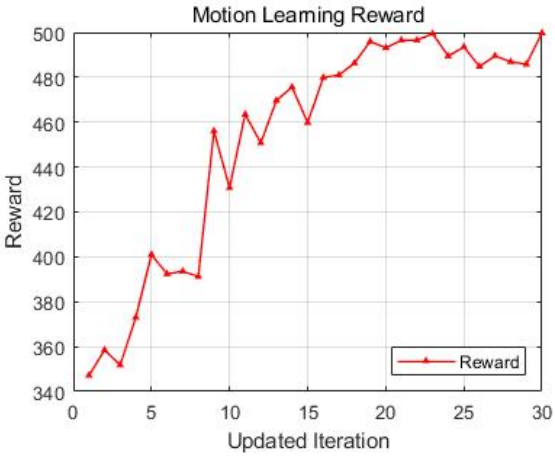
6. 결 론

본 논문은 보틀 플리핑에 대한 강화 학습을 수행하는 과정을 보상 함수의 설계에 중점을 두어 연구하였다. 강화 학습의 알고리즘으로는 경사법인 FDM을 사용했다.

보상 함수는 주관에 따라 적용하기가 달라 기존 연구들에서 크게 다루지 않았다. 학습할 임무가 복잡한 경우에는 어떤 경우에 높은 보상 값을 줄 것인지 판단하기 힘들기 때문에 주



(a)



(b)

[Fig. 13] Reward plot of the bottle flipping task learning

관이 개입 될 수밖에 없다. 따라서 본 논문에서는 첫번째로 병의 중간 거동을 고려여 주관적 판단이 들어간 보상 함수, 두번째로 병의 최종 순간만을 고려하여 직관적이고 단순한 두 가지 보상함수를 설계했고 그 결과를 분석했다. 학습 결과에 따르면, 보상 함수는 임무의 성공과 관련하여 직관적이고 간단한 보상 함수의 학습에서 수렴 속도를 단축시킨 효과가 있었다.

학습의 결과로 보틀 플리핑을 반복 수행했을때 성공률은 약 60%정도였다. 착지할 때 각도가 최대 20도 까지 차이가 났기 때문이다. 보틀 플리핑의 결과에 가장 큰 영향을 끼치는 요인은 그리퍼가 병을 놓는 순간의 병의 운동 상태다. 로봇과 물병에 다양한 요인들이 작용하는데 이를 완벽하게 통제하기는 어렵다. 이 때문에 [Fig. 12], [Fig. 13]의 학습 과정에서 보이는 것과 같이 보상값이 감소하는 방향으로 잘못 학습할 가능성도 존재한다.

로봇 강화 학습은 초기 동작과 보상 함수에 따라 학습 속도 및 학습 성공 여부에 극명한 결과가 나올 수 있으며, 특히 보상 함수는 학습 결과에 지대한 영향을 끼치지만, 여러 임무에

한번에 적용하거나 주관이 개입되지 않는 일반화된 보상함수를 구하기는 어렵다. 향후 로봇 강화 학습에서의 연구 방향은, 강화 학습뿐 아니라 인공신경망 등 여러 기계학습 알고리즘을 이용하여 이러한 보상 함수의 일반화를 이끌어 내야 한다.

References

[1] R. S. Sutton and A. G. Barto, "Introduction," *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2017, ch. 1, sec. 1-7, pp.1-18.

[2] J. Kober and J. Peters, "Learning Motor Primitives for Robotics," *2009 IEEE International Conference on Robotics and Automation*, Kobe, Japan, pp. 2112-2118, 2009.

[3] *Machine learning with applications to robotics*, [Online], http://lasa.epfl.ch/research_new/ML/index.php, Accessed: August 24, 2018

[4] Y. S. Liang, D. Pellier, H. Fiorino, and S. Pesty, "Evaluation of a Robot Programming Framework for Non-Experts using Symbolic Planning Representations," *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Lisbon, Portugal, pp. 1121-1126, 2017.

[5] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469-483, May, 2009.

[6] D. Hong, D. Lee, and J. Han, *DARwIn OP: Open Platform Humanoid Robot for Research and Education*, [Online], http://www.romela.org/darwin-op-open-platform-humanoid-robot-for-research-and-edu, Accessed: August 24, 2018.

[7] J. Kober and J. Peters, "Policy Search for Motor Primitives in Robotics," *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, 2009.

[8] P. Kormushev, S. Calinon, and D. G. Caldwell, "Robot Motor Skill Coordination with EM-based Reinforcement Learning," *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, pp. 3232-3237, 2010.

[9] P. Kormushev, S. Calinon, R. Saegusa, and G. Metta, "Learning the skill of archery by a humanoid robot iCub," *2010 10th IEEE-RAS International Conference on Humanoid Robots*, Nashville, TN, USA, pp. 417-423, 2010.

[10] M. Riedmiller, T. Gabel, R. Hafner, and S. Lange, "Reinforcement Learning for Robot Soccer," *Autonomous Robots*, vol. 27, no.1, pp. 55-73, July, 2009.

[11] J. Peters and S. Schaal, "Policy Gradient Methods for Robotics," *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Beijing, China, pp. 2219-2225, 2006.

[12] S. H. Lee, "Designing an efficient reward function for robot reinforcement learning of the water bottle flipping task," M.S thesis, Sogang University, Seoul, Korea, 2018.

양 영 하

2009 서강대학교 기계공학과(학사)

2018 서강대학교 기계공학과(석사과정)

관심분야: Reinforcement Learning

이 상 혁

2010 서강대학교 기계공학과(학사)

2016 서강대학교 기계공학과(석사)

관심분야: Reinforcement Learning

이 철 수

1990 KAIST 산업공학과 박사

현재 서강대학교 기계공학과 교수

관심분야: Reinforcement Learning