반응형 에이전트의 효과적인 물체 추적을 위한 베이지안 추론과 강화학습의 결합

민현정°, 조성배
연세대학교 컴퓨터과학과
{solusea, sbcho}@cs.yonsei.ac.kr

Hybrid of Reinforcement Learning and Bayesian Inference for Effective Target Tracking of Reactive Agents

Hyeun-Jeong Min and Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

에이전트의 '물체 따라가기'는 전통적으로 자동운전이나 가이드 등의 다양한 서비스를 제공할 수 있는 기본적인 기능이다. 여러 가지 물체가 있는 환경에서 '물체 따라가기'를 하기 위해서는 목적하는 대상이 어디에 있는지 찾을 수 있어야 하며, 실제 환경에는 사람이나 차와 같이 움직이는 물체들이 존재하기 때문에 다른 물체들을 피할 수 있어야 한다. 그런데 에이전트의 최적화된 피하기 행동은 장애물의 모양과 크기에 따라 다르게 생성될 수 있다. 본 논문에서는 다양한 모양과 크기의 장애물이 있는 환경에서 최적의 피하기 행동을 생성하면서 물체를 추적하기 위해 반응형 에이전트의 행동선택을 강화학습 한다. 여기에서 정확하게 상태를 인식하기 위하여 상태를 추론하고 목표물과 일정거리를 유지하기 위해 베이지안 추론을 이용한다. 베이지안 추론은 센서정보를 이용해 확률 테이블을 생성하고 가장 유력한 상황을 추론하는데 적합한 방법이고, 강화학습은 실시간으로 장애물 종류에 따른 상태에서 최적화된 행동을 생성하도록 평가함수를 제공하기 때문에 베이지안 추론과 강화학습의 결합모델로 장애물에 따른 최적의 피하기 행동을 생성할 수 있다. Webot을 이용한 시뮬레이션을 통하여 다양한 물체가 존재하는 환경에서 목적하는 대상을 따라가면서 이종의 움직이는 장애물을 최적화된 방법으로 피할 수 있음을 확인하였다.

1. 서론

이동로봇의 활용 범위가 변화와 다양성이 내포된 실제 환경으로 확대됨에 따라 다양한 서비스와 기능이 요구되었다[1]. '물체 따라가기'는 전통적으로 탄도 미사일과 같은 군 응용 프로그램과 항공시스템, 인공위성, 이동로봇 등에 많이 활용되었는데, 자동 운송수단이나 로봇의 보호자 기능 또는 구조작업에까지 응용될 수 있다. 이러한 기능은 사람과의 상호작용을 기반으로 예측되지 않은 실제 상황의 변화에도 적절한 기능을 수행할 수 있어야 한다. 따라가기의 정확성을 요구했던 기존의 시스템은 레이더, 초음파, 적외선 등의 센서로 거리와 위치 정보를 이용하여 행동을 생성했다. 그러나 다양한 종류의 움직이는 물체가 존재하는 상황은 이러한 센서만으로 목표와 장애물을 감지하기 어렵기 때문에 카메라와 같은 더 많은 정보를 줄 수 있는 센서가 필요하다[2].

카메라 센서는 물체에 대한 정보를 줄 수 있지만 거리에 대한 정보를 얻을 수 없고 에이전트와의 거리나 위치에 따라 다른 모양으로 정보가 들어오기 때문에 정확한 정보를 추출하기 어렵다. 최근 불확실한 센서정보에서 필요한 정보를 추출하는 연구로 퍼지이론이나 신경망을 이용한 방법으로 다양하게 진행되고 있다. 본 논문에서는 다양한 물체가 있는 상황에서 '장애물 피하기'와 '물체 따라가기' 기능을 생성하기 위해 베이지안 추론과 강화학습을 하이브리드 한다. 이 방법에서 환경 정보를 수집하기 위해 카메라 센서와 적외선 센서를 함께 이용하고, 제안한 방법은 센서에서 수집할 수 있는 정보의 불확실성을 보정하고 비교적 정확한 상태(에이전트가 직면한 상황)를 추측할 수 있다. 하이브리드 모델에서 베이지안 추론은 불확실한 정보로부터 확률 테이블을 생성하여 확률값을 제공함으로써 가장 유력한 상황을 추론할 수 있도록 한다.

에이전트가 처한 상황을 정확히 추론한다고 해도 다양한 물체가 있는 환경에서는 "물체 따라가기"기능을 완수하기 어렵다. 예를 들어서 사람들이 많이 있는 공간에서 에이전트가 목표를 따라가고 있다면 에이전트에게 관심을 갖는 사람들이나 임의로 에이전트와 목표물 사이를 지나가는 사람의 행동을 제지할 수 없고 에이전트는 이런 상황에서도 목표물을 잃지 않고 따라갈 수 있어야 한다. 이렇게 에이전트와 목표물 사이에 지나가는 대상이 사람이나 또는 다른 물건이 될 수도 있기 때문에 그 대상의 크기와 종류에 따라서 에이전트는 자율적으로 최적화된 피하기와 따라가기 행동을 선택할 수 있어야 한다. 제안한 방법에서 강화학습은 다양한 장애물을 피하기 위한 반응형 에이전트의 최적화된 행동선택을 학습할 수 있도록 하며, 상태-행동의 실시간 온라인 학습을 위해 다중 에이전트나 다양한 환경에 적합한 학습방법이다[3]. 이 강화학습의 학습률을 높이기 위해 이전에 받았던 점수를 가감하여 학습의 효과를 높이도록 하였다.

본 논문에서는 Webot 시뮬레이션을 통해 제안하는 방법으로 다양한 장애물 종류에 따라 최적화된 방법으로 움직이는 장애물을 피하면서 목적 대상을 따라갈 수 있음을 보인다. 학습 결과의 분석을 통하여 얼마나 적절하게 에이전트가 처한 상태를 추론할 수 있었는가와 장애물의 종류와 크기에 따라서 어떤 피하기 행동을 학습하였는지를 보인다.

2. 반응형 에이전트의 추론 및 학습

실제환경에서 센서는 부분적으로 관찰가능하고 불확실한 정보에 대한 오차가 존재한다. 따라서 에이전트가 정확한 행동을 생성하기 위해서는 불확실한 센서 정보로부터 상태를

정확하게 인지하고 그 상태에서 최적화된 행동을 선택할 수 있어야 한다. 에이전트의 카메라 센서와 적외선 센서는 실제 환경에서 항상 예측하기 어려운 오차가 존재하기 때문에 베이지안 네트워크를 이용하여 정보를 추론하여 결과값을 얻는 방법이 유용하다. 실제로 베이지안 네트워크를 이용해서 센서를 퓨전하는 방법은 제안되었지만 이 방법으로 최적의 장애물 피하기와 따라가기 행동이 시도되지는 못했다[4].

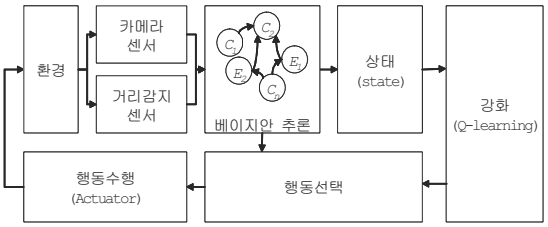


그림 1. 반응형 에이전트 추론 및 학습의 하이브리드 구조

본 논문에서는 베이지안 추론과 강화학습의 하이브리드 구조를 제안하여 장애물 종류에 따른 최적의 피하기 행동과 함께 물체를 따라가는 것을 보이고자 한다. 에이전트의 최적화된 행동 선택은 학습을 통해 가능한데 실시간으로 학습하기 위해 강화학습이 필요하다[3]. 베이지안 네트워크는 불확실한 정보로부터 비교적 정확한 결과 값을 추론할 수 있는 robust 한 방법이고 최근 다양하게 연구되고 있다[5]. 제안하는 방법의 하이브리드 구조는 그림 1 과 같다.

2.1. 상태 및 행동의 베이지안 추론

DAG 로 표현되는 베이지안 네트워크는 확률 값으로 표시된 현상들의 조건들로 원인과 결과의 관계를 네트워크로 구성하여 원인들로부터 결과를 추론하도록 구성된다. 그림 2 는 에이전트가 직면한 상태와 목표의 방향과 속도를 추론하기 위해 설계된 베이지안 네트워크의 구조이다. 이 베이지안 네트워크는 상태를 추론하고 따라가기 행동을 위하여 목표의 방향과 속도를 예측한다. 불확실한 센서 정보로부터 원인으로 구성되는 노드가 '카메라의 왼쪽', '카메라의 오른쪽', '거리감지 센서'이고, 이전 정보를 이용해서 상황 예측에 필요한 원인으로 구성된 노드가 '이전 목표의 진행방향', '이전 로봇의 진행방향'이다. 초기값으로 모든 원인정보에 대한 확률은 각 노드에 있는 조건의 개수 n 에 대하여 $1/n$ 로 정규화하였다. 불확실한 센서에 대한 예측정보도 중요하지만 다양한 환경을 예측하기 위해서는 '이전 목표의 진행방향', '이전의 로봇 진행방향'과 장애물의 유무를 추론하기 위해 '장애물과의 거리'의 예측 정보가 중요한 인자로 작용한다.

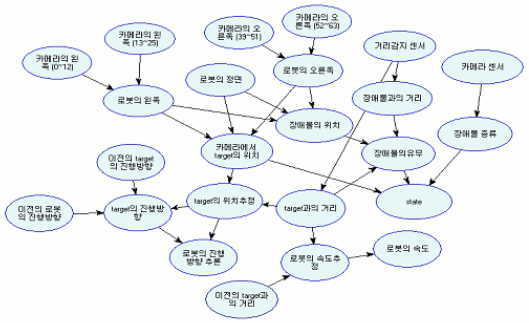


그림 2. 에이전트의 상태와 목표의 방향 및 속도를 추론하기 위한 베이지안 네트워크

그림 2 에서 각 노드의 정보는 실험을 통해 실제 센서 정보로 계산된다. 카메라 센서의 픽셀 정보를 5 등분으로 나누어 왼쪽에 2 개, 오른쪽에 2 개, 정면으로 정의하였다. 이 카메라 센서와 거리감지 센서의 정보에 따라 목표의 위치와 진행방향의 확률 테이블 값이 변경된다. '로봇의 속도' 확률 테이블은 거리감지와 카메라 센서의 위치에 따라 에이전트와 목표물의 실제 거리를 측정하며 이 거리가 일정거리와 얼마나 차이가 있는지 추론하고 결과로 에이전트의 속도를 결정하도록 한다.

2.2. 장애물 피하기의 강화학습과 평가

강화학습은 기대하는 행동을 수행하도록 하기 위해 피드백을 받아서 연산하는 trial-and-error 학습방법으로, 학습하는 에이전트는 환경과의 상호작용을 통해 지속적으로 평가를 받고 실패를 통해 강화된다. 따라서 강화학습에서 상태와 행동의 정책과 더불어 평가 기준이 중요한 요인이 된다. 상태에서 행동의 선택을 빠르게 학습시키기 위하여 처음에 모든 행동이 선택될 수 있도록 기회를 주고 다음에는 높은 점수를 받은 행동이 더 많이 선택되도록 하였다. 베이지안 네트워크에서 추론된 상태정보를 센서정보로, 상태의 전이를 목적으로, 그리고 에이전트가 선택하는 각 행동을 행동선택 테이블로 구성하고 상태가 전이되었을 때 행동을 평가한다. 학습은 로물체 따라가기를 수행하면서 장애물이 들어왔을 경우에 가장 효과적으로 피하기 행동을 생성하기 위한 상태와 행동의 연결을 찾는 것이다. 실험에서 정의된 상태와 행동은 표 1 과 같다.

표 1. 강화학습을 위한 상태와 행동의 정의

상태	
Nothing	Only_Target
NT_Stick_Left	Stick_Left
NT_Stick_Right	Stick_Right
NT_Circle_Left	Circle_Left
NT_Circle_Right	Circle_Right
NT_Square_Left	Square_Left
NT_Square_Right	Square_Right
행동	
Waiting	Following
AvoidLeft	FindTarget
AvoidRight	

강화 학습을 위한 평가 기준은 에이전트가 행동을 한 다음에 충돌이 있는지, 목표물과의 거리는 얼마나 되는지 그리고 그 행동을 수행하는데 걸린 시간은 얼마인가 이다. 장애물을 피하고 목표를 따라가기 위한 최적의 정책을 결정하기 위해 Q-learning 평가방법을 이용한다. 식 (1)은 강화학습에 적용된 평가를 나타낸다. 이 식에서 a, b, c 는 각각의 평가 함수에 대한 가중치이며 $a+b+c=1$ 이고, 세 평가 함수 중에서 충돌이 평가에 좀 더 많은 영향을 주도록 하였다. δ 는 $R(b)$ 의 평균을 -1 에서 1 로 정규화시킨 값으로 이전에 평가를 더해서 더 좋은 행동에 더 많은 점수를 주고 더 많이 선택되도록 하였다. 식 (2), (3), (4)는 각각 충돌에 대한 평가, 목표물과 에이전트와의 거리, 그리고 행동을 수행하는데 걸린 시간에 대한 평가함수이다.

- d_{target}: 목표물과 에이전트의 거리
- d_{obst}: 장애물과 에이전트의 거리
- Θ: 임계값, M_d, M_t: 거리와 시간의 최대값
- m_d, m_t: 거리와 시간의 최소값, $\delta = \eta \frac{\sum_b R(b)}{n}$

$$R(b,t) = aR_{dist}(b,t) + bR_{time}(b,t) + cR_{collison}(b,t) + \delta \quad (1)$$

$$R_{collison}(b,t) = \begin{cases} 10 & \text{if } d_{obst} < \theta \\ -10 & \text{if } d_{obst} \le \theta \end{cases} \quad (2)$$

$$R_{dist}(b,t) = \begin{cases} \dfrac{-20(d_{target} - \theta)}{M_d - \theta} & \text{if } d_{target} > \theta \\ \dfrac{20(d_{target} - \theta)}{\theta - m_d} & \text{if } d_{target} < \theta \\ 10 & \text{if } d_{target} = \theta \end{cases} \quad (3)$$

$$R_{time}(b,t) = \begin{cases} \dfrac{\text{avr} - t}{\text{avr} - m_t} & \text{if } t > \text{avr} \\ \dfrac{\text{avr} - t}{M_t - \text{avr}} & \text{if } t < \text{avr} \\ 0 & \text{if } t = \text{avr} \end{cases} \quad (4)$$

$$A(s,b) = \sum_t R(s,t) \quad (5)$$

학습 알고리즘은 상태 s 와 행동 b 의 연결을 위한 함수를 학습하도록 하였다. 이 함수가 행동을 선택하기 위해 사용되며 행동 선택은 반복적으로 수행하면서 5 번까지 모든 행동이 선택될 수 있도록 하였고 그 이후는 높은 평가를 받은 행동을 더 많이 선택하도록 하였다. 테이블로 구성되며 함수는 식 (5)와 같다. 이 식에서 A(s, b)는 기본적인 5 번 선택 이후의 행동 선택 기준이 된다. 5 번의 기본 행동을 선택하도록 한 이유는 상태에서의 모든 행동이 공평하게 점수를 받을 수 있도록 하기 위함이다.

3. 실험 및 결론

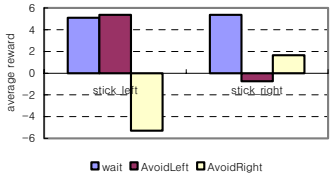
실험은 webot 시뮬레이터를 이용하고 실험 로봇은 8 개의 적외선 센서와 8 개의 빛 감지 센서 그리고 1 개의 CCD 카메라 센서를 가지고 있다. 실험 환경은 4 면이 각각 울타리로 막혀 있으며, 제안된 방법으로 행동을 생성하는 로봇과 따라가기를 할 목표가 있고 원형과 막대형의 움직이는 장애물이 있다. 여기서 막대형의 길이를 바꿔가며 실험하였다. 목표 로봇과 장애물은 임의로 돌아다니도록 되어있고 실험 로봇은 목표 로봇을 일정간격을 유지하면서 따라다닌다. 실험 로봇은 임의의 위치에서 시작되며 실험을 반복하면서 각 상태마다 5 번씩 학습할 수 있도록 하였고 그 이후에는 높은 점수를 받은 행동이 더 많이 선택되도록 구성하였다.

표 2. 강화학습을 통해 선택된 상태와 행동의 연결 예

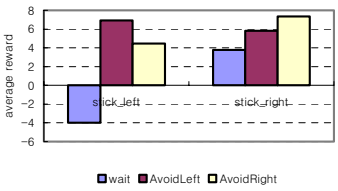
상태	행동
Nothing	FindTarget
Only_Target	FollowTarget
Stick_Left	AvoidLeft
Stick_Right	AvoidRight
Circle_Left	AvoidLeft
Circle_Right	Wait

장애물 피하기의 최적화된 행동을 학습하기 위해 제안된 방법인 반응형 에이전트의 행동선택에 대한 학습결과를 분석하기 위해 먼저 각 상태에서 선택한 행동을 분석해 보았다. 이 실험을 위해 각 상태에서 5 분씩 반복실험 하였고 행동 선택은 2.2 절에서와 같이 처음 5 번은 모든 행동이 선택되어 평가를 받을 수 있도록 하였고 다음에는 높은 점수를 받은 행동이 더 많이 선택되도록 하였다. 표 2 는 각 상태에서 선택된 행동이다.

장애물의 길이에 따른 피하기 행동을 분석하기 위해 막대 모양 장애물에 대하여 각각 1.5와 2.5길이에서 피하기 행동을 비교하였다. 길이가 비교적 짧은 1.5 장애물에서 로봇은 기다리기 행동이 길이 2.5 장애물에서보다 높은 점수를 받았다. 그러나 비교적 길이가 긴 2.5 장애물에서 로봇은 왼쪽에서 다가올 경우에는 왼쪽으로 피하기 행동이 오른쪽에서 다가올 경우에는 오른쪽으로 피하기 행동이 높은 점수를 받았음을 그림 3 의 결과에서 알 수 있다.



(a)



(b)

그림 3. 막대형 장애물의 길이 1.5(a)와 2.5(b)에서의 피하기 행동 학습 결과

실제 환경과 같은 다양한 장애물이 존재하고 변화하는 환경에서 물체 따라가기 기능을 수행하기 위해서는 상황의 정확한 추론과 적합한 장애물 피하기 행동을 생성하는 것이 중요하다. 본 논문에서는 상태-행동의 정책을 학습시켜 최적의 장애물 피하기 행동과 물체 따라가기 행동을 생성하도록 제안하였다. 그러나 에이전트의 불확실한 센서정보로 상황을 정확하게 추론하지 않으면 만족할만한 결과를 얻기 어렵기 때문에 상황을 추론하고 물체를 예측하여 따라가기 행동을 수행하도록 베이지안 추론을 적용하였다. 실험을 통해 목표와 일정 간격을 유지하면서 목표를 따라갈 수 있었고 상태의 변화에 따라 최적의 방법으로 피하기 행동을 생성하면서 다시 따라가기를 할 수 있음을 확인할 수 있었다.

감사의 글

본 연구는 프론티어 연구사업의 지원에 의한 것임.

참고문헌

[1] R. R. Murphy, *Introduction to AI Robotics*, The MIT Press, 2000.
[2] N. Zhang and J. Weng, "A developing sensory mapping for robots," In *Proc. IEEE 2nd Int. Conf. on Development and Learning*, pp. 13-20, 2002.
[3] R. C. Arkin and G. A. Bekey, *Robot Colonies*, Kluwer Academic Publishers, 1997.
[4] C. Coue, T. Fraichard, P. Bessiere, and E. Mazer, "Using Bayesian programming for multi-sensor multi-target tracking in automotive applications," In *Proc. IEEE Int. Conf. on Robotics & Automation*, pp. 2104-2109, 2003.
[5] H.-J. Min and S.-B. Cho, "Bayesian inference driven behavior-network architecture for intelligent agent to avoid collision with moving obstacles," *Journal of KISS*, vol. 31, no. 8, 2004.