상태 표현 방식에 따른 심층 강화 학습 기반 캐릭터 제어기의 학습 성능 비교*

손채준[0], 이윤상
한양대학교
{thscowns, yoonsanglee}@hanyang.ac.kr

Comparison of learning performance of character controller based on deep reinforcement learning according to state representation

Chaejun Sohn, Yoonsang Lee
Hanyang University

요약

물리 시뮬레이션 기반의 캐릭터 동작 제어 문제를 강화학습을 이용하여 해결해 나가는 연구들이 계속해서 진행되고 있다. 강화학습을 사용하여 문제를 풀기 위해서는 네트워크 구조, 하이퍼파라미터 튜닝, 상태(state), 행동(action), 보상(reward)이 문제에 맞게 적절히 설정이 되어야 한다. 많은 연구에서 다양한 조합으로 상태, 행동, 보상을 정의하였고, 성공적으로 문제에 적용하였다. 상태, 행동, 보상을 정의함에 다양한 조합이 있다 보니 학습 성능을 향상시키는 최적의 조합을 찾기 위해서 각각의 요소들이 미치는 영향을 분석하는 연구도 진행되고 있다. 우리는 지금까지 이뤄지지 않았던 상태 표현 방식에 따른 강화학습에 미치는 영향을 분석하였다. 첫째로, root attached frame, root aligned frame, projected aligned frame 3가지로 좌표계를 정의하였고, 이에 대해 표현된 상태를 이용하여 강화학습에 미치는 영향을 분석하였다. 둘째로, 상태를 정의할 때, 관절의 위치, 각도로 다양하게 조합하는 경우에 학습성능에 어떠한 영향을 미치는지 분석하였다.

1. 서론

물리 시뮬레이션 기반의 사람 캐릭터 동작 제어는 상당히 어려운 문제로 여겨져 왔다. 그 이유 중 하나로 사람 캐릭터가 많은 관절을 가지고 있어 높은 자유도(degrees of freedom, DOFs)를 가지기 때문에 상태(state)의 차원이 크다는 것을 들 수 있다. 이러한 어려움을 효과적으로 극복할 수 있는 방법으로서, 최근에는 심층신경망 정책을 기반으로 하는 심층강화학습을 캐릭터 동작 제어 문제에 적용하는 연구들이 많이 발표되고 있다. 강화학습을 사용하여 문제를 풀기 위해서는 네트워크 구조, 하이퍼파라미터 튜닝(hyperparameter tuning), 상태, 행동(action), 보상(reward)이 문제에 맞게 적절히 설정이 되어야 사람 캐릭터가 걷거나 뛰는 동작을 만들어 낼 수 있다.

강화학습이 다루는 문제의 기본적인 형태를 살펴보자면 매 스텝마다 환경은 상태와 보상을 제공하고 에이전트(agent)는 이 상태에 근거하여 최적의 행동을 취한다. 강화학습의 목표는 가장 높은 누적보상을 갖게 하는 행동을 출력하는 에이전트 정책(policy)을 찾는 것이다. 캐릭터 제어 문제에서 상태는 일반적으로 관절 위치(joint position) 또는 관절 각도(joint angle), 속도, 위상(phase) 등을 이용하여 표현된다. 행동은 PD제어(PD control)를 위한 관절 각도 또는 토크(torque)가 이용되며, 보상은 보통 참조 동작(reference motion)과 시뮬레이션된 캐릭터의 차이와 에너지 최소화 항 등으로 구성된다. 이러한 상태, 행동, 보상은 각 연구마다 조금씩 다른 조합으로 정의되고 있다. 예를 들어 [1]에서는 관절 위치, 관절 속도, 각속도, 위상 정보들을 이용해서 상태를 정의하여 캐릭터 동작제어 문제를 풀었고, [2]에서는 관절 각도, 관절 속도, 말단부(end-effector) 위치 정보 및 미래의 참조 동작을 상태로 정의하여 문제를 성공적으로 풀었다. 이렇듯 다양한 조합으로 문제가 정의될 수 있다 보니 캐릭터 동작 제어 문제를 강화학습을 이용하여 풀 때 학습성능에 영향을 미치는 요소들을 알아볼 필요성이 있었다. 이에 [3]에서는 강화학습을 통해 학습시킨 정책의 성능에 행동의 설계(design)가 미치는 영향에 대해 분석하였고, [4]에서는 상태, 행동, 보상의 설계가 학습에 미치는 영향을 분석하였다.

상태의 경우, 같은 속성을 상태로 정의하더라도 어떤 좌표계에 대해 표현된 상태 값인지에 따라 그 특성이 달라질 수 있다. 예를 들면 평면에 투영된 좌표계에 의해 표현된 상태 값은 높이에 대한 정보가 포함되므로, 투영되지 않은 좌표계에 의해 표현된 상태와 특성이 다를 수 있다. 그러나 기존 연구들은 상태를 표현하기 위한 좌표계의 선택에 집중하여 분석하지는 않았다. 이 연구

* 구두 발표논문, 요약 논문(Extended Abstract)
* 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2019R1C1C1006778, NRF 2019R1A4A1029800).

에서는 다양한 좌표계에서 표현된 상태가 강화학습을 통한 문제 해결에 어떠한 영향을 미치는지 알아볼 것이다. 또한, 캐릭터의 상태를 나타내기 위해 관절 위치 혹은 관절 각도가 사용될 수 있다. 이에 [4]에서 상태를 관절 위치, 각도를 이용하여 정의할 때 학습에 미치는 영향을 보여주었다. 그러나 이는 두가지 모두를 사용한 것과 각도만 사용하였을 경우 학습성능을 비교하였고, 참조 동작을 따라하도록 하는 문제에 대해서는 학습 성능을 비교하지 않았다. 따라서 우리는 참조동작을 따라하도록 하는 캐릭터 제어문제를 강화학습을 이용하여 풀 때, 관절의 위치와 각도를 다양하게 조합하는 경우의 학습 성능을 분석할 것이다. 이를 통해 우리는 물리 시뮬레이션 기반의 사람 캐릭터 동작 제어 문제를 강화학습을 이용하여 풀 때, 상태를 정의하는 가장 좋은 방법 및 상태를 표현하는 좌표계의 선택에 중요한 요소가 무엇인지 보이고자 한다.

2. 관련연구

물리 시뮬레이션 기반의 캐릭터 제어 문제에 강화학습을 적용하는 연구는 계속해서 늘어나고 있다. 효과적으로 강화학습을 이 문제에 적용하기 위해서는 상태, 행동, 보상을 적절히 설계해야 한다.

상태는 크게 캐릭터의 동작을 나타내는 정보인 동역학적 상태(dynamic state), 캐릭터가 따라 가야할 동작에 대한 정보인 추적 정보(tracking information), 캐릭터가 움직이는 방향, 속도 등을 설정하기 위한 사용자 지정 정보(user-specified goal), 추가 정보(additional information)로 나눌 수 있다. 첫째로, 동역학적 상태를 정의하기 위해 [1]에서는 관절 위치, 관절 각도, 속도, 각속도를 이용하여 정의하였다. [2] 에서는 관절 각도, 속도, 말단부(end effector) 정보를 이용하여 정의하였다. [5] 에서는 시뮬레이션 캐릭터의 속도, 무게중심의 속도를 이용하였다. [6] 에서는 속도, 상체의 기울기(up-vector of the trunk), 말단부 위치, root의 높이가 사용되었다. 둘째로, 추적 정보를 [1]에서는 위상 정보를 이용하였고, [2]에서는 현재, 미래의 동작을 이용하여 추적정보를 구성하였다. [5]에서는 참조 동작의 위치, 속도와 시뮬레이션 되는 캐릭터와의 위치, 속도의 차이를 이용하였다. 사용자 지정 정보는 어떠한 일을 수행하는가에 따라 다르게 구성되기도 한다. 예를 들어, [1]에서 사용된 목표 방향(target direction), 무게중심의 속도가 사용되기도 하고, [2], [5]처럼 root의 목표(desired) 속도 정보를 이용하여 문제를 풀기도 하였다. 그 외의 추가 정보로는 [2]에서 사용한 캐릭터 크기 변수(body shape parameter), [5]에서 이전 스텝의 에이전트가 선택한 행동을 이용한 경우가 있었다.

행동으로는 일반적으로 PD 제어 목표 자세, 토크, 근육 활성화 정도 등이 사용된다. [1]에서는 목표 자세를 사용하여 행동을 정의하였다. [2], [5], [6]에서는 목표 자세를 구성하기 위한 참조동작으로부터의 차이(offset)를 이용하여 정의하고 있다. [7] 에서는 근육을 통해 움직이는 캐릭터를 모델링하였고, 근육의 활성정도를 행동으로 이용하여 캐릭터의 움직임을 생성하였다.

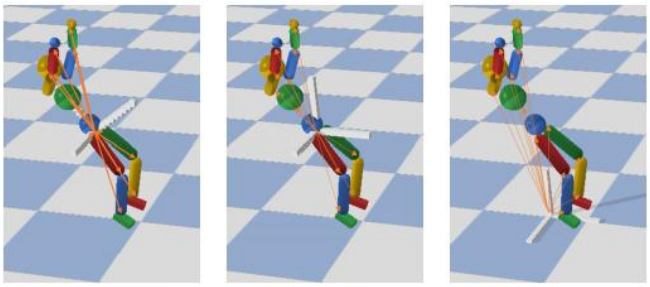
보상의 구성요소는 크게 3가지로 나눌 수 있었다. 참조동작과 캐릭터가 비슷하도록 하는 모방 항(imitation term), 특정 일을 수행하도록 하는 수행 항(task term), 최소한의 에너지를 사용하도록 하는 에너지 항(energy term)으로 구성할 수 있다. 뿐만 아니라, 이러한 항들을 어떻게 조합할지에 대한 방법들이 제안되어왔다. [1]와 [2]에서는 모방항을 정의할 때, 참조동작과 캐릭터의 관절 각도, 관절의 속도, 말단부 위치를 이용하였다. 에너지 항은 [2]에서 사용되었고, 가능하면 최소한의 에너지를 사용하도록 캐릭터를 제어할 수 있게 한다. 마지막으로 이렇게 정의된 항들을 조합하여 최종보상으로 구성하는 방법에는 [1],[5],[2]에서 사용된 항들을 더하는 방식과 [6]에서 사용된 모든 항들을 곱하는 방식이 존재한다. [6]에서는 최종 보상을 곱셈으로 정의하는 것이 모든 항에서 높은 보상을 가져야 높은 누적보상을 가질 수 있기 때문에 더하는 방식보다 더 좋은 성능을 가질 수 있다고 한다. 이렇게 물리 시뮬레이션 기반의 사람 캐릭터 동작 제어 문제를 강화학습을 이용하여 풀기 위해 고려해야할 사항들이 많이 존재한다.

많은 연구에서 강화학습을 이용하게 되면서 강화학습의 성능에 영향을 미치는 요소들의 적용 여부에 따른 결과를 비교 분석하는 연구들도 많이 수행되었다. [8]에서는 강화학습에 영향을 미치는 요소들을 코드수준 최적화기법(code-level optimization)이라고 정의하고 알고리즘별로 이러한 최적화 요소의 존재 여부에 따른 성능을 비교 분석한다. 또한, 문제의 환경에 따른 영향을 분석한 연구들도 존재하였다. [3] 연구에서는 행동을 어떻게 정의하는 지에 따른 강화학습의 성능을 분석하고 있다. 또, [4]에서는 환경을 정의하는 방식과 그 외의 다른 요소(episode termination, control frequency)에 따른 학습된 정책의 성능을 분석하였다. 환경의 상태를 다양하게 정의해서 분석하는 연구들이 존재했지만, 상태의 표현방식에 따른 성능을 분석하는 연구는 부족했고, 이 논문에서 상태의 표현방식에 따른 학습된 정책의 성능을 분석할 것이다.

3. 방법

우리는 물리 시뮬레이션 기반의 캐릭터를 참조동작을 따라하도록 제어하는 문제를 강화학습을 이용하여 풀 때, 상태를 여러가지 다른 방식으로 정의하고 이것이 학습에 어떠한 영향을 미치는지 확인하였다. 이를 위해 우리는 [1]에서 제안한 강화학습 기반 캐릭터 제어 프레임워크의 구조를 사용하였다. 우리는 [1]에서 정의된 것처럼 관절 위치, 관절 각도, 속도, 각속도, 위상, root의 높이를 상태로 사용하였다. 첫째로, 우리는 상태가 서로 다른 좌표계에 대해 표현되었을 경우에 학습 성능을 비교하였다. 둘째로, 상태로 관절 위치와 각도의 다양한 조합을 사용하는 경우의 학습 성능을 비교하였다. 3.1 에서는 상태를 정의하기 위한 3가지 다른 좌표계와 좌표계에 대해 표현되는 상태를 설명할 것이다. 3.2에서는 관절의 위치 및 관절 각도가 상태에 포함되는지 여

부에 따른 실험을 설명할 것이다.



(a) Attached Frame (b) Aligned Frame (c) Projected Frame

Figure 1: The three reference frames used in the experiments. The orange arrows depict joint positions represented in each frame.

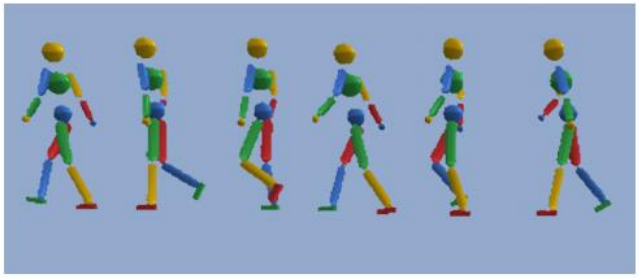
3.1. 상태를 나타내는 좌표계

어떤 좌표계에 대해 표현된 상태 값인지에 따라 특성이 다르기 때문에 우리는 총 3가지의 좌표계를 이용하여 상태를 다르게 표현하여 사용하였다. Figure 1에 표현한 것처럼, (1) Root attached frame, (2) Root aligned frame, (3) Projected root aligned frame 세 가지 좌표계를 이용하였다.

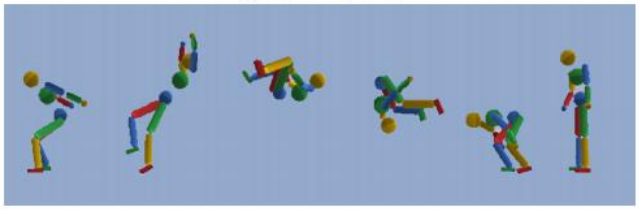
Root attached frame은 root의 로컬 좌표계를 의미한다. Root aligned frame은 상체가 바라보는 방향과 y축 및 이 두가지를 외적 한 것을 각 축으로 가지는 좌표계를 의미하고 위치는 root의 위치로 정의한다. Projected root aligned frame은 root aligned frame의 원점을 xz 평면에 projection 시킨 좌표계이다. Figure 1에서처럼 우리는 관절 위치 뿐 아니라 관절 각도, 속도, 각속도를 각각의 좌표계에 대해서 표현하여 상태를 정의하였다. 우리는 앞서 설명한 총 3가지의 좌표계를 이용하여 관절 위치, 관절 각도, 속도, 각속도를 각각의 좌표계에 대해서 표현하여 상태를 정의하였다.

3.2. 관절의 위치 혹은 각도의 조합

캐릭터의 위치 상태를 나타내기 위해 관절 위치, 관절 각도 중 하나만 알아도 된다. 이전 연구들에서 상태를 정의할 때 관절 위치, 관절 각도를 모두 사용한 연구[1]도 존재하였고, 둘 중 하나만 사용한 연구[2]도 존재했다. 이러한 경우들 간의 성능을 비교하는 연구[4]도 존재하였으나 관절 위치, 각도 모두 사용한 것과 관절 각도를 사용한 것의 비교였고, 참조 동작을 추적하는 문제가 아닌 사람이 걷는 문제를 강화학습을 이용해 학습하는 경우 성능을 비교하였다. 우리는 관절 위치만 사용한 것, 관절 각도만 사용한 것, 관절 위치, 각도 모두 사용한 것 세가지로 나누어서 참조 동작을 추적하는 문제의 강화학습을 진행하여 결과를 비교했다. 이 실험에 대해서 모든 상태는 root aligned frame에 의해 표현된다.

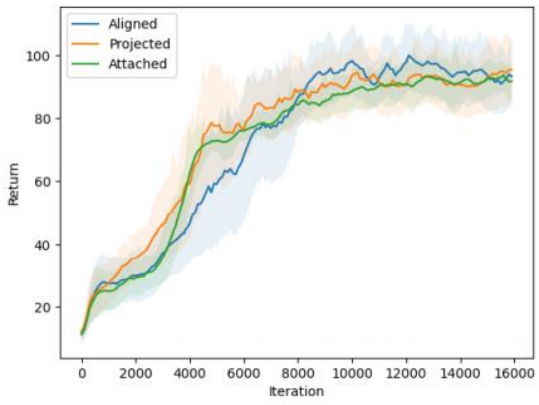


(a) walk sequence

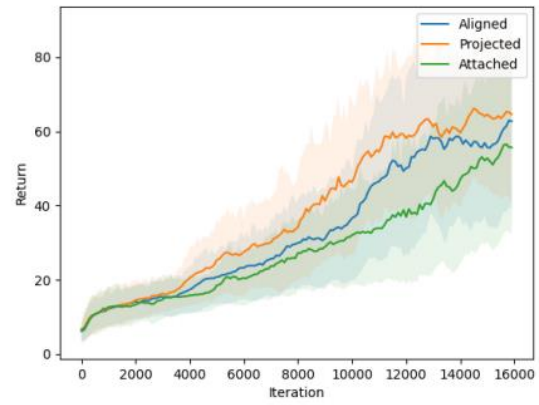


(b) backflip sequence

Figure 2: Reference motion sequence for each environment



(a) WalkEnv



(b) BackflipEnv

Figure 3: This is a learning graph when the state is represented by 3 frames(root aligned, projected root aligned, root attached).

4. 결과

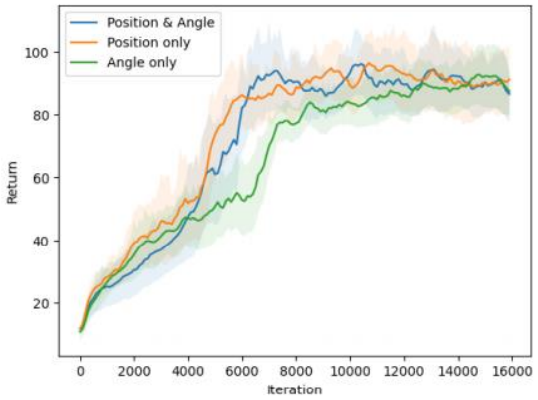
시뮬레이션은 bullet physics[9]를 이용하였다. 우리는 pybullet에 정의되어 있는 deepmimic 환경을 이용하였다. 기본적으로 [1]에서 사용된 것처럼 aligned frame에 대해 표현된 상태를 사용하였고, PPO를 이용하여 학습을 진행하였다. hyperparameter는 [1]에서 사용한 값을 그대로 사용하였고, HumanoidDeepMimicWalkBulletEnv(이하 WalkEnv)와 HumanoidDeepMimicBackflipBulletEnv(이하 BackflipEnv) 환경을 학습시켰다. 각 환경마다 16000 iteration(65,000,000 sample)씩 학습을 진행하였고, 100 iteration 마다 네트워크 파라미터를 저장하였다. 저장된 네트워크마다 50번씩 에피소드를 시뮬레이션 하여 누적보상의 평균과 표준편차를 계산하였고, 이를 이용하여 그래프를 만들었다(Figure 3,4). 그래프에서 실선은 누적보상의 평균을 의미하고, 음영은 각 시점의 평균에 해당 시점의 표준편차를 더한 값과 뺀 값 사이의 영역으로 각 시점에서 나타날 수 있는 누적보상의 범위를 의미한다.

4.1. 상태를 나타내는 좌표계에 따른 성능 비교

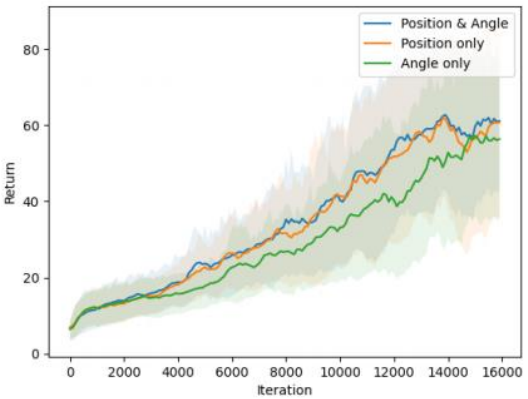
WalkEnv, BackflipEnv에 대해 상태를 각각 attached frame, aligned frame, projected aligned frame에 대해 표현하여 학습을 진행하였다. Figure 3에서 알 수 있듯이 WalkEnv에 대해서는 상태를 각 프레임에 따라 표현하는 것이 크게 의미가 있지는 않았다. 그러나, BackflipEnv에서는 상태를 각각 다른 좌표계에 대해서 표현하는 것이 큰 영향을 미치는 것을 확인할 수 있었다. 이유에 대해 분석하기에 앞서, 학습한 환경들의 참조동작(Figure 2)을 살펴보면 WalkEnv의 경우 root의 높이는 거의 변하지 않고, 회전이 큰 동작이 없이 앞으로 나아가기만 한다. 반면 BackflipEnv의 경우 root의 높이가 계속해서 변하고, 회전이 큰 동작이다. 이러한 환경의 특성 차이로 인해 Figure 3a와 3b에서 나타난 경향성에 차이가 난다.

Figure 3b를 보면 BackflipEnv를 학습시킬 때 projected frame에 의해 표현된 상태를 이용하여 학습시킨 결과가 다른 좌표계를 이용하였을 경우보다 더 좋은 성능을 보여준다. projected frame에 의해 표현되는 관절의 위치 정보에는 나머지 좌표계에 의해 표현되는 상태에 비해 root의 높이가 더해져서 상태로 정의된다. 그렇기에 root의 높이 정보가 계속해서 달라지는 BackflipEnv의 경우 projected frame에 대해 상태를 표현하였을 때, 캐릭터의 동역학적 상태를 가장 잘 표현하고, 결과적으로 학습도 가장 잘 되는 것으로 볼 수 있다. 또한 Figure 3b에서 보면 상태를 aligned frame에 대해 표현하는 경우가 attached frame에 대해 표현하는 경우보다 학습성능에 미치는 영향이 큰 것을 알 수 있다. 이는 aligned frame 기준으로 표현된 상태가 attached frame에 의해 표현된 상태보다 회전에 대한 정보를 더 잘 표현하기 때문으로 볼 수 있다. 예를 들면, 점프의 중간단계에서 상태를 aligned frame에 대해 표현한다면 aligned frame에서 y축은 고정이 되어있기 때문에 몸이 기울어진 정도를 바로 알 수 있지만, attached frame은 캐릭터가 회전을 하면 좌표계도 같이 회전을 하게 되고 따라서 attached frame에 대해 표현된 상태는 그만큼 회전에 대한 정보가 부족하다고 할 수 있다. 이러한 이유로 aligned frame에 의해 정의된 상태를 이용하여 학습시킨 결과가 더 좋은 성능을 보여주는 것으로 추측된다.

이를 통해 최적의 학습을 위한 상태 정의를 위한 좌표계의 선택은 참조동작이 가지는 특성에 따라 달라져야 한다는 것을 알 수 있다. 즉, 높이의 변화가 큰 부분이 많은 참조 동작에 대해서는 projected frame을 사용하는 것이 유리하고, 몸 전체의 회전 동작이 많은 참조 동작에 대해서는 projected frame 혹은 aligned frame을 사용하는 것이 유리하다고 볼 수 있다.



(a) WalkEnv



(b) BackflipEnv

Figure 4: This is a learning graph when the dynamic state is defined as Position only or Angle only or Position and Angle

4.2. 관절의 위치 혹은 각도의 조합에 따른 성능 비교

WalkEnv, BackflipEnv에 대해 각각 관절 위치만 사용한 경우, 관절 각도만 사용한 경우, 관절 위치, 각도 모두 사용한 경우로 상태를 정의하여 학습을 진행하였다. Figure 4에서와 같이, WalkEnv와 BackflipEnv에서 모

두 최종적으로는 비슷한 수준의 누적보상을 보여주었다. 두 환경 모두 위치 정보를 사용하여 학습한 경우가 각도만 사용한 경우보다 빠르게 최종수준에 도달하였다. 이를 통해 환경의 참조 동작이 갖는 특성, 즉 동작의 회전과 root 높이 변화에 상관없이 캐릭터 관절의 위치정보가 학습에 더 큰 영향을 주는 것을 알 수 있었다. 강화학습에서 에이전트가 캐릭터의 상태를 학습하여 가장 높은 누적보상을 갖도록 행동을 만들어내는데, 이 때 캐릭터의 위치정보가 에이전트가 학습을 더 쉽게 할 수 있도록 만드는 것으로 여겨진다

5. 결론

물리 시뮬레이션 기반의 캐릭터 제어 문제에 강화학습을 적용하는 시도들이 계속해서 이루어지고 있다. 최근에는 참조동작을 따라하도록 환경을 구성하고 이를 강화학습을 이용하여 푸는 연구들이 주를 이룬다. 이에 우리는 이러한 문제 정의속에서 상태를 표현하는 방식이 강화학습 에이전트의 학습에 미치는 영향에 대해서 알아보았다. 실험 결과로부터 참조동작이 가지는 특성에 따라서 상태를 정의하는 좌표계를 선택해야 학습 속도 및 성능을 높일 수 있다는 것을 알 수 있다. 또한, 상태를 정의함에 관절의 위치 정보와 각도가 학습하는 데에 미치는 영향을 알아보았다. 이 연구를 통해 상태 표현 방식이 강화학습에 미치는 영향을 알아볼 수 있었다.
이 연구에서는 두 가지 참조동작만을 이용하여 실험을 진행하였으나, 향후 더 많은 참조동작들에 대해 분석을 수행할 계획이다. 또한, 이 연구에서는 학습 속도 측면에서 가장 유리한 참조동작 별 상태 표현 좌표계의 선택에 대해 정성적인 분석만을 수행하였으나, 향후 이를 정량적으로 분석할 수 있는 방법에 대한 연구를 진행할 계획이다.

참고문헌

[1] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, “Deepmimic,” ACM Transactions on Graphics, vol. 37, no. 4, p. 1–14, Aug 2018. [Online]. Available:
[2] J. Won and J. Lee, “Learning body shape variation in physics-based characters,” ACM Trans. Graph., vol. 38, no. 6, Nov. 2019. [Online]. Available: https://doi.org/10.1145/3355089.3356499

[3] X. B. Peng and M. van de Panne, “Learning locomotion skills using deeprl,” Proceedings of the ACM SIGGRAPH / Euro graphics Symposium on Computer Animation, Jul 2017. [Online]. Available: http://dx.doi.org/10.1145/3099564.3099567

[4] D. Reda, T. Tao, and M. van de Panne, “Learning to locomote: Understanding how environment de sign matters for deep reinforcement learning,” Motion,
Interaction and Games, Oct 2020. [Online]. Available: http://dx.doi.org/10.1145/3424636.3426907

[5] K. Bergamin, S. Clavet, D. Holden, and J. R. Forbes, “Drecon: Data-driven responsive control of physics-based characters,” ACM Trans. Graph., vol. 38, no. 6, Nov. 2019. [Online]. Available: https://doi.org/10.1145/3355089.3356536

[6] S. Park, H. Ryu, S. Lee, S. Lee, and J. Lee, “Learning predict-and-simulate policies from unorganized human motion data,” ACM Trans. Graph., vol. 38, no. 6, 2019.

[7] S. Lee, M. Park, K. Lee, and J. Lee, “Scalable muscle-actuated human simulation and control,” ACM Trans. Graph., vol. 38, no. 4, July 2019. [Online]. Available: https://doi.org/10.1145/3306346.3322972

[8] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry, “Implementation matters in deep policy gradients: A case study on ppo and trpo,” 2020.

[9] “Bullet physics library,” https://pybullet.org/, 2015