강화학습을 이용한 밸런싱 로봇 제어 시스템

김남훈*, 권영서*, 강산희*, 엄단경*, 박진현*, 송진우*
세종대학교 지능기전공학부*

Self-Balancing Robot Control System Using Reinforcement Learning

Nam Hoon Kim*, Yeong Seo Kwon*, San Hee Kang*, Eom Dan Gyeong*, Park Jin Hyeon*, Jin Woo Song*
School of Intelligent Mechatronics Engineering Sejong University*

Abstract - PID 제어 기법은 현대에도 널리 사용되는 제어기법이지만 비선형 모델 제어 시에 최적의 이득 값을 찾는 데 어려움이 있어 이득 값을 최적화하기 위한 다양한 방법론이 연구되고 있다. 이에 본 논문에서는 강화학습 모델인 DDPG를 이용하여 PD 제어기의 이득 값을 조절하는 적응형 PD 제어기를 설계하였다. 적응형 PD 제어기를 이용하여 비선형 밸런싱 로봇을 제어하고 시뮬레이션을 진행하여 기존 제어 시스템과의 비교를 통해 성능을 검증한다.

1. 서 론

PID 제어기(Proportional Integral Derivative Controller)는 적응, 퍼지, LQR 제어를 비롯하여 시스템 제어 분야에서 사용되는 대표적인 제어 기법 중 하나이다. PID 제어기는 비례, 적분, 미분 항으로 구성되어있다. PID 제어기가 최적화된 성능을 발휘하기 위해서는 이득 값의 세부적인 조정이 요구된다. 선형모델의 제어 이득 값을 조정하는 것은 비교적 쉽다. 하지만 비선형 모델이나 불확실성이 존재하는 모델은 시스템을 해석하기 힘들기 때문에 특성에 따른 적합한 이득 값을 찾기 어렵고, 찾는다 하더라도 제어성능을 유지하기 힘들 수 있다. 또한, 계수 조정이 잘되지 않는다면 시스템의 불안정성이 커지고 반응이 느려진다는 문제점이 있다. 이러한 문제를 해결하기 위해 PID 제어기의 이득 값을 최적화하는 방법이 많이 연구되었으며 최근에는 자동으로 조정하는 연구도 많이 이루어지고 있다.[1]

본 논문에서는 밸런싱 로봇과 같은 비선형 모델을 제어하기 위한 방법으로 강화학습을 이용한 적응형 PD 제어기를 제안한다. 강인한 제어를 위해 강화학습 알고리즘 DDPG (Deep Determinant Policy Gradient)[2]를 사용하여 현재 자세 오차에 따른 이득 값을 조절해 비선형 모델 제어의 문제점을 해결하며 PD 제어기와 성능을 비교하고자 한다.

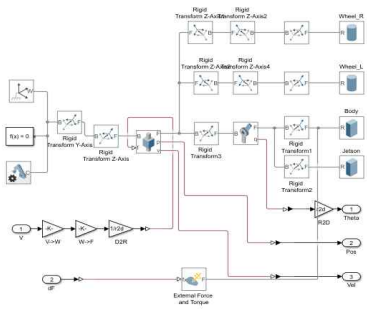
본 논문의 구성은 기존의 강화학습을 응용한 제어 기법들을 소개하고, PID 제어와 강화 학습 모델에 대한 간단한 개념을 설명한다. 이어 밸런싱 로봇의 모델, 강화학습 모델의 구성을 구체적으로 제시한다. 보다 직관적이고 모듈화된 작업환경과, 추후 임베디드 보드에 제어기를 탑재하여 실제 밸런싱 로봇 환경에서 실험 하기위해 MATLAB®을 이용하여 강화학습 모델과 3D 밸런싱 로봇 모델을 설계하고 모의실험하며, 제안하는 제어기와 기존의 PD 제어기의 비교하고 결론을 맺는다.

2. 밸런싱 로봇 제어 시스템

2.1 밸런싱 로봇 모델

밸런싱 로봇은 비선형 모델로 단일자유도를 가진 3D 밸런싱 로봇 모델을 매트랩을 사용하여 그림 1과 같이 구성하였다. 제어 입력에 따른 로봇의 기울기의 관계를 알기 위해 모델의 입력 값은 힘으로, 출력 값은 로봇의 피치 각(θ)으로 설정하였다. 밸런싱 로봇의 수학적 모델을 선형화하면 수식(1)과 같다. 이때, U는 수평방향의 힘, M,m은 각각 바퀴, 로봇 몸체 질량, b는 마찰계수, I는 로봇의 질량 관성 모멘트, l은 로봇의 무게중심까지의 높이이다.

$$\frac{\Theta(s)}{U(s)} = \frac{\frac{ml}{q}s}{s^3 + \frac{b(I+ml^2)}{q}s^2 + \frac{(M+m)mgl}{q}s - \frac{bmgl}{q}} \left[\frac{rad}{N}\right] \quad (1)$$



<그림 1> 밸런싱 로봇 모델

2.2 PID

PID제어는 오차 값 $e(t)$, 오차의 적분, 오차의 미분 값에 비례하는 제어 값 $u(t)$을 계산하는 제어 기법이다. PID제어는 구조가 간단하고 쉽게 적용이 가능하지만 비선형 모델을 제어 하기 위해서는 모델을 선형화해야 이득 값 K_p, K_i, K_d를 쉽게 찾을 수 있다. 따라서 밸런싱 로봇과 같은 비선형모델에서 이득 값을 찾기 어렵기 때문에 본 논문에서는 강화 학습을 이용하여 PD 제어기의 최적의 이득 값을 구한다. 시스템 특성상 적분오차가 시스템에 미치는 영향이 작으므로 적분 항은 제외하고 성능 비교를 위해 PD제어만을 수행하였다. PD 제어기의 수학적 모델링은 수식(2)과 같다.

$$u(t) = K_p e(t) + K_d \frac{de(t)}{dt} \quad (2)$$

3. 강화학습을 이용한 비선형 모델 PD 제어 이득 값 결정

3.1 강화 학습

강화 학습의 기본적인 구성 요소를 살펴보면, 크게 환경(environment), 에이전트(agent), 정책(policy), 관찰 값(observation), 동작(action), 보상(reward)으로 구성되어 있다. 환경에서 에이전트는 Q-function을 이용하여 현재 관찰 값을 고려하였을 때 가장 적절한 정책에 따라 동작을 취한다. 그 결과로 관찰 값이 변화하며 보상을 반환한다. 모델은 높은 보상을 받을 수 있는 가장 최적화된 정책을 수립하기 위해 학습을 반복한다.

3.2 PD 제어 이득 값 결정 기법

환경은 제어하고자 하는 3D 밸런싱 로봇이며 로봇으로부터 나오는 현재 자세정보와 자세의 오차, 위치를 관찰 값으로 설정하여 최적의 PD 제어기 이득 값을 구한다. 각도 오차에 대한 관찰 값은 오차의 $\sin, \cos$값으로 구성되며 범위는 [-1, 1]이다.

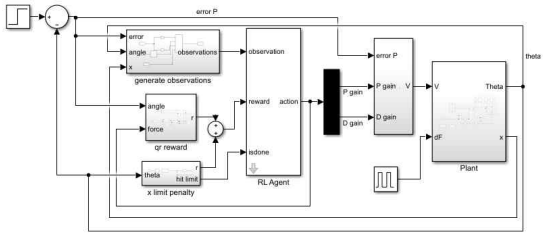
각도 θ는 $[-\pi, \pi]$, 위치 x값은 (-inf, inf)의 범위를 가진다. 로봇의 균형을 잡기위한 보상은 로봇의 자세와 PD 제어기 이득 값을 기준으로 판단하였다. 로봇이 넘어지지 않도록 하는 것이 가장 중요하므로, 로봇의 피치 각이 10도보다 클 경우 큰 패널티를 주었다. 10도 보다 작을 경우, 0도에 가까울수록 더 큰 양의 보상을 주어 자세의 오차를 줄여나갈 수 있게 보상을 설정하였다. 또한, 자세 균형을 맞추기 위해 최소한의 제어를 수행하도록 이득 값의 크기에 비례하는 음의 보상 값을 주었다. 보상 값에 따라 학습 결과가 매우 달라지고 시스템 응답을 결정하는 데 큰 영향을 미치므로 적절한 보상 값을 결정하는 것이 중요하다. 학습 회차(episode)를 종료 조건으로는 로봇의 피치 각이 10도보다 큰 경우 복원력을 상실한 것으로 간주하여 종료하게 설계하였다.

에이전트는 주어진 관찰 값을 기반으로 적절한 동작을 취해 환경과 상호작용하여 상태를 변화시킨다. 알고리즘은 DDPG를 이용한다. DDPG는 DPG에 DQN의 심층 신경망 등의 성공적인 아이디어를 접목한 Model-Free, off-policy, actor-critic 알고리즘 모델이다. 기존의 discrete space에서 발생하는 데이터의 손실을 막기 위해 continuous space를 구성하고, 심층 신경망을 적용하여 연산량 문제를 해결하였다. 따라서 연속적인 제어가 필요한 현실 세계의 로봇을 대상으로 더욱 정밀한 학습을 기대할 수 있다. 이에 본 논문에서는 DDPG를 이용하여 제어 모델을 설계한다. DDPG는 Actor 네트워크와 Critic 네트워크로 구성되어 있으며, 전자는 정책 함수를 학습하며, 후자는 해당 정책의 가치(Value)를 평가한다. 이를 통해 에이전트가 최적의 정책을 수립하고 동작할 수 있도록 한다. 에이전트의 동작으로 K_p, K_d 두 개의 값이 출력되며, [-5e+03, 5e+03]의 범위 내의 이득 값을 선택한다.

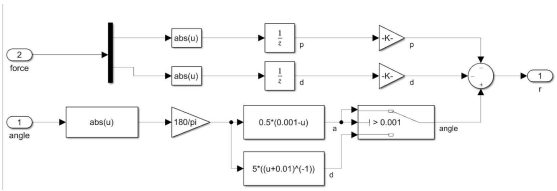
4. 실 험

4.1 강화학습 모델

강화학습을 통한 최적의 이득 값 결정하기 위해 MATLAB®을 이용하여 아래와 같이 강화학습 모델과 보상함수를 구성하였다. 실험에서 초기 자세는 0도로 설정하였고 학습을 위해 0.17N/s의 외란을 0.4초 동안 밸런싱 로봇에 가했으며 20초 동안의 시스템 응답을 기반으로 학습하였다.



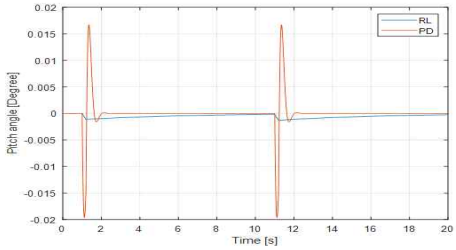
〈그림 2〉 전체 시스템 구성도



〈그림 3〉 리워드 구성도

4.2 강화학습 결과를 이용한 밸런싱 로봇 제어

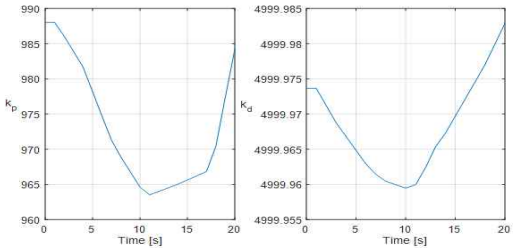
강화학습을 이용한 적응형 PD 제어의 성능을 검증하기 위해 PD 제어기와 비교분석하였다. 이 때 적응 PD 이득은 그림 5와 같다. 0.21N/s의 외란이 0.2초 동안 가해졌을 때 K_p=68, K_d=860인 PD 제어기와 비교 실험을 수행하였다. 아래 <그림 4>와 같은 시스템 응답과 표 1을 통해 적응형 PD 제어기의 RMS 오차가 작고 외란에도 강인하게 제어됨을 확인하였다. 자세가 수렴한 이후에 PD 제어 이득 값이 커지지만 자세오차가 작아 로봇이 안정적으로 제어된다.



〈그림 4〉 제어에 따른 시스템 응답

〈표 1〉 제어기에 따른 RMS 오차

	강화학습을 이용한 PD 제어기	PD 제어기
RMS 오차	1.1299e-05	8.1693e-05



〈그림 5〉 이득 값 K_p, K_d

5. 결 론

본 논문에서는 강화학습을 이용하여 비선형 밸런싱 로봇을 제어하기 위해 변하는 동특성에 따라 PD 제어 이득 값을 조정을 하는 학습모델을 설계하였다. 강화학습을 이용한 PD제어기에 외란을 주어 실험을 하였고, PD 제어기와 비교했을 때 오버슈트가 크지 않고 RMS오차가 낮아 비선형 모델인 밸런싱 로봇이 안정적으로 제어가 되는 것을 확인하였다.

본 연구를 바탕으로 단일 자유도를 가진 비선형 모델에서 확장하여 다 자유도 모델의 제어 이득 값을 강화학습으로 구할 수 있으며, 3축에서 이동하는 비선형모델을 PID제어 기법을 이용해 제어할 때 시간에 따른 최적의 제어 이득 값으로 조정할 수 있도록 활용될 수 있다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2020-2018-0-01423)

김남훈, 권영서, 강산희는 공동 1저자로서 연구에 참여하였음.

[참 고 문 헌]

[1] Sun, Q., Du, C., Duan, Y. et al. Design and application of adaptive PID controller based on asynchronous advantage actor–critic learning method. Wireless Netw (2019).

[2] Lillicrap, Timothy P., et al. "Continuous control with deep reinforcement learning." arXiv preprint arXiv:1509.02971 (2015).