이동 로봇을 위한 행위 기반 제어 및 학습 구조의 설계와 구현

Design and Implementation of a Behavior-Based Control and Learning Architecture for Mobile Robots

이 상 훈, 김 봉 오, 서 일 홍*
(Sanghoon Lee, Bong Oh Kim, and Il Hong Suh)

Abstract : A behavior-based control and learning architecture is proposed, where reinforcement learning is applied to learn proper associations between stimulus and response by using two types of memory called as short Term Memory and Long Term Memory. In particular, to solve delayed-reward problem, a knowledge-propagation (KP) method is proposed, where well-designed or well-trained S-R(stimulus-response) associations for low-level sensors are utilized to learn new S-R associations for high-level sensors, in case that those S-R associations require the same objective such as obstacle avoidance. To show the validity of our proposed KP method, comparative experiments are performed for the cases that (i) only a delayed reward is used, (ii) some of S-R pairs are preprogrammed, (iii) immediate reward is possible, and (iv) the proposed KP method is applied.

Keywords : behavior-based, reinforcement learning, delayed reward, knowledge propagation

I. 서론

자율적이며 지능적인 시스템의 제작을 위하여 인지 기능, 추론 기능, 학습 기능 등 여러 방면에서의 많은 연구가 되어 왔다[1-3]. 인간의 지능 수준에 비교할 때 지능적일 것 같지 않은 하등 동물들의 행동이 그들의 환경과 구조를 고려하면 충분히 지능적인 행동임을 알게 되었다. 이러한 하등 동물의 행동을 모방한 행위 기반 구조를 적용한 로봇들이 80년대 중반 이후 크게 주장되었다[4-7].

행위 기반 구조의 특징은 로봇이 행할 수 있는 행위들의 집합과 로봇이 맞닿을 수 있는 상태 집합 안에서 로봇의 상태가 결정되면 인지나 추론의 계획과정(Plan) 없이 바로 해당 상태에 연결된 행위들이 선택되고, 선택된 행위 중 우선권을 갖는 행위가 다른 행위를 누르고 동작으로 수행되는 Subsumption[7]과 선택된 행위가 합쳐져 동작으로 수행되는 스키마(Schema)[6]가 있다. 행위 집합과 상태 집합 사이의 연결 관계가 고정된 로봇은 상태에 연관된 행동을 변경하지 못하기 때문에 불확실한 환경 속에서는 지능적인 행동을 유지 할 수 없다. 따라서, 로봇이 두 집합 사이의 연결 관계를 스스로 조절하는 방법으로 강화 학습을 사용할 수 있다[8-11].

강화 학습이 적용된 행위 기반 구조로 자율적이며 지능적인 로봇을 만들려 할 때 여러 가지 사항을 고려하여야 한다. 예를 들어 설계자는 변화하는 환경 속에서 로봇에게 필요한 상태에 대한 행동(S-R 행동)들을 설계해야만 하는바, 여기에는 계층적[7], 수평적[6], 파라미터 조절적[12][13] 행위 분해 방법이 있다. 한편, 로봇 사용자(또는 설계자)는 아직 정해지지 않은 상태 집합과 행동 집합 사이의 필요한 연결관계가 반복 학습되기까지 많은 시간을 기다려야만 한다[14][15]. 특히, 로봇이 학습을 하기 위해서 필요한 보답을 전달 받는 시점이 중요하다. 보답을 전달하는 방법은 즉각 보답과 지연 보답으로 나눌 수 있다. 즉각 보답은 상태에 대한 행동이 올바른 정도를 행동 이후 바로 계산하기 때문에, 강화 학습을 상대적으로 쉽고 빠르게 진행할 수 있었다. 그러나 행동에 대한 보답을 즉시 전달하기 위하여 보답을 관리 할 내부 또는 외부의 관찰자가 로봇의 행동을 항상 지켜봐야 한다[16]. 그리고 보답을 잘못 했을 경우에 대한 대처가 부족하다. 한편, 지연 보답은 현재에 받은 보답을 가지고 과거 행동들을 평가한다[17]. 따라서 최적의 행동을 관찰자가 알고 있을 필요도 없으며, 항상 로봇을 지켜 볼 필요도 없다. 그러나 학습이 어렵고 학습 시간이 더 많이 걸린다[16].

본 논문에서 제안한 행위 기반 제어 구조는 외부 환경의 상태를 자극으로 입력 받아 행위 선택 모듈에서 내부 상태와 입력 받은 자극의 영향을 고려하여 알맞은 행위를 선택하며 이 행위 선택 모듈은 외부 자극에 대해 알맞은 행위를 선택 할 수 있도록 강화 학습을 한다. 강화 학습은 입력된 자극과 현재 선택된 행동들을 저장하는 ST(Short Term)메모리와 보상(reward)을 받았을 때 ST메모리에 저장된 자극들과 행동들간의 신뢰도 계산하여 자극과 행동의 쌍들의 신뢰도를 저장하는 LT(Long Term)메모리 구조를 가지고 있다. 또한 강화 학습의 지연 보상

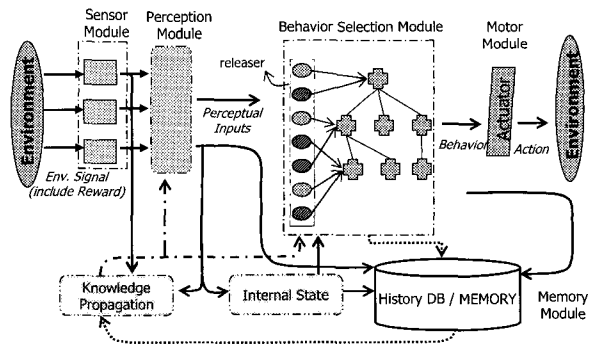


그림 1.행위 기반 제어 구조

Fig. 1. Behavior-Based Control Architecture

* 책임저자(Corresponding Author)
논문접수 : 2002. 9. 11., 채택확정 : 2003. 3. 25.
이상훈, 김봉오, 서일홍 : 한양대학교
(shlee@incorl.hanyang.ac.kr/bokim@ incorl.hanyang.ac.kr/ihsuh@hanyang.ac.kr)
※ 본 연구는 산업자원부 차세대신기술개발사업(퍼스널 로봇을 위한 제어/인식 기술)에 의하여 연구되었습니다.

(delayed reward)문제를 풀기 위해 로봇이 이미 학습되거나 설계되어있는 S-R 행동들의 지식을 이용하여 학습하는 지식 전파(knowledge propagation) 방법을 제시 하고 이를 검증하기 위해서 지연 보상만 이용한 경우, 학습할 자극의 수를 줄이는 경우 그리고 즉각 보상(immediate reward)을 위해 관찰자를 두는 경우들에 대해 비교 실험 하였으며 제시한 지식 전파 방법이 학습 속도 및 설계자가 환경이 동적으로 변하는 모든 상황에 대해 고려를 하지 않아도 됨을 알 수 있었다. 본 논문의 구성은 다음과 같다. 행위 기반 제어 구조와 강화 학습 방법에 대해 2장에서 살펴 보며 3장에서는 지연 보상 문제를 풀기 위해 제안한 지식 전파 방법을 설명하며 실험 내용은 4장에서 다루고 이어서 5장의 결론 내용으로 이루어져 있다.

II. A Behavior-Based Control and Learning Architecture

1. General Architecture

제안한 행위 기반의 제어 구조를 살펴 보면 그림 1과 같이 나타낼 수 있다.

행위 기반 제어 구조를 크게 나누어 보면 외부 환경 상태 정보를 추출하는 센서 모듈, 센서 모듈에서 얻어진 정보를 필터링하고 행위 선택 모듈로 현재의 외부 상태를 전달하는 인식 모듈과 전달된 외부 상태를 근거로 하여 행동을 선택 하는 행위 선택 모듈 그리고 선택된 행동을 수행하는 모터 모듈로 나누어 볼 수 있다. 또한 내부적으로 메모리와 학습 모듈이 동작 하게 된다.

센서 모듈(sensor module) – 센서 모듈은 실제 로봇에 장착되어 정보를 추출하는 물리적 센서와 물리적 센서의 출력을 바탕으로 로봇이 인지 할 수 있는 자극으로 바꾸어주는 논리적 센서로 구분을 하였으며, 이는 행위 기반 제어 구조의 개방성을 향상 시킬 수 있게 된다.

인지 모듈(perception module) – 인지 모듈은 센서 모듈로부터 받은 정보를 행동 선택 모듈로 전달 하는 역할을 하게 된다. 인지 모듈은 내부적으로 자극인식 필터가 있어 정보의 필터링과 가중치를 주어 전달하게 된다.

메모리 모듈(memory module) – 메모리 모듈은 ST(short-term)메모리와 LT(long-term)메모리로 구성 되며, ST메모리는 현재 시각에 결정된 행위와 행위의 대상을 매 시간 마다 저장하는 저장 공간 이며, 이 정보를 연합 학습을 통해 LT메모리에 저장 시킨다.

모터 모듈(motor module) – 선택된 행위를 수행하기 위해 직접 모터를 제어하는 역할을 하게 된다.

행위 선택 모듈(behavior selection module) – 행위 네트워크는 행위를 활성화 시키는releaser들, 로봇이 갖고 있는 행위들, releaser와 행위 간의 연결 관계, 그리고 행위들 사이의 관계를 담고 있다. 인지 모듈의 출력은 모든 releaser들에게 전달되고 releaser는 현재의 인지 모듈의 출력(로봇의 현재 상태)과 자신과 연결된 행위 사이의 관련 정도를 계산한다. 하나의 releaser는 하나의 행위에 연결되어 S-R 행동을 이루고 있다[5]. 행위의 값 결정 방법에는 releaser로의 입력만을 이용하거나, releaser의 입력뿐만 아니라 내부 상태까지 고려하여 계산하는 방법[8]-[11]이 있다. 행위의 선택 방법에는 계산된 행위들의 값들 중 최대값을 갖는 하나의 행위를 선택[18] 하거나, 일정 기준 값(threshold)보다 높은 값을 가지는 행위들을 선택하는 방법[19] 그리고 행위들이 서로 경쟁을 하여 결정 하는 방법[20]이 있다. 최대값을 선택하는 방법은 선택 후 바로 행위를 수행 할 수 있고, 기준 값보다 높은 값을 갖는 행위들을 선택하는 방법은 선택된 복수개의 행위들에서 로봇의 동작 수행을 위한 조치가 필요하다. 그리고 행위들간의 경쟁에 의한 방법은 행위 선택에 있어서 한번에 결정하지 못하고 행위 상호간의 경쟁 과정이 추가되어 상대적으로 복잡도가 증가 된다. 이러한 행위 선택의 방법들은 모두 행위의 수가 많아지게 되면 행위를 선택할 때 적은 행위의 수를 갖은 경우 보다 많은 시간이 필요하다. 또 다른 방법으로 행위들간에 우선순위를 두어 행위를 선택하는 방법도 있다. 그러나 이들 방법은 모두 로봇의 상태에 대한 행동을 선택할 때 계획이나 추론의 과정 없이 정해진 선택 알고리즘을 거쳐 바로 행위가 선택된다. 본 논문에서는 기준 값 이상의 행위 중 최대값을 갖는 행위가 선택되는 방법을 사용한다.

한편, 행위 선택 모듈의 학습을 통해서 네트워크(network)의 연결 강도(weight)가 바뀔 수 있으며, 이는 연결된 자극과 행위가 바뀔 수 있는 것이다. 그리고 학습을 통해 인지 모듈의 출력인 자극에 대해 연관된 행위가 없는 경우 자극에 적절한 행위의 연결 강도가 높아질 경우, 해당 행위가 새로이 네트워크(network) 구조에 포함될 수 있도록 하는 구조로 설계되어 있다. 여기서, 행위 선택 모듈이 자극에 적절한 행위를 학습하기 위해 강화 학습 방법이 사용된다.

2. Memory for Learning

2.1. Behavior Exploration

탐색은 로봇에 아직 기억되어 있지 않은 상태에 대한 적절한 행동을 찾기 위한 경우와 알려진 상태에 대하여 보다 최적의 행동을 찾기 위한 탐색으로 나눌 수 있다.

확인되지 못한 상태에 대한 행동을 찾기 위해서 일반적으로 임의에 행동을 취한 후 받게 되는 보상(reward)을 참조하게 된다. 임의의 행동을 취한 것에 대한 기록은 ST(short-term)메모리에 기록한다. 보상을 받은 경우 ST 메모리에 등록된 기록을 참조하여 LT(long-term)메모리로 기록을 옮겨 적는다. LT메모리에 기록을 옮겨 적는 과정에서 기존에 등록된 관계인 경우에는 신뢰도(reliability)를 갱신하게 된다. 신뢰도가 일정 기준 값(threshold)을 만족하는 행동과 상태에 대해서는 이후 행동 선택 모듈의 행동 선택 과정에 등록된다.

최적의 행동을 찾기 위한 탐색은 행동 선택 모듈에서 제시하는 행동에 대하여 epsilon 만큼의 확률로 주변 행동을 취하는 방법과 모터 모듈에서 연관된 엑츄에이터(actuator)의 입력 값을 변화시키는 방법(skill learning)이 있는바, 여기서는 epsilon 방법을 사용한다.

2.2. Short-Term and Long-Term Memory Operation

메모리 모듈이 학습에 어떠한 영향을 미치는 지를 살펴 보기로 하겠다. 메모리 모듈의 구조를 먼저 살펴 보면 그림 2와 같이 구성 되어있다.

ST메모리에는 과거의 경험을 시간 순서에 따라 저장한다. 저

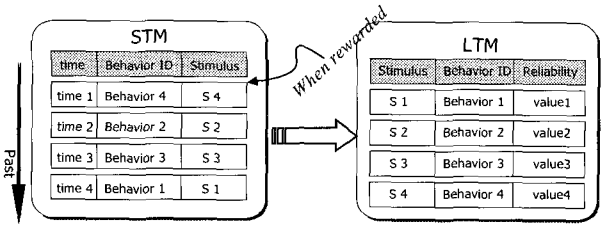


그림 2. 단기 기억 메모리와 장기 기억 메모리

Fig. 2. STM and LTM

장되는 요소는 현재 선택된 행동과 그에 영향을 주는 자극 인식 필터가 되며, 행동을 선택 할 때마다 저장하게 된다.

ST메모리에 저장된 내용은 단순히 시간에 따른 자극과 행동만이 저장될 뿐 행동에 직접 영향을 줄 수 있는 어떠한 정보도 저장되지 않는다. 이 내용이 행동에 영향을 주기 위해서는 LT메모리로 등록 된 후에 가능하게 된다. 또한 ST메모리는 유한한 기억공간을 가지며, ST메모리의 용량은 행동 선택 엔진이 기억할 수 있는 최대 개수의 행동 순서이다.

ST메모리에 저장된 내용은 강화를 받는 시점에서 LT메모리로 등록된다. LT메모리에는 자극과 행동 그리고 그 신뢰도 값이 기록되며, 신뢰도 값은 보상을 받는 빈도와 행동이 보상을 받는 시점에 대해 얼마나 과거에 이루어진 행동인가에 관한 시간적 지연도 연관된다. 따라서, LT메모리에 저장되는 내용은 보상을 받기까지의 과정을 나타내게 된다.

ST메모리의 내용을 LT메모리에 등록 시키고 LT메모리의 신뢰도(reliability)를 갱신함으로써 학습이 이루어진다. 이러한 자극과 행동간의 학습을 연관 학습(association learning)이라고 하며 고전적 조건 형성(classical conditioning)과 도구적 조건(operant conditioning) 형성의 2 가지의 종류가 있다[21][22]. 고전적 조건 형성은 두 개 이상의 자극 사이의 관련성을 학습하는 것이며, 도구적 조건 형성은 반응과 그 결과 사이의 관계를 학습 하는 것이다. 제안된 행위 기반 제어 시스템은 이러한 학습들 중 도구적 조건 형성을 하게 된다. 그림 2는 이러한 ST메모리 내용이 LT메모리에 등록 되는 과정을 보여준다.

보상을 받는 시점에서 ST메모리에 저장된 모든 내용은 보상을 받기 위한 사전 행동으로 가정한다. 보상을 받은 시점에서 시간적으로 가까운 자극과 행동의 쌍은 시간적으로 먼 자극과 행동의 쌍에 비해 중요도가 크며, 그리고 그 값은 일정한 값으로 수렴하도록 하였다. 그리고 이미 LT메모리에 등록 되어 있는 자극과 행동의 쌍은 신뢰도만 갱신을 하게 된다. 이러한 신뢰도 갱신은 (1)과 같이 한다.

$$V_{ij} = V_{ij(t-1)} + \eta \frac{1 - V_{ij(t-1)}}{d_k} \quad (1)$$

η : learning rate

V_{ij} : reliability of between stimulus and behavior at time t

$V_{ij(t-1)}$: reliability of between stimulus and behavior at time t-1

d : time difference on ST memory

i : index of stimulus

j : index of behavior

k : index of ST memory

III. Knowledge Propagation

1. Behavior Learning by Delayed Reward

일반적으로 실제 로봇에 강화 학습을 적용할 경우 보상(reward)을 즉시 줄 수 없다. 이는 로봇이 스스로 한 행동이 자기에게 어떤 영향을 미치는지를 판단을 하기가 어렵기 때문이고, 사람이 판단을 해서 보상을 준다고 해도 사람이 로봇을 지켜봐야 한다는 어려움이 존재한다. 이처럼 지연 보상(delayed reward)를 통한 학습을 하는 경우 로봇이 빠른 시간에 적절한 자극과 행동간의 관계를 학습하기가 어렵다. 이러한 지연 보상 문제를 풀기 위한 방법을 생각해 볼 수 있다.

우선 첫 번째로 로봇이 학습을 마칠 때까지 계속해서 동작을 시키는 것이다. 이러한 경우는 당연히 학습시간을 예측 할 수 없을 뿐만 아니라 적절한 학습을 할 것이란 보장이 없다. 이는 학습하고자 하는 자극에 대해 아주 많은 시행 착오(trial-and-error)를 통해 학습을 해야 하고 학습하고자 하는 자극이 나타날 수 있는 가능성도 매우 희박하기 때문이다.

두 번째는 RP(reinforcement program)을 통한 방법이다[16]. 개발자가 컴퓨터 프로그램(computer program)을 통하여 인공적인 트레이너(artificial trainer)를 제공하는 방법이다. 이 인공적인 트레이너(artificial trainer)는 즉각 보상(immediate reward)을 제공하게 되며 학습자(learner)는 인공적인 트레이너(artificial trainer)의 강화(reinforcement)에 의해 강화 학습을 하게 되는 것이다. 일반적으로 로봇에 강화학습을 적용하기 위해 RP 방법이 가장 많이 사용되고 있으나, RP 방법에 있어서 설계자는 로봇의 환경에 대해 정확히 알아야 하며 어떤 상황에서 강화를 받을 것인지에 대해 고려를 해야 한다

세 번째로는 DP(direct program)방법으로 학습해야 할 자극이 많은 경우 각 자극에 대해 알맞은 행동을 학습 할 때까지 많은 탐색(explore)과 수행을 되풀이 해야 한다. 학습의 결과(자극과 행동의 쌍) 중 일부를 제공해 줌으로써 보다 빨리 학습을 할 수 있게 하는 것이다.

우리는 지연 보상 문제를 해결하기 위해서 다음과 같은 방법을 제시하고자 한다. 로봇에 이미 학습된 어떤 자극과 행동의 쌍들이 존재하고 학습해야 할 새로운 자극에 대한 적절한 행위의 목표가 기존의 자극과 행동 쌍들의 목표와 연관 관계를 고려하는 방법이다. 만약 로봇이 어떤 자극에 대해 알맞은 행동이 학습 되어 있다고 하자. 이러한 자극은 로봇이 갖고 있는 센서에 의해 제공될 것이다. 그런데 학습해야 할 새로운 자극이 이미 학습된 자극과 행동의 쌍과 어떤 연관 관계가 있다면, 이 연관 관계를 통해서 새로운 자극을 빠르게 학습 할 수 있을 것이다. 이는 새로운 센서가 로봇에 장착되어 새로운 자극이 생길 때에 필요한 행위를 만들어야 하는 경우로 볼 수 있다. 예를 들면, 로봇은 초음파(sonar) 센서에 의해 장애물을 피하는 행동들이 학습되어 있거나 innate한 행동들로 이미 갖고 있다고 하자. 이때, 로봇에 비전(vision) 센서가 새로 장착 된 경우 로봇은 sonar센서에 의해서 장애물을 피하면서도 비전(vision) 센서로부터 새로운 자극을 동시에 받게 될 것이다. 이렇게 어떤 자극과 행동의 쌍이 새로운 자극과의 연관 관계가 있음을 찾아 새로운 자극을 학습 하는데 사용을 하여 새로운 자극을 빠르게 학습할 수 있다. 우리는 이러한 학습을 지식전파 (knowledge propagation)

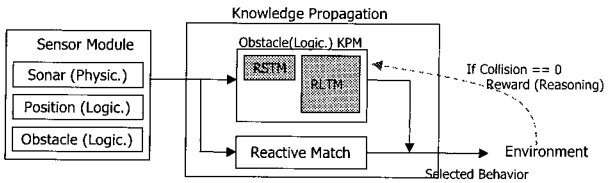


그림 3. 지식 전파 모듈

Fig. 3. Knowledge Propagation Module

라 부르기로 한다. 지식 전파(knowledge propagation)는 이미 알고 있는 지식 즉, 자극과 행동의 쌍에 관한 정보를 새로운 자극의 학습에 이용하는 것이 다. 그림 3은 이러한 지식 전파(knowledge propagation) 모듈 개념도를 보여 주고 있다.

2. Knowledge Propagation Algorithm

그림 3에서 보는 것과 같이 지식 전파(knowledge propagation) 모듈도 앞에서 설명한 그림 2와 같은 구조의 메모리 모듈을 갖고 있으며 이를 RST메모리 와 RLT메모리라 하였다. 그림 2의 ST메모리는 현재 선택된 행동과 이 행동이 선택될 수 있게 한 자극의 쌍을 저장하나 RST메모리는 학습할 자극이 존재하는 경우에 대해서 해당 자극과 행동의 쌍을 저장 한다. 이는 학습할 자극에 대해 어떤 행동을 취했는지에 대한 자극과 행동의 history가 된다. RLT메모리는 그림 2의 LT메모리와 같은 역할을 하게 된다. 그림 4는 이러한 지식 전파(knowledge propagation) 모듈에 대한 pseudo code를 보여준다.

로봇은 입력된 자극에 대해 reactive match를 통해서 S-R 행동을 하게 되며 이러한 S-R 행동의 결과로 로봇이 보상(reward)을

```
SensorModuleUpdate()
PerceptionModuleUpdate()
ReactiveMatch()
SelectBehavior()

If( existKPMPercept() )
  SaveKPMMemory( KPMPercept, CurrBehvaior )

If( receivedReward() )
{
  ReasoningKPM()
  UpdateKPMMemory()
}
```

그림 4. 지식 전파 의사 코드

Fig. 4. Pseudo Code of Knowledge Propagation

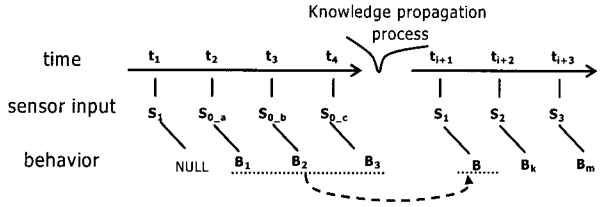


그림 5. 지식 전파 과정

Fig. 5. Knowledge Propagation Process

받았을 경우 지식 전파(knowledge propagation) 모듈에서는 reasoning을 하게 되며 이것은 RST메모리에 있는 자극과 행동에 대한 신뢰도(reliability)를 계산해서 RLT메모리의 신뢰도 값을 갱신 하는 것이다. LT메모리의 신뢰도 계산식(1)을 사용 하게 된다. 그림 5에서 보면, S0는 충분히 준비된 자극이고, S1은 학습되지 않은 자극이다. 즉, 로봇이 S0의 자극을 받으면 당연히 해당되는 S-R행동을 취하는데, 설계된 S-R행동을 갖지 않는 S1의 자극에 대해서 S0의 행동 연결과 그 결과를 전파 받아 학습된 이후에는 S1의 자극에 대한 B의 행동이 로봇의 새로운 S-R행동으로 등록된다. S1에 대한 새로운 S-R행동은 ReasoningKPM() 과정에서 결정되며, B1, B2, B3의 행위를 조절하여 만들어진다. 본 논문에서는 이동 로봇을 대상으로 실험하였기 때문에 B1, B2, B3의 수행 결과에 해당하는 이동 로봇의 상태로 S1 자극의 시점에서 빠르게 전이 될 수 있도록 이동 방향의 벡터 합을 새로운 행동으로 만들 수 있었다. 일반적인 경우를 위하여 ReasoningKPM()의 과정은 S0의 S-R행동 결과 상태로 S1 자극의 상황에서 만들 수 있는 행동에 가중치를 부여 할 수 있다.

시간의 거리에 따라서 가중치를 주어 계산 하는데 이는 학습되어 질 자극에 대해 가장 최근에 행해진 행동이 더 중요함을 나타내며, RST메모리에 동일한 자극과 행동의 쌍이 여러 번 저장되어질 경우에는 해당 자극과 행동의 쌍의 신뢰도는 한번만 갱신 한다.

IV. 실험

1. AmigoBot Robot and Control System

실험에 사용된 로봇은 ActivMedia Robotics사의 AmigoBot[23] 이며 로봇의 제어를 위해 그림 6과 같이 구성 하였다.

로봇 제어를 위해 Host PC로 Pentium 233MHz가 사용되었으며 로봇과는 900MHz 대역의 RF 모뎀을 통하여 시리얼 통신을 하며, vision 데이터는 2.2GHz 대역의 A/V receiver를 사용한다.

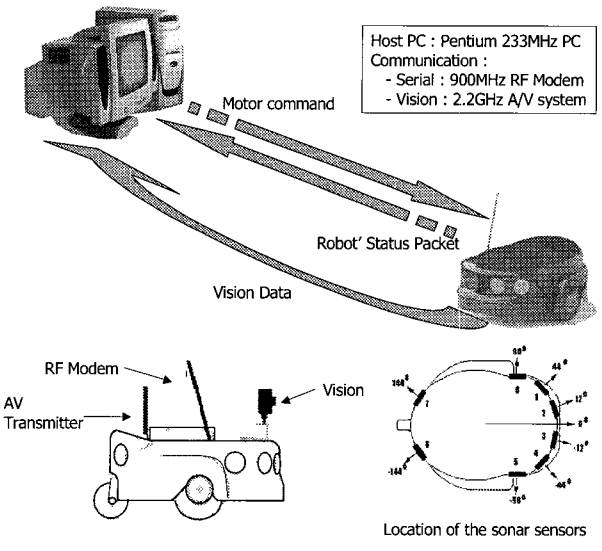


그림 6. 제어 시스템과 모바일 로봇

Fig. 6. Control System and Mobile Robot

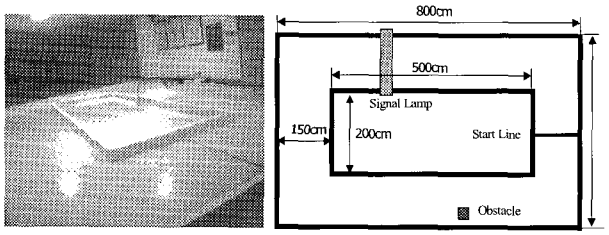


그림 7. Environment

Fig. 7. 실험 환경.

2. Learning and Knowledge Propagation

본 논문에서 제안한 행위 기반 제어 구조와 학습 알고리즘의 성능 및 유효성을 알아보기 위해서 그림 7과 같이 환경을 구성하였다.

로봇의 Task는 그림 7과 같이 구성된 환경에서 트랙 전체를 완주하는 것이다. 완주를 했을 경우 로봇은 보상(reward)을 받게 된다.

로봇 설계자가 로봇의 행위와 상태를 설계할 때 로봇이 동작하는 환경의 모든 상황을 알지 못한다는 것을 가정하기 위해 구성 환경에서 변화되는 요소와 변화되지 않는 요소를 구별하였다. 변화하지 않는 요소는 트랙 전체의 모양과 크기이며 변화되는 요소는 트랙 중간에 존재하는 신호등과 장애물이다. 실험에서는 변화되지 않는 요소만을 고려하여 로봇의 행위와 상태를 그림 8과 같이 설계 하였다.

Stimulus	Behavior
Corner	LeftTurn90
Left Wall	LeftWallAvoid
Right Wall	RightWallAvoid
Collision	Back

그림 8. 초기 지식

Fig. 8. Innate Knowledge

표 1과 같이 설계된 로봇이 트랙을 주행하면서 신호등과 장애물 회피에 대해 학습을 하여 주어진 임무를 완수 하는지 모의 실험 및 실험을 통하여 알아보았다.

표 1. 초기 지식만으로 실험한 결과

Table 1. Experiment Result in case Robot has only Innate Knowledge

	모의 실험	실험
성공 회수	5	2
전체 episode	250	100

실험 결과 표 1에 나타난 것과 같이 모의 실험 실험에서 아주 많은 episode를 수행 하였으나 로봇은 주어진 환경에 대해 학습하는 것이 불가능함을 보았으며, 실제 로봇에 적용한 실험도 같은 결과를 보였다. 이는 실제 완주 후에 보상을 받게 되어 있고 이는 앞에서도 언급했던 지연 보상(delayed reward)에 대한 문제가 발생 했기 때문이다. 우리는 앞서 이 지연 보상 문제를 해결 하기 위한 방법들을 살펴 보았으며 지식 전파(knowledge propagation) 방법을 제시 하였다.

지연 보상 문제를 풀기 위해 제시한 KP(knowledge propagation) 방법의 효용성을 보기 위해 DP(direct program), RP(reinforcement program) 그리고 지식 전파(knowledge propagation)에 의한 방법들에 대해 실험을 하고 비교 하였다. DP 방법은 신호등의 신호들에 대해서 적절한 행동을 프로그램 해주는 것이며, RP 방법은 로봇이 적절한 행동을 했을 경우에 즉각 보상(immediate reward)을 주기 위한 인공적인 트레이너(artificial trainer)를 컴퓨터 프로그램(computer program)으로 만들어 로봇이 보상을 즉시 받을 수 있도록 하는 것이며, 지식 전파(knowledge propagation) 방법을 위해서 로봇에 S-R 행동들을 설계하고 S-R 행동의 결과에 따라서 보상을 받을 수 있도록 하였다. 위 세가지 방법 모두 트랙을 충돌 없이 완주하였을 경우 보상을 받는다.

실험은 신호등의 신호는 주기적으로 신호가 바뀌게 되며 장애물은 그림 7과 같이 트랙의 오른쪽에 위치시켜 실험을 하였다. 로봇은 DP(direct program), RP(reinforcement program) 그리고 KP(knowledge propagation)의 방법에 대해 50번의 주행을 실시하였다. 표 2는 실험을 마친 후 로봇 메모리 모듈의 LT메모리의 신뢰도(reliability) 값을 기록한 것으로, 신뢰도는 자극과 행동의 연결 관계를 보여주는 것이다. 여기서 신호등과 그 외의 자극에 대한 행동의 신뢰도는 생략했다. 표 2에서 M – M 은 장애물(obstacle) 로직컬 센서의 자극이 포지션(position) 로직컬 센서와 비전(vision) 센서 모두 middle의 자극임을 나타낸다(4.3.1절 참조). 그리고 DP에서 1이란 값은 DP에 의한 S-R 행동을 나타낸다.

표 2. 장기 기억 메모리의 신뢰도.

Table 2. Reliability of Long Term Memory

Behavior ID	Obstacle Logical Sensor														
	M-M			M-R			R-L			R-M			R-R		
	DP	RP	KP	DP	RP	KP	DP	RP	KP	DP	RP	KP	DP	RP	KP
goforward1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
goforward2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
goforward3	0	0	0	1	0.33	0.73	0	0	0	0	0	0	0	0	0
goforward4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
goforward5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
stop	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
left turn 90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
right turn 90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
left turn 75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
right turn 75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
left turn 60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
right turn 60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
left turn 45	0	0	0	0	0	0	0.08	0	0	0.12	0	0	0	0	0
right turn 45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
left turn 30	0	0	0	0	0	0	0.29	0.42	0.67	0.17	0.36	0.48	0.04	0.29	0.33
right turn 30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
left turn 15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
right turn 15	1	0.18	0.27	0	0	0	0	0	0	0	0	0	0	0	0
left wall avoid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
right wall avoid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
back	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

실험 결과 자극과 행동의 신뢰도가 지식 전파(knowledge propagation)을 이용한 방법이 가장 높게 나타남을 볼 수 있다. 같은 횟수의 episode를 수행 하고서 신뢰도가 높다는 것은 학습의 진행이 빨리 이루어 졌으며 자극과 행동과의 강도가 높다는 것을 보여준다. 따라서 지식 전파(knowledge propagation) 방법이 가장 좋고 RP 그리고 DP의 순으로 볼 수 있다. 이는 지식 전파(knowledge propagation) 방법은 이미 알고 있는 지식, 즉 설계된 S-R 행동에 의해서 장애물을 피할 수 있고, 따라서 완주에 의한 양의 보상(positive reward)을 받으며, 장애물(obstacle) 로직컬 센서의 자극들에 대해서 구별할 수 있어서, 지연 보상을 받더라도

각 자극 입장에서는 즉각 보상의 효과를 갖기 때문이다. RP는 로봇이 자극에 대해서 알맞은 행동을 할 경우 인공적인 트레이너(artificial trainer)에 의해서 즉각 보상을 받으나 자극에 대해 적절한 행동을 찾기 위해 로봇이 갖고 있는 행동들에 대해서 탐색(explore)을 많이 해야 한다. DP의 경우에는 직접 프로그램을 행한 자극과 행동에 대한 S-R 행동들 이외의 새로운 자극들에 대해서는 RP와 마찬가지로 로봇이 가지고 있는 행동들에 대해서 탐색을 해야 하며 지연 보상(delayed reward)을 받기 때문에 학습이 가장 느리게 되는 것이다.

로봇을 설계할 경우 RP는 환경에 대해서 설계자가 많은 시간과 노력을 투자해 고려를 해야 하며 DP는 학습할 대상이 많은 경우 학습할 자극의 수를 줄여주는 효과는 있으나 학습 해야 할 자극들에 대해서는 여전히 지연 보상 문제가 남아 있다. 그러나 지식 전파(knowledge propagation) 방법은 로봇이 이미 학습되어 있거나 설계된 자극과 행동의 쌍들의 지식을 통하여 학습을 하여 설계자는 환경의 동적인 변화에 대해서 전부 고려를 하지 않아도 된다.

3. Preprocessing for Experiments (Details in Experiments)

본 실험을 위한 센서 모듈, 인지모듈 그리고 행위 선택 모듈의 행위들을 설계를 하여 DP(S-R program), RP(reinforcement program) 그리고 지식 전파(knowledge propagation) 방법을 이동로봇에 적용하였다. 본 논문에서는 센서를 통해 외부 환경에 대한 정보 classify하는 능력은 센서 모듈에서 갖고 있는 것으로 간주하며 로봇 스스로 모든 외부 환경 정보를 classify하는 능력은 본 논문의 범위에서 벗어 난다.

3.1. Properties of Robot

센서 모듈에서 물리적 센서로는 로봇에 장착되어 있는 초음파(sonar), 비전(vision) 그리고 엔코더(encoder) 센서를 사용하며, 로직컬 센서는 물리적 센서의 정보로부터 로봇이 이해 할 수 있는 자극으로 변환을 위해 포지션(position 로직컬 센서와 충돌 로직컬 센서 그리고 object information 로직컬 센서를 그림 9와 같이 구성하였다.

포지션(position) 로직컬 센서는 트랙에서의 위치정보를 추출하는데 이 위치 정보는 로봇의 절대 위치가 아닌 트랙의 폭에 대해서 3개의 영역으로 나누었을 경우 로봇이 왼쪽, 오른쪽, 중앙 또는 코너의 위치 중 어느 곳에 있는지를 그리고 트랙의 벽에 가까이 있을 경우에 대한 정보를 추출 한다. 그리고 object information 로직컬 센서는 color object에 대해 존재 여부와 object와의 거리를 추출하게 된다. 특히 장애물(obstacle)에 대해서는 포지션(position) 로직컬 센서와 비전(vision) 센서에서

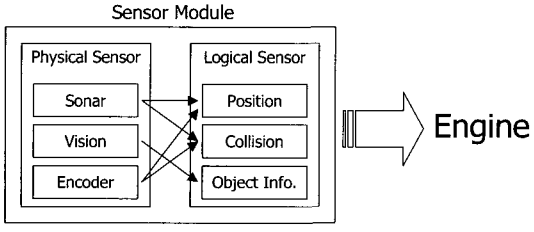
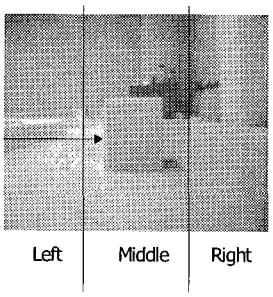
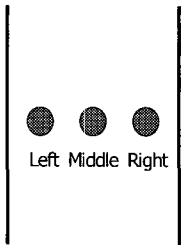


그림 9. 센서 모듈

Fig. 9. Sensor Module



(a) Vision Physical Sensor



(b) Position Logical Sensor

그림 10. 장애물 로직컬 센서

Fig. 10. Obstacle Logical Sensor.

의 장애물(obstacle) 정보를 종합적으로 사용하여 Obstacle에 대해서는 9개의 자극이 나타날 수 있도록 그림 10과 같이 구성하였다.

인지 모듈은 센서 모듈에서 받은 정보를 필터링과 가중치를 두어 행위 선택 모듈로 전달하는 역할을 하게 되며 행위 선택 모듈은 입력된 자극에 대해서 적절한 행동을 선택하게 된다. 본 실험에서는 인지 모듈의 구성요소와 행위 선택 모듈에서 선택될 수 있는 행동들을 그림 11과 같이 구성 하였다.

Percepts		
Position		Left, Right, Middle, Corner
		Left Wall
		Right Wall
Object Information	Signal Lamp	Red
		Yellow
		Blue
	Obstacle	Pos_L Vision_L
		Pos_L Vision_M
		Pos_L Vision_R
		Pos_M Vision_L
		Pos_M Vision_M
		Pos_M Vision_R
		Pos_R Vision_L
		Pos_R Vision_M
		Pos_R Vision_R
Collision		

(a) Percepts

Behaviors
Goforward1
Goforward2
Goforward3
Goforward4
Goforward5
Stop
Left Turn 90
Right Turn 90
Left Turn 75
Right Turn 75
Left Turn 60
Right Turn 60
Left Turn 45
Right Turn 45
Left Turn 30
Right Turn 30
Left Turn 15
Right Turn 15
Left Wall Avoid
Light Wall Avoid
Back

(b) Behaviors

그림 11. 인지된 자극과 행동

Fig. 11. Percepts and Behaviors

3.2. DP (S-R Program), RP, and Knowledge Propagation

3.2.1. Direct Program (S-R program) :

DP는 자극과 알맞은 행동을 지식으로 주는 것이다. DP를 이용한 실험을 하기 위해서 초기 로봇이 갖고 있는 지식 이외의 자극과 행동에 대한 정보를 추가하였다. 이는 신호등 신호 자극들에 대한 적절한 행동들과 장애물을 피하기 위한 행동들이다. 장애물을 피하기 위한 장애물(obstacle) 로직컬 센서의 출력에 대한 자극들에 대해서 적절한 행동을 학습해야 한다. 그러나 장애물(obstacle) 로직컬 센서의 출력, 즉 학습해야 할 자극의 수가 많으므로 이에 대해서 DP(S-R program)을 하였다. 위에 설명한 DP(S-R)에 대한 내용은 그림 12에 나타내었다. DP를 하지 않은 장애물(obstacle) 로직컬 센서의 자극들은 로봇이 트랙을 주행하면서 학습을 해야 하는 자극들이다.

DP를 위해서 우리들은 로봇의 사용자가 로봇 제어 프로그램을 직접 작성하지 않을 수 있게 로봇의 configuration 파일을

Stimulus		Behaviors
Signal Lamp	Red	Stop
	Yellow	Stop
	Blue	Goforward4
Obstacle	Pos_L Vision_L	Goforward3
	Pos_L Vision_M	Goforward3
	Pos_L Vision_R	Goforward3
	Pos_M Vision_L	Goforward3
	Pos_M Vision_M	Left Turn 15
	Pos_M Vision_R	Goforward3
	Pos_R Vision_L	
	Pos_R Vision_M	
	Pos_R Vision_R	

그림 12. DP 방법의 S-R 행동들
Fig. 12. S-R Behaviors of Direct Program

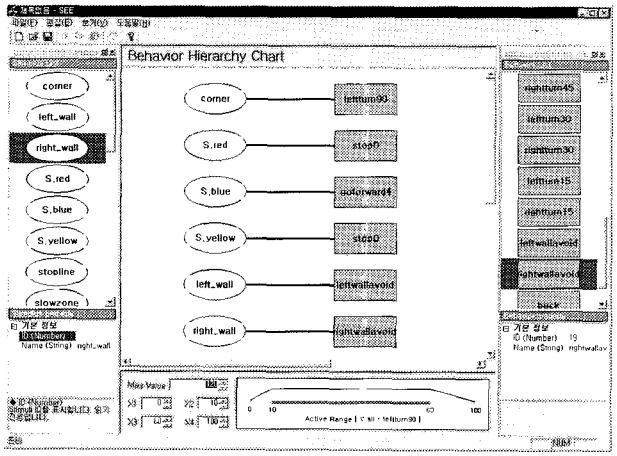


그림 13. DP를 위한 S-R 에디터
Fig. 13. S-R Editor Tool for DP

만들어주는 tool을 그림 13과 같이 개발 하였으며 로봇은 이 configuration 파일로부터 환경의 자극과 적절한 행동에 대한 S-R 지식을 가질 수 있도록 하였다.

3.2.2. Knowledge Propagation:

지식 전파(knowledge propagation)를 위해서 로봇에 innate하게 초음파(sonar) 센서와 포지션(position) 로직컬 센서의 상태에 따른 행동들을 즉, S-R 행동들을 설계를 하였다. 그림 14는 S-R 행동들 중 하나의 예를 나타낸다.

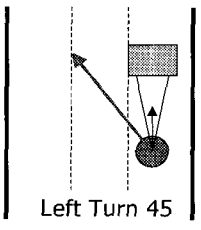


그림 14. S-R 행동의 예
Fig. 14. An Example of S-R Behavior

로봇은 S-R 행동들에 의해서 장애물을 만났을 경우 장애물을 회피를 할 수 있게 되며 이 경우에 대해서 지식 전파(knowledge propagation) 모듈은 reasoning을 통해서 S-R 행동의 자극과 행동 쌍들과 장애물(obstacle) 로직컬 센서의 자극들의 연관관계를

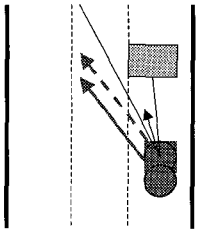


그림 15. 지식 전파로 생성되는 행동.
Fig. 15. Behavior Pattern of Knowledge Propagation

학습하게 된다. RST메모리에는 장애물(obstacle) 로직컬 센서의 출력이 존재하는 경우에만 자극과 현재 선택된 행동을 저장 하게 되며 보상(reward)을 받으면 reasoning에 의해서 RLT메모리의 자극과 행동과의 신뢰도(reliability)를 갱신하게 된다. 신뢰도가 높아 지면 S-R 행동에 의해 장애물을 피하던 로봇은 그림 15와 같이 장애물(obstacle) 로직컬 센서의 자극에 의해서 S-R 행동의 자극에 의해 장애물을 회피할 때보다 더 먼 거리에서 한번에 효과적인 방향 전환 행동을 선택했고, 따라서 S-R 행동에 의한 회피 방법 보다 지능적으로 로봇이 장애물을 회피하였다. 이것으로 지식 전파(knowledge propagation) 방법이 이뤄 졌음을 알 수 있었다.

3.2.3. Reinforcement Program [16] :

강화 프로그램(reinforcement program)을 위해서 컴퓨터 프로그램(computer program)을 통해 인공적인 트레이너(artificial trainer)를 두고 실험을 하였다. 인공적인 트레이너(artificial trainer)는 로봇이 트랙을 주행하면서 신호등과 장애물에 대해서 어떠한 행동을 해야 하는지를 알고 있고 로봇은 각 자극에 대하여 적절한 행동을 보일 경우 인공적인 트레이너(artificial trainer)에 의해서 즉각 보상(immediate reward)을 받게 되며 강화학습에 의해 로봇은 자극과 자극에 알맞은 행동을 학습하게 하였다.

V. 결론

본 논문에서는 행위 기반 제어 구조를 제안하였으며 제안된 구조는 외부의 자극을 센서모듈과 인지 모듈을 통해 자극이 행위 선택 모듈에 전달 하여 행위 선택 모듈에서는 전달된 자극에 대한 알맞은 행동을 선택할 수 있도록 하였다. 로봇은 자극에 대해서 적절한 행동을 선택해야 하는데 자극에 적절한 행동을 선택하도록 행위 선택 모듈은 강화학습을 통해 자극과 행동의 관계를 학습 하게 된다.

강화 학습을 통해 학습을 하는 로봇은 환경 또는 트레이너(trainer)가 주는 강화(reinforcement)에 의해 학습을 하게 되는데 강화 학습에서 지연(delayed reward)에 대한 문제가 발생함을 보였다. 우리는 이러한 지연 보상 문제를 해결하기 위해 지식 전파(knowledge propagation) 학습 방법을 제안하였으며 이 방법의 유효성을 DP(direct program)방법, RP(reinforcement program)방법과 비교하는 실험을 통하여 검증하였다. 지식 전파(knowledge propagation) 방법을 통하면, 로봇의 설계자가 고려하지 않은 상황에 대해서도 로봇이 이미 갖고 있는 지식, 즉 자극과 자극에 대한 적절한 행동의 쌍에 대한 정보를 사용하여 새로운 자극에

대해 학습할 수 있을 뿐만 아니라 지연 보상 상태에서도 학습이 이루어질 수 있다. 이는 환경에 대해 개발자가 모두 고려를 하지 않아도 되며 로봇은 더욱 survive할 수 있는 능력을 갖는 것이다.

로봇이 더욱 지능적인 행동 선택을 하기 위해서는 외부 환경에 대한 정보를 로봇 스스로 classify하여 새로운 자극으로 인지할 수 있는 능력이 요구된다. 로봇은 주어진 임무를 수행하기 위해 일련의 과정을 거치게 되는데 이러한 임무 수행에 대한 연속적인 행동을 학습할 수 있다면 로봇은 더욱 지능적인 행동을 보일 수 있으며 이러한 연속적인 행동의 학습에 대한 연구와 이를 실제 로봇에 적용하는 연구를 진행 중이다.

참고문헌

[1] D. B. Fogel, Evolutionary Computation, The IEEE Press, 2000.

[2] J. K George, and B. Yuan, Fuzzy Sets and Fuzzy Logic Theory and Applications, Prentice Hall, 1995.

[3] S. Maykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, 1999.

[4] R. C. Arkin, Behavior-Based Robotics, The MIT Press, Cambridge, 1998.

[5] R. R. Murphy, Introduction to AI Robotics, The MIT Press, Cambridge, 2000.

[6] R. C. Arkin, "Towards cosmopolitan robots: Intelligent navigation in extended man-made environments," Ph.D. Dissertation, COINS Tech, Rpt., 97-80, Univ. of Massachusetts, Dept. of Computer and Information Science, pp. 143-177, 1987.

[7] R. A. Brooks, "A robust layered control system for a mobile robot," *IEEE J. Robotics and Automation,* vol. RA-2, no. 1, pp. 14-23, 1986.

[8] S. D. Touretzky, and L.M. Saksida, "Skinnerbots," *Proceedings of The Fourth International Conference on Simulation of Adaptive Behavior (SAB96),* pp. 285 – 294, 1996.

[9] B. Blumberg, "Old Tricks, New Dogs: Ethology and Interactive Creatures," The Media Lab, MIT, Cambridge, Ph.D. Dissertation, 1996.

[10] S. Y. Yoon, "Affective Synthetic Characters," The Media Lab, MIT, Cambridge, Ph.D. Dissertation, 2000.

[11] J. Pauls, "Pigs and People," *Project Report, Division of Information,* University of Edinburgh, 2001.

[12] P. Maes, "The dynamics of action selection," *Proceedings of International Joint Conference On Artificial Intelligence,* Detroit, MI, pp. 991-997, 1989.

[13] A. Saffiotti, K. Konolige, and E. Ruspini, "A multivalued logic approach to integrating planning and control," *Artificial Intelligence 76,* pp. 481-526, 1995.

[14] A. F. R. Araujo, and A. P. S. Braga, "Reward-Penalty Reinforcement Learning Schema for Planning and Reactive Behavior," *Proceedings of IEEE International Conference on System, Man, and Cybernetics,* vol. 2, pp. 1485-1490. 1998.

[15] R. Genov, S. Madhavapeddi, and G. Cauwengerghs, "Learning to Navigate from Limited Sensory Input: Experiments with the Khepera Microrobot," *Proceedings of International Conference on Neural Networks,* vol. 3, pp. 2061-2064, 1999.

[16] M. Dorigo, and M. Colombetti, Robot Shaping: An Experiment in Behavior Engineering, The MIT Press, Cambridge, 1998.

[17]C. J.C. H. Watkins, "Learning from delayed rewards," Ph.d. Thesis, Cambridge University, Cambridge, England, 1989.

[18] K. Lorenz, "The comparative method in studying innate behavior patterns," *Symposia of the Society for Experimental Biology, 4,* pp. 221-268, 1950.

[19] P. Maes, "How to do the right thing," *Connection Science,* 1, 291-323, 1989.

[20] A. Ludlow, "The evolution and simulation of a decision maker," In: Analysis of Motivational Process, Academic Press, 1980.

[21] I. P. Pavlov, Selected works, Foreign languages Publishing House, Moscow, 1950.

[22] B. F. Skinner, The behavior of organisms: An experimental analysis, Englewood Cliffs, NJ: Prentice Hall, 1938.

[23] ActivMedia, AmigoBot User's Guide, ActiveMedia Robotics, 2000.

[24] R. C. Arkin, and J. Diaz, "Line-of-sight constrained exploration for reactive multiagent robotic teams," *7th International Workshop on Advanced Motion Control,* pp. 455-461, 2002.

[25] M. Likhachev, M. Kaess, and R.C. Arkin, "Learning behavioral parameterization using spatio-temporal case-based reasoning," *Proceedings of International Conference on Robotics and Automation,* vol. 2, pp. 1282-1289, 2002.

[26] J. B. Lee, M. Likhachev, and R. C. Arkin, "Selection of behavioral parameters: integration of discontinuous switching via case-based reasoning with continuous adaptation via learning momentum," *Proceedings of International Conference on Robotics and Automation,* vol. 2, pp. 1275-1281, 2002.

[27] M. Likhachev, and R. C. Arkin, "Spatio-temporal case-based reasoning for behavioral selection," *Proceedings of International Conference on Robotics and Automation,* vol. 2, pp. 1627-1634, 2001.

[28] J. B. Lee, and R. C. Arkin, "Learning momentum: integration and experimentation," *Proceedings of International Conference on Robotics and Automation,* vol. 2, pp. 1975-1980, 2001.

[29] Y. Endo, and R.C. Arkin "Implementing Tolman's schematic sowbug: behavior-based robotics in the 1930's," *Proceedings of International Conference of Robotics and Automation,* vol. 1, pp. 477-484, 2001.

이상훈

1994년 한양대학교 이과대학 수학과(이학사). 1997년 한양대학교 산업대학원 전자계산학과(공학석사). 2000년 ~ 현재 한양대학교 전자전기제어계측과 박사과정 재학중. 관심분야 : 지능로봇의 행동선택 및 학습

김봉오

1975년 3월 14일생. 2001년 경남대학교 전자공학과 졸업. 2003년 한양대학교 전자전기제어계측공학과(공학석사). 관심분야는 인공생명, 지능제어 및 로보틱스

서일홍

1977년 서울대학교 졸업. 1982년 한국과학기술원 졸업(공학박사). 1982년~1985년 대우 중공업 기술연구소 근무. 1987-1988년 미국 미시간대 객원 연구원. 1985년~현재 한양대학교 교수.