강화 학습을 사용한 동적 게임 환경에서의 빠른 경로 탐색

이승준° 장병탁
서울대학교 바이오지능 연구실
{sjlee°, btzhang}@bi.snu.ac.kr

Fast Navigation in Dynamic 3D Game Environment Using Reinforcement Learning

Seung Joon Yi° Byoung Tak Zhang
Biointelligence Lab, Seoul National University

요 약

연속적이고 동적인 실세계에서의 경로 탐색 문제는 이동 로봇 분야에서 주된 문제 중 하나였다. 최근 컴퓨터 성능이 크게 발전하면서 컴퓨터 게임들이 실제에 가까운 연속적인 3차원 환경 모델을 사용하기 시작하였고, 그에 따라 보다 복잡하고 동적인 환경 모델 하에서 경로 탐색을 할 수 있는 능력이 요구되고 있다. 강화 학습 기반의 경로 탐색 알고리즘인 평가치 반복(Value iteration) 알고리즘은 실시간 멀티에이전트 환경에 적합한 여러 장점들을 가지고 있으나, 문제가 커질수록 속도가 크게 느려진다는 단점을 가지고 있다. 본 논문에서는 연속적인 3차원 상황에서 빠르게 동적 변화에 적응할 수 있도록 하기 위하여 작은 세상 네트웍 모델을 사용한 환경 모델 및 경로 탐색 알고리즘을 제안한다.

3차원 게임 환경에서의 실험을 통해 제안된 알고리즘이 연속적이고 복잡한 실시간 환경 하에서 우수한 경로를 찾아낼 수 있으며, 환경의 변화가 관측될 경우 이에 빠르게 적응할 수 있음을 확인할 수 있었다.

1. 서 론

경로 탐색 문제는 AI, 이동 로봇, 가상 현실, 에이전트 시뮬레이션 등 다양한 분야에서 중요한 문제이며, 특히 게임 AI 분야에서 빈번하게 요구되는 문제 중 하나이다 [1]. 과거의 단순한 형태의 게임의 경우 2차원의 격자 상에서 등장 인물들이 이동하며 단순한 경로 탐색 알고리즘을 사용하여 움직였으나, 넓고 복잡한 연속된 3D 환경을 갖는 최근의 게임의 경우 실세계에서 사용되는 것과 같은 높은 수준의 경로 탐색 알고리즘이 요구된다. 특히, 최근의 경우 현실감을 높이기 위해 이동 가능한 많은 장애물들이 환경에 포함되고 있는 추세인데, 이를 위해서는 동적인 환경에 대응할 것이 요구된다.

동적인 환경에서의 경로 탐색을 위해 적용될 수 있는 알고리즘으로는 강화 학습 (Reinforcement Learning) 이 있다. 강화 학습 에서는 학습의 주체인 에이전트(Agent)는 환경(World)과 상호작용하며 최대의 보상(Reward)를 주는 상태(State)와 행동(Action)의 함수인 정책(Policy)를 학습하려 한다. 전통적인 RL 프레임웍에서는 환경을 이산적인 시간과 공간으로 이루어진 마르코프 결정 프로세스 (Markov Decision Process)으로 정의하고, Q-Learning 과 같은 강화 학습 알고리즘에서는 상태와 행동 공간을 테이블의 형태로 가정하고 모든 상태-행동의 평가치(Value function)을 구해서 최적의 정책을 결정하게 된다.

이러한 강화 학습 알고리즘은 확률적으로 최적의 경로를 찾을 수 있으며, 환경이 변화하는 경우에도 새로운 환경에 적응이 가능하며 당장 근사 경로를 출력할 수 있는 온라인 알고리즘의 특성을 가진다. 하지만 이러한 강화 학습 알고리즘을 연속적인 공간에서의 경로 탐색에 사용하는 데에는 몇 가지 문제점이 따른다. 우선 대부분의 강화 학습 알고리즘에서는 환경을 이산적으로 가정하기 때문에, 연속적인 시간과 공간을 갖는 환경에 바로 적용하기가 힘들다. 주로 쓰이는 대안은 신경망과 같은 함수 근사장치를 사용하는 방법으로 많은 성공적인 결과가 있어 왔으나, 이러한 근사 장치를 사용하더라도 문제가 복잡해짐에 따라 학습해야 할 파라미터가 지수적으로 증가하는 차원성의 저주는 피할 수가 없다. 즉, 환경이 복잡해질수록 강화 학습을 사용하여 경로를 찾아내는 데에는 오랜 시간이 걸리게 되며, 따라서 동적인 환경에서 환경의 변화에 빠르게 적응하기 어렵게 된다.

본 논문에서는 이러한 문제를 해결하기 위하여 작은 세상 성질(Small World Property)[2]를 가지는 네트웍 모델에 기반한 강화 학습 알고리즘을 사용한다. 작은 세상 성질이란 실세계의 네트워크에서 자주 보이는 성질로 네트워크의 크기가 커지더라도 임의의 두 노드 간에 상대적으로 짧은 경로가 존재한다는 것을 의미한다. 이러한 작은 세상 성질을 가지는 네트워크를 이용할 경우 문제가 복잡해지더라도 임의의 두 노드 간에 짧은 경로가 존재하게 되며, 따라서 수렴 속도를 증가시킬 수 있어 동적인 환경에 빠르게 적응할 수 있게 된다.

2. 강화 학습을 이용한 경로 탐색

2.1 강화 학습을 이용한 경로 탐색

강화 학습 프레임웍에서는 학습의 주체인 에이전트가 상태를 관측한 뒤 현재의 정책에 따라 행동을 취하게 된

다. 그 결과로 상태가 새로운 상태로 이동하게 되고, 행동의 결과인 보상을 받게 된다. 에이전트는 장기적인 보상의 기댓값을 최대가 되게 하는 정책을 학습하게 된다. 대표적인 강화 학습 알고리즘으로는 상태 s나 행동 a에 평가치 $V(s)$ 혹은 $Q(s,a)$을 책정하고 매 행동시마다 다음과 같이 이들을 수정해 나가는 방법을 들 수 있다[3].

$$V(s) \leftarrow V(s) + \alpha[r + \gamma \max_{s'} V(s') - V(s)] \quad (1)$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s,a') - V(s)] \quad (2)$$

평가치의 수정이 끝난 후에는 매 상태마다 최대의 평가치를 갖는 인접 상태나 행동을 선택함으로써 장기적인 보상을 최대화할 수 있다. 경로 탐색 문제에 이를 적용하기 위해서는 목적지에만 보상을 주고, 알려지지 않은 공간을 탐색하기 위해서는 가보지 않은 공간에 보상을 줌으로써 단일 프레임웍 하에서 두 문제를 동시에 해결할 수 있다. 또한 동적인 환경의 경우, 각 상태나 행동에 특정 상태가 막혀 있을 확률 p를 할당하고 시간에 따라 이 p를 감소시키는 방식으로 간단히 적용이 가능하다.

평가치 반복 알고리즘의 장점은 당장 해를 구할 수 있는 실시간 알고리즘이고, 평가치가 수렴할 경우 전역 최적치를 구할 수 있으며, 모든 상태에 대해 정책을 계산하기 때문에 멀티 에이전트 상황에 바로 적용이 가능하며 에이전트의 위치가 계획에서 벗어나도 바로 대처가 가능하다는 것들이 있다. 반면 이러한 방식의 단점은 이산적인 상태를 가정하기 때문에 연속적인 공간에 바로 적용하기가 힘들며, 문제 크기가 증가함에 따라 학습 시간이 크게 증가한다는 점이다[4].

2.2 ITPM

연속적인 공간에 강화 학습을 적용하기 위해서는 상태 공간을 이산화하거나 함수 근사장치를 사용하는 방식이 일반적이다. ITPM(Incremental Topology Preserving Map)[5]은 이러한 함수 근사장치의 일종으로, 알려진 상태 공간을 일정한 영역을 가지는 노드들의 영역으로 나눈다. 노드들의 위치는 자기조직화를 사용해 재배치되게 되고, 각 공간간의 연결 관계는 노드들 간의 에지로 표시된다.

이 ITPM의 경우 공간을 보로노이 다이어그램(Voronoi Diaram)으로 분할하고, 분할된 공간간의 연결관계도 얻을 수 있어 얻어진 노드들과 에지로 이루어진 그래프 상에서 위상적 경로탐색(Topological navigation) 이 가능하기 때문에 경로 탐색 문제에 적합한 특징을 갖는다. 하지만, 복잡한 환경을 나타내기 위해서는 많은 노드들이 요구되는데 이 경우 근사장치를 사용하더라도 학습 시간이 문제 크기에 따라 증가하게 된다.

2.3 SW-ITPM

앞서 말한 바와 같이 함수 근사장치를 사용할 경우에도, 상태 공간이 복잡해질 경우 학습해야 하는 함수 근사장치의 파라미터 수가 지수적으로 늘어나 학습하는 데 걸리는 시간이 크게 늘어나게 됨이 알려져 있다. 본 논문에서는 이에 대처하기 위해 고안된 작은 세상 성질(Small World Property)을 가진 자기 조직화 성장 신경망 모델인 SW-ITPM[6]을 사용한다. SW-ITPM이 하는 일은 다음과 같다.

1. 행동 a를 행하고 다음 상태 x'와 보상 z를 받는다.
2. ITPM에서 x'에 가장 가까운 노드 b'를 찾는다.
3. x'가 b'에서 γ이상 떨어져 있을 경우 새로운 노드를 그 위치에 생성하고 5번으로 간다.
4. b'의 Q값을 사용해서 다음의 행동 a'를 선택한다.
5. RL 알고리즘을 사용해서 기존의 가장 가깝던 노드 b의 Q값을 수정한다.
6. 자기조직화:b'의 연결 상태와 위치를 수정한다.

그림 1. SW-ITPM 알고리즘

1. 새로운 노드 u가 추가되었을 경우
   (a)u와 b',u와 b''를 연결하는 에지를 만든다.
   (b)b'와 b''간의 에지를 제거한다.
   (c)노드 v를 다음의 확률분포에 따라 선택한다.
      MODEL 1: $distance(u,v)^{-p}$
      MODEL 2: $d(v)$
   (d) u와 v를 연결한다.
   추가되지 않았을 경우
   (e) b'와 b''를 연결하는 에지를 만든다.
2. b'와 b'의 인접 노드들 r을 x'쪽으로 이동시킨다.

$$w_{b'} \leftarrow w_{b'} + \delta(x' - w_{b'})$$

$$w_r \leftarrow w_r + \delta(x' - w_r)$$

그림 2. SW-ITPM의 자기조직화 알고리즘

3. 실험 및 결과

3.1 실험 환경

실험 환경으로는 연속적인 3차원 환경을 제공하는 게임인 Half-Life 2[7]와 50*150*10m 사이즈의 2층 실내 환경을 모델링한 dm_lockdown 맵을 사용하였다. 에이전트는 크기가 1*1*1m이고 환경 안에서 임의의 방향으로 최대 7m/s로 이동할 수 있다. 에이전트가 사용 가능한 정보는 현재의 위치와 주위 8방향의 1m 내의 장애물 유무이고, 매 1/20초마다 평가치를 수정하고 측정 정보를 바탕으로 다음의 진행 방향을 결정하게 된다. 여러 대의 에이전트가 동시에 행동할 수 있으며, 이 경우 이들 에이전트는 ITPM을 공유하게 된다.

3.2 실험 결과

우선 에이전트의 공간 탐색 능력을 테스트하기 위해 에이전트들을 맵의 여러 군데에 투입하였다. 파라미터는 $r=0.75m, \delta=0.0002, \delta_r=0.00002, p=\log_2 5, \varepsilon=0.1,$ $\alpha=0.5, \gamma=0.9, p_l=0.1$을 사용하였다. 각 에이전트들은 자기 주위의 공간부터 탐색해 나가, 아래의 그림 3과 같이 전체 공간을 무사히 학습하였다.

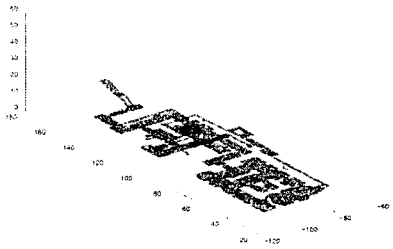


그림 3. 학습된 ITPM

그림 4는 작은 세상 모델을 적용했을 경우와 그렇지 않을 경우의 정보 전파 속도를 비교한 그래프이다. 작은 세상 모델을 적용할 경우 정보의 전파 속도가 훨씬 빠름을 알 수 있다.

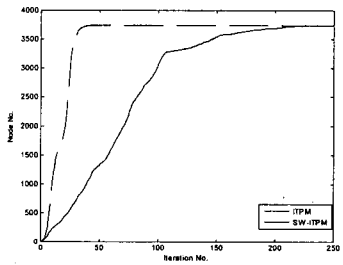
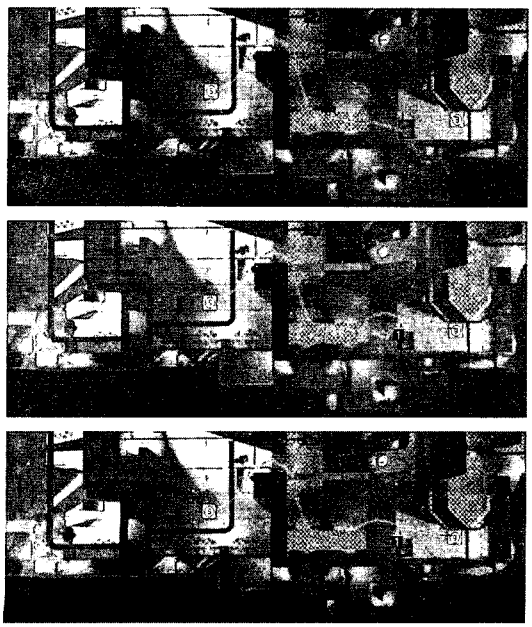


그림 4. 정보 전파 속도의 비교.

마지막으로 동적으로 변하는 환경에 대처하는 능력을 보기 위해, 장애물 1,2를 차례로 설치하였다. 실험 결과 장애물의 존재를 인지한 에이전트는 돌아가 새로운 최단 경로를 찾아내고, 그 후에는 더 이상의 시행착오 없이 새로운 경로를 사용하는 것을 알 수 있었다.



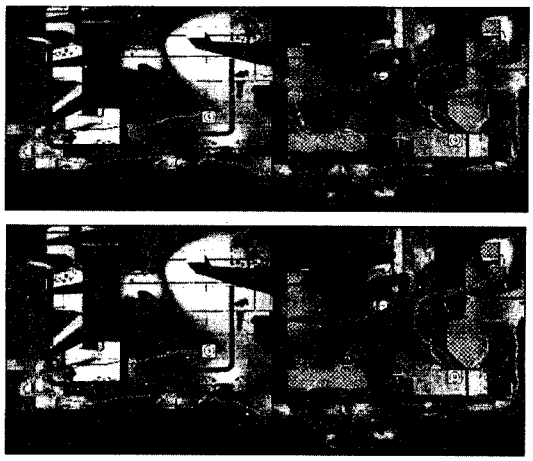


그림 5. 에이전트의 이동 경로.

4. 결론

동적으로 환경이 변하는 3D 게임 환경에서의 실시간 경로 탐색을 위해 작은 세상 네트웍 모델을 적용한 강화학습 알고리즘을 사용하였다. 실험 결과 환경을 학습한 뒤 환경의 변화를 관측하고 그에 적응하여 새로운 경로를 찾는 것을 확인할 수 있었고, 작은 세상 모델을 적용하지 않았을 때 보다 바뀐 환경에 적응하는 시간이 크게 줄어듦을 알 수 있었다.

감사의 글

이 논문은 교육인적사업부의 BK21 사업과 산업자원부에 의해 지원되었음.

참고 문헌

[1]Rabin. S. AI fame programming wisdom, Charles River Media, Inc.

[2]Kleinberg, J. Small-world phenomena and the dynamics of information.

[3]Sutton, R.S. Barto, A.G. Reinforcement learning: an introduction. MIT press, 1998.

[4]Turan, S. Learning metric-topological maps for mobile robot navigation. Artificial Intelligence, 99, 21-71 , 1998.

[5]Millan, D.R., Posenato, D., Dedieu, E. Continuous-action q-learning. Machine Learning, 49, 241-265, 2002.

[6]이승준, 장병탁. 복잡계 네트워크를 이용한 강화 학습 구현, 한국정보과학회 가을학술발표 논문집, pp.232-234, 2004.

[7] Hodgson, D. Half-life 2: raising the bar. Prima Games.