Proceedings of the 12th
KACC, October 1997

강화 학습에 의한 소형 자율 이동 로봇의 협동 알고리즘 구현

A Reinforcement Learning-based Method for the Cooperative Control of Mobile Robots

0김재희*, 조재승**, 권인소***

* 현대자동차㈜ 상용기술연구소(Tel : +82-2-958-3465; Fax : +82-2-960-0510; paulj@chollian.net)
** 한국과학기술원 자동화및설계공학과(Tel : +82-2-958-3415; Fax : +82-2-960-0510;)
*** 한국과학기술원 자동화및설계공학과(Tel : +82-2-958-3415; Fax : +82-2-960-0510; iskweon@eekaist.kaist.ac.kr)

Abstracts This paper proposes methods for the cooperative control of multiple mobile robots and constructs a robotic soccer system in which the cooperation will be implemented as a pass play of two robots. To play a soccer game, elementary actions such as shooting and moving have been designed, and Q-learning, which is one of the popular methods for reinforcement learning, is used to determine what actions to take. Through simulation, learning is successful in case of deliberate initial arrangements of ball and robots, thereby cooperative work can be accomplished.

Keywords Mobile Robot, Robot Soccer, Q-learning, Reinforcement learning, cooperative control

1. 서론

최근 로봇 축구에 대한 연구들을 내용에 따라 분류를 하면, 실제 경기 출전을 위한 시스템 구성에 주안점을 둔 연구[1]와 로봇의 제어에 대한 관점에서 신경 회로망과 학습의 개념을 포함시킨 연구[2] 및 시뮬레이션 소프트웨어에 관한 연구[3] 등이 주종을 이루고 있으며 강화 학습 방법을 이용하여 학습을 시도한 연구[4][5]도 이루어지고 있다.

지금까지의 연구 중에서는 로봇의 협동 작업의 개념과 관련된 내용은 많지 않고 고전적인 if-then-else 구조의 판단 트리(decision-tree)에 의한 접근 방식[2] 등만이 제시되었다.

각 로봇의 행동을 결정하는 작업은 모델링에 의한 해결이 어려운데다가 대전(對戰)하는 상대에 따라 그에 맞는 적절한 전술을 구사하는 것이 유리하므로 행동 결정 과정은 학습에 의해 이루어지는 것이 바람직하다고 하겠다.

본 연구에서는 각 로봇의 기초적인 행동에 대한 제어로부터 시작하여 상위 수준의 행동 결정을 위해 경험에 의한 시행 착오(trial-and-error)를 학습하는 강화 학습(Reinforcement learning)을 사용하여 여러 대의 로봇들이 기초적인 협동 작업을 하기 위한 알고리즘을 제안하였다.

2. 로봇의 기초행동제어

2.1 *로봇의 제어계*

Fig.1 은 축구로봇의 제어기 전체에 대한 블록선도를 나타낸 것인데, 판단을 위한 행동결정부분(Action Selector), 기본기를 수행하는 행동제어기(Motion Controller), 그리고 강화학습과 신경회로망 학습모듈로 구성되어 있다.

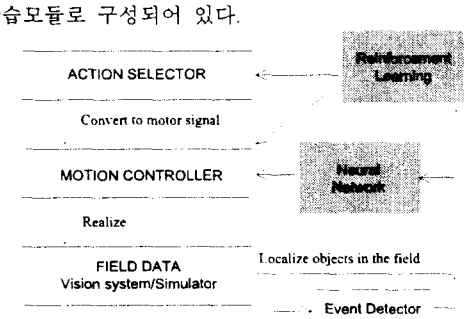
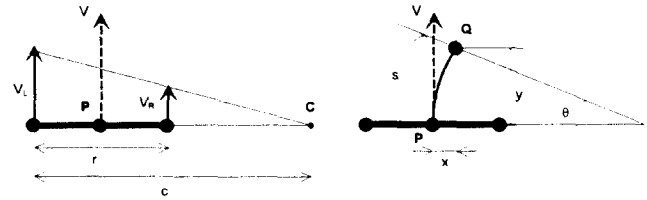


Fig.1 Block diagram of robot controller

2.2 *로봇의 구동*

로봇의 위치는 (x, y) 좌표에 의해 표현되고, 이 좌표를 입력으로 하여 로봇 제어 모듈에서는 $(\dot{x}, \dot{\theta})$와 같은 병진/회전 운동 성분의 속도를 출력하게 되며, 이를 받아 시뮬레이터에서 로봇의 위치를 갱신하게 된다. 그러나 로봇의 구동 기구는 보통 두 개의 모터가 양 바퀴를 각각 구동하는 구조로 되어 있고 따라서 로봇 제어 모듈에서의 출력값은 양 바퀴의 각속도로 변환하여 로봇의 위치를 갱신한다.



(a) (b)

Fig.2 Motion control signals (a) Linear velocities of either side of wheel (b) Displacement and steering angle

Fig.2의 왼쪽 그림에서 두 바퀴의 선속도를 각각 V_L과 V_R이라고 하면 강체 운동의 순간 중심은 점 C에 위치한다. 병진 운동의 속도 성분 V는

$$V = \frac{2\left(V_L^2 + V_R^2\right)}{V_L + V_R} \tag{1}$$

로 얻을 수 있다.
샘플링 주기를 Δt라고 하면 Δt동안 진행한 거리 s는 $V\Delta t$와 같고 순간 중심에 대해 회전한 각 θ는

$$\theta = \frac{D}{c - \frac{r}{2}} = \frac{4\Delta t}{r}\frac{\left(V_L - V_R\right)\left(V_L^2 + V_R^2\right)}{\left(V_L + V_R\right)^2} \tag{2}$$

와 같이 된다.
반대의 경우, 즉 실제 움직인 거리 s와 조향각 θ를 알고 있을 때 각 바퀴의 선속도를 구하면

$$V_L = \frac{2s + r\theta}{4s^2 + r^2\theta^2} sV \tag{3}$$

$$V_R = \frac{2s - r\theta}{4s^2 + r^2\theta^2} sV \tag{4}$$

와 같이 된다.
한편 공과 로봇의 충돌 후 속도 및 방향변화는 다음과 같은

과정으로 구한다. 우선 문제를 단순화하기 위해 첫 번째, 로봇의 경우 바퀴와 지면 사이의 미끄러짐이 없다고 가정하여 충돌 후에도 진행 방향이 바뀌지 않는다고 가정한다. 두 번째, 또한 로봇은 충돌 후 각운동량 성분을 가지지 않는다고 가정한다. 세 번째, 공의 회전 운동 성분은 무시하며 병진 운동만이 존재한다고 가정한다. 네 번째, 공과 지면과의 마찰은 무시한다. 마지막으로 충돌에서 변형시와 반발시의 충격의 크기의 비인 반발 계수 (coefficient of restitution, e)는 실험적으로 구한다.

위의 가정하에 세 개의 미지수 (u_2, v_2, θ_2)를 연립 방정식에 의해 풀면,

$$u_2 = \frac{(m-en)u_1 + (1+e)nv_1\cos\theta_1}{m+n} \quad (5)$$

$$v_2 = \sqrt{{v_1}^2\sin^2\theta_1 + e^2(u_1 - v_1\cos\theta_1)^2 + {u_2}^2 + 2eu_2(u_1 - v_1\cos\theta_1)} \quad (6)$$

$$\theta_2 = \tan^{-1}\left(\frac{v_1\sin\theta_1}{e(u_1 - v_1\cos\theta_1) + u_2}\right) \quad (7)$$

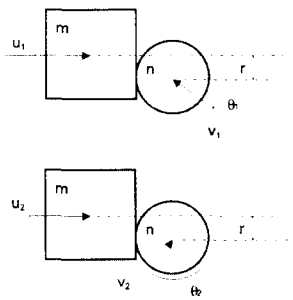


Fig.3 Impact between ball and robot (a) incidental (b) reflective

2.3 *기초 행동 제어*

로봇축구에서는 여러가지 기초행동들이 있을 수 있으나 그중에서도 이동하고 있는 공을 원하는 방향으로 차는 기초행동이 가장 복잡하고도 빈번하게 발생하므로 다음과 같은 방식으로 구현하였다. 공이 정지한 경우에는 Fig.4 에서 F_2에 의해 공과 목표를 잇는 직선상으로 움직이면서 F_1에 의해 공으로 다가가 충돌하면 된다. 계수 k_1은 진행 방향을 목표 지점으로 향하도록 하는 경향(조준 동작; Aiming)의 크기를, 계수 k_2는 공에 접근하려는 경향(접근 동작; Approaching)의 크기를 조절한다.

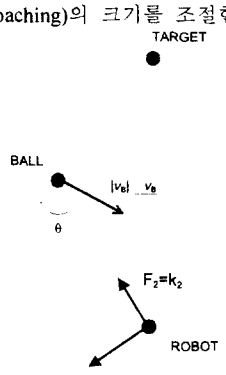


Fig.4 Parameters for shooting action

공이 이동할 경우에는 움직일 위치를 예측하여 미리 그곳을 기준으로 조준 동작을 해야 한다. 이와 같은 효과는 공의 속도를 감안하여 Fig.4 의 각 θ에 적당한 값 α를 더해 줌으로써 얻을 수 있다. 공이 빨리 움직일 수록 큰 값을 더해 보정해 주어야 하므로 α는 대체로 θ의 크기에 비례하는 성질을 가지지만 계산에 의해 얻기는 매우 어렵다. 따라서 본 논문에서는 공과 로봇, 목표 지점과의 상대 각도와 거리 등을 입력으로 하고 보정값 α를 출력으로 하는 신경 회로망을 구성하여 학습에 의해 보정을 하는 방법을 사용하였다.

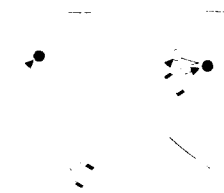


Fig.5 Shooting a moving ball

로봇의 위치를 3 가지, 로봇의 방향을 3 가지, 공의 위치를 3 가지, 공의 진행 방향은 4 가지, 공의 속도 2 가지를 모두 서로 조합하여 216 가지의 입력 패턴을 만들고 이 패턴을 5000 회 학습하였다. 설명된 패턴에 의한 학습을 검증하기 위해 공과 로봇의 위치, 방향 및 초기 속도를 무작위로 설정한 후 슈팅을 시켜본 결과는 Fig.7 과 Fig.8 에 표시하였다.

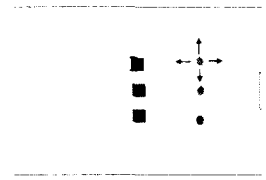


Fig. 6 Input pattern for learning a shooting action

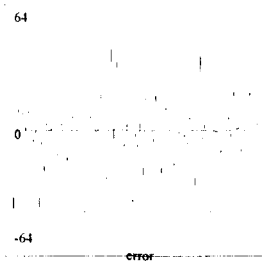


Fig.7 Shooting error versus testing steps

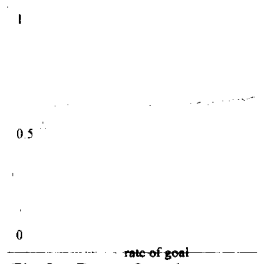


Fig.8 Rate of goal versus testing steps

학습 패턴에 의한 성공률과 대체로 비슷한 60%~70% 사이의 성공률을 보이고 있으며, 이는 무작위로 배치된 공과 로봇의 위치들 가운데 로봇의 속도의 한계에 의해 도저히 골을 성공시킬 수 없는 위치에 있을 때의 실패 때문이다.

3. 로봇의 협동제어

각각의 로봇은 앞에서 설명한 방식으로 여러 가지 기초적인 동작을 학습하여 수행할 수 있게 되었지만 전체적인 움직임과 상황을 판단하여 각 로봇에게 어떠한 동작을 수행해야 할 지를 결정해 주는 일이 필요하다. 이러한 문제에 적용할 수 있는 것으로써 강화 학습의 시간차(temporal difference)에 의한 방법 중 하나인 Q-Learning 이 있다[8]. Q-값이란 한 상태에서 어떠한 행동을 취했을 경우에 기대할 수 있는 보상 값에 대한 척도인데, Watkins 의 알고리즘[8]에 의하면 Q-값을 갱신하기 위한 식은 아래와 같다.

$$Q(s,a) := (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a' \in A} Q(s',a')\right) \quad (8)$$

식(8)에서 α는 학습 속도(learning rate)이고 γ는 절감율(discount factor), r은 그 순간의 보상 값(reward)이며, 상태 s는 상태 공간(state space) S, 행동 a는 행동 공간(action space) A의 원소(element)이다. 식에서 보듯이 Q-값은 상태 s와 행동 a의 함수인데, 이 함수 관계를 구현하기 위해 모든 s와 a에 대해 Q-값을 배열로 저장하기 위해서는 $O(|S||A|)$ 만큼의 기억 장소가 필요하게 되는데 이는 많은 기억용량을 요구하게되는 문제점이 있다. 또한 낮은 정밀도의 이산화된 양으로 구성한다면 이는 결과적으

로 다양한 상황에 대한 기술이 불가능하게 하여 학습이 무의미해지게 된다. 그러므로 본 논문에서는 상태 변수 s를 입력으로 하고 그에 따른 Q-값을 출력으로 하는 신경 회로망으로 이를 대체하였다[7]. Fig.9를 보면 신경 회로망의 입력 층(input layer)은 상태를 표시하는 13개의 절점(node)과 행동을 나타내는 8개의 절점으로 구성되어 있다. Q-값은 상태 s와 행동 a의 함수이므로 Fig.9와 같은 회로망을 기초 행동의 개수만큼 만들었다

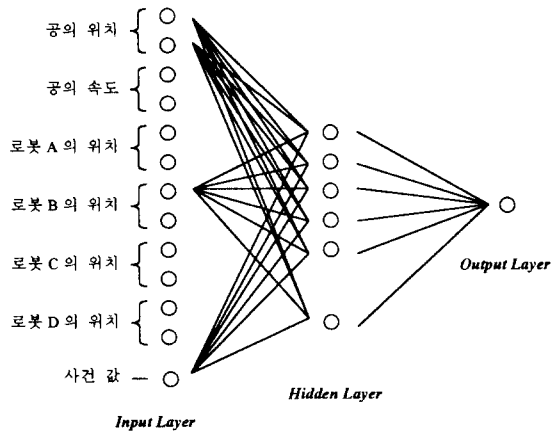


Fig.9 Value function as a neural network

Fig.9의 회로망을 사용하여 행동값(action value) Q(s, a)를 구하는 방법은 다음과 같다. 상태 s를 나타내기 위한 공의 위치와 속도, 및 로봇의 위치는 0과 1사이의 값으로 정규화(normalize)하여 입력하고, 사건 값은 뒤에서 설명할 사건(event)를 수치화하여 대입하여 회로망의 입력 층에 들어갈 값을 만들며, 이 입력값을 행동 a에 해당하는 'a번째' 회로망에 입력하여 출력을 얻는다. 이때의 출력값이 상태 s(공과 로봇의 배치)에서 행동 a(Table 1 참고)를 했을 때의 기대 보상값(expected reward)이다. 행동 a에 따라 별도의 회로망을 만든 것은 신경 회로망의 출력값 Q(s, a)가 각 s 값과 a 값에 대해 모두 독립적으로 계산되는 것이 아니기 때문에 어떠한 행동 a_0에 대한 학습 값의 갱신에 의해 인접한 다른 행동 a_1의 행동 값이 영향을 크게 받을 수 있기 때문이다. 이것은 상태 값 s에 대해서도 마찬가지 문제가 될 수 있지만, 상태 값 s는 그 값이 비슷한 경우 역시 비슷한 상황으로 인정할 수 있기 때문에 상태 s_0의 학습에 의해 인접한 상태 s_1의 값이 영향을 받는다 해도 오히려 비슷한 상황에 대한 학습을 한 효과가 될 것이다. 입력 층에 들어간 행동 1에서 행동 N은 앞에서 설명한 기초 행동들로 Table 1에서 정의하였다.

Table 1 Definition of elementary actions

일련 번호	기초 행동	설명
1	CONT	이전의 행동을 계속한다
2	FOLLOW	공의 뒤를 따라간다
3	SHOOT	공을 골문을 향해 찬다
4	BLOCK	공의 앞을 막는다
5	PASS1	지정된 위치(1)를 향해 찬다
6	PASS2	지정된 위치(2)를 향해 찬다
7	MOVE1	지정된 위치(1)로 이동한다
8	MOVE2	지정된 위치(2)로 이동한다

Table 1에서 '지정된 위치'란 Fig.10과 같이 공격 진영에서 골과 얼마 떨어진 위치에서 지면에서 보아 위/아래를 각각 위치 1, 2로 정의하였다. 움직이고 있는 공을 맞추어 정확하게 원하는 장소로 보내는 일이 쉽지 않기 때문에 패스를 위한 목표 지점을 상대편 진영의 양쪽 두 군데로 선정하였다. 마지막으로 행동 FOLLOW와 행동 BLOCK은 수비 동작을 위해 추가하였다.

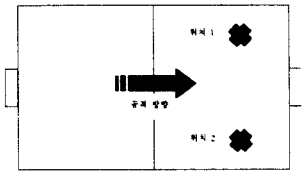


Fig.10 Designated positions for pass play

Table.2 Definitions of events

종류	명칭	설명
전체적 사건	GOAL	공이 오른쪽 골에 들어감
	GOAL2	공이 왼쪽 골에 들어감
	TIMEOUT	지정된 시간 초과
	HITBALLTARGET	공이 골이 있는 벽과 충돌
	HITBALLWALL	공이 골이 없는 벽과 충돌
해당 로봇에 관련된 사건	HITAGENTBALL	로봇이 공과 충돌
	HITAGENTAGENT	로봇이 다른 로봇과 충돌
	HITAGENTWALL	로봇이 벽과 충돌
	BALLNEAR	공이 일정 거리 내에 있음
	BALLFRONT	공이 로봇의 전방에 있음
	TARGETFRONT	목표 지점이 로봇의 전방에 있음
	ARRIVE	로봇이 목표 지점에 도착함

한편 각 행동들을 선택하여 각 로봇에게 전달하는 방법은 공의 진행 방향의 변화, 로봇 사이의 충돌, 앞서 받은 명령의 완수 등의 상황의 변화를 알 수 있는 사건(event)을 감지하여 이 때 새로운 상태에 대한 평가를 내려 새로운 행동을 지시한다.(Table 2 참고)

사건 GOAL과 사건 GOAL2가 발생하면 현재 진행을 중단하고 초기 위치로 돌아간다. 그 밖의 전체적 사건(Global events)이 발생하면 모든 로봇들에게 새로운 명령을 지시하고 로봇과 관련된 사건(Agent-specific events)이 발생하면 그 사건에 해당하는 로봇에게만 새로운 명령이 전달된다. 새 명령을 전달하기 전에는 이전 명령의 수행 결과에 대한 평가로써 Q-값을 갱신하게 된다. Fig.11에 사건 인식으로부터 학습이 이루어 지기까지의 흐름도를 나타내었다

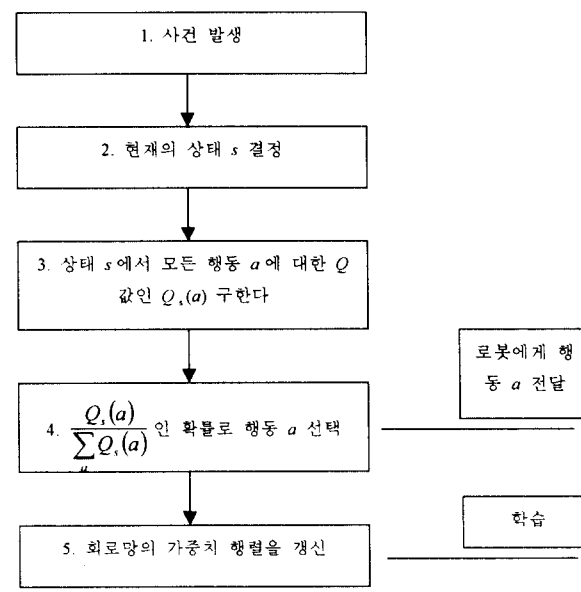


Fig.11 Flow chart of decision making and strategic learning

흐름도 내의 단계 4에서는 구해진 Q-값들로부터 행동 a를 결정하기 위해서 Q-값을 a에 따라 더한 누적 확률 분포 함수를 얻은 후 난수를 발생하여 이 값이 확률 분포 함수 내에 위치한

구간을 찾게 된다. 이것은 강화 학습의 exploration/exploitation 특성을 구현하기 위한 것이다[9]. 단계 5 에서는 식 (8)을 적용하여 새로운 Q-값을 계산한 후 이 값을 신경 회로망의 출력 층의 참값으로 생각하고 역전파(back-propagation)를 수행하여 회로망의 가중치 행렬(weight matrix)을 새로 계산한다.

4. 시뮬레이션 결과

시뮬레이션이 시작되면 각 로봇은 8 개의 기초 행동 중 한 가지를 선택하여 수행하게 된다. 어떠한 한 행동이 선택되어질 확률은 Table 1 에 열거된 현재의 상태에서의 각 기초 행동에 대한 행동 값(Action value)과 같다. 즉 큰 행동 값을 가진 행동이 선택될 확률도 높다. 모든 기초 행동들은 학습이 시작될 때 0 에 가까운 값을 가지고 시작하게 되지만 학습이 시작된 후 골이 성공하면 보상 값(reward) 1 을, 공을 찬 후 골에 성공하지 못하면 보상 값 0 을 받기 때문에 식 (8)에 의해 새로운 Q-값을 얻은 후 회로망을 학습시키게 된다.

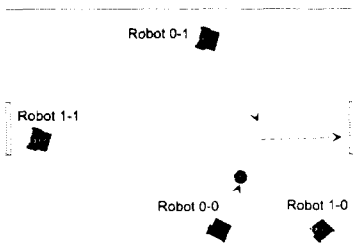


Fig.12 Situation : Pass and Shoot

Fig.12 와 같은 배치에서는 로봇 0-0 이 로봇 0-1 에게 공을 패스하고 그 공을 로봇 0-1 이 슛을 하는 상황을 생각할 수 있다. 반복하여 학습한 결과 로봇 0-0 과 로봇 0-1 의 기초 행동에 대한 행동 값(Action value)을 각각 Fig.13 와 Fig.14 에 나타내었다.

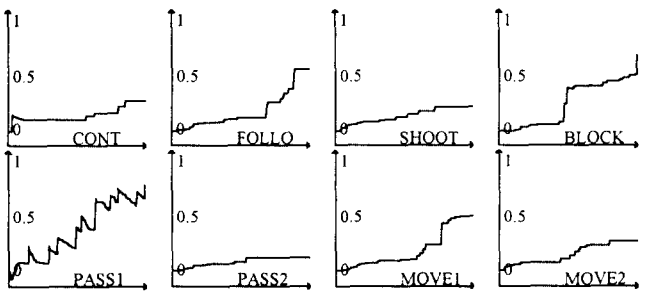


Fig.13 Action values of 8 elementary actions of robot 0-0 versus learning epoch

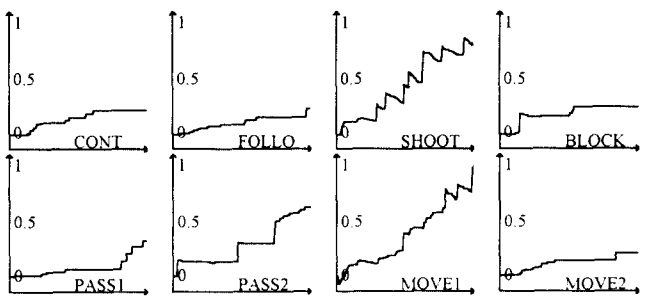


Fig.14 Action values of 8 elementary actions of robot 0-1 versus learning epoch

Fig.13 에서 보여지는 로봇 0-0 의 기초 행동 중에서는 행동 PASS1 과 FOLLOW, BLOCK, MOVE1 이 각각 상대적으로 높은 값을 가지는데, 행동 PASS1 및 FOLLOW, BLOCK, MOVE1 모두 로봇 0-1 이 위치한 방향으로 공을 보낼 수 있었기 때문이다. 반면 Fig.14 에서 보여지는 로봇 0-1 의 기초 행동 값의 변화를 살펴보면 행동 SHOOT 의 상승 폭이 다른 행동에 비해 크게 나타나는데, 이는 로봇 0-0 으로부터 건네 받은 공을 행동 SHOOT 에 의해 골로 성공한 경우가 많았기 때문이다.

5. 결론

이동 로봇의 행동의 대상이 되는 물체와 활동 공간의 특성에 따라 기초 행동을 결정하기 위해 슈팅 동작, 이동 동작 등의 알고리즘을 제안하였고, 단순한 수학적 모델링에 의해 정확한 동작을 구현하기 힘든 기초 동작들은 반복적 학습에 의해 이루어 내기 위해서 움직이는 공에 대한 슈팅 동작을 신경 회로망을 통한 학습으로 구현하였다. 또한 다른 이동 로봇들을 포함한 주위 환경에서의 상황에 따라 적절한 기초 행동들을 적절한 시점에서 선택하기 위해 강화 학습의 알고리즘을 적용하였다. 현 단계에서는 공과 로봇이 적절히 배치된 제한된 상황에서 축구 로봇은 자신이 처한 상황에 따라 패스 또는 슈팅 등의 행동 중에 적절한 것을 선택할 수 있는 가능성을 보였다.

지금까지 제시한 학습 알고리즘들은 시뮬레이션 환경에서 타당성을 알아 보았으나 실제 실험 장치에 적용하여 발생하는 문제점등을 알아 보아야 하겠다. 또한 앞에서 제시한 문제처럼 강화 학습에 의한 팀 전술의 학습은 실제 경기 중 발생할 수 있는 무수히 많은 상황들에 대해 대처할 수 있어야 하기 때문에 일반적인 공-로봇의 배치에 대해 학습이 가능하여야 하며 또한 그에 따라 오랜 시간이 소요되는 학습 속도를 빠르게 하는 것과 수렴성의 확보가 과제로 남는다. 마지막으로 팀 전술의 학습은 기초 행동의 성능에 따라 실패할 수 있으며 Q-learning 을 구현하기 위한 신경 회로망을 본 문제에 맞게 구성하는 것 또한 중요한 향후 과제이다.

참고문헌

1. H.S.Shim, M.J.Jung, H.S.Kim, I.H.Choi, W.S.Han, J.H.Kim, "Designing Distributed Control Architecture for Cooperative Multiagent Systems", Proc. of Micro-Robot world cup soccer tournament(MIROSOT), 1996
2. Peter Stone, Manuela Veloso, Sorin Achim, "Collaboration and Learning in Robotic Soccer", Proc. of Micro-Robot world cup soccer tournament(MIROSOT), 1996
3. John Harvey, Chao Cheng, Dennis Michaelson, "High-Level Design of a MIROSOT simulator", Proc. of Micro-Robot world cup soccer tournament(MIROSOT), 1996
4. Eiji Uchibe, Minoru Asada, Koh Hosoda, "Strategy Classification in Multi-agent Environment - Applying Reinforcement Learning to Soccer Agents -", ICMASS'96 workshop 2: RoboCup Workshop: Soccer as a Problem for Multi-Agent Systems, 1996
5. Minoru Asada, Shoichi Noda, Sukoya Tawaratsumida, and Koh Hosoda, "Purposive Behavior Acquisition On A Real Robot By A Vision-Based Reinforcement Learning", Proc. Of MLC-COLT Workshop on Robot Learning, 1994
6. P. Stone and M. Veloso, "Multiagent Systems: A Survey from a Machine Learning Perspective", IEEE Trans. Knowledge and Data Engineering, June 1996
7. Richard S. Sutton, "Reinforcement Learning II: Learning Values"
8. C.J.C.H.Watkins, "Learning from Delayed Rewards", PhD thesis, King's College, Cambridge, 1989
9. Leslie Pack Kaelbling, "Learning to Achieve Goals", Computer Science Department, Brown University
10. 김종환 外, "MIROSOT 에 대한 소개", KAIST 마이크로 로봇 축구대회 논문집, 1996