論文97-34S-8-7

강화학습과 분산유전알고리즘을 이용한 자율이동로봇군의 행동학습 및 진화

(Behavior Learning and Evolution of Collective Autonomous Mobile Robots using Reinforcement Learning and Distributed Genetic Algorithms)

李 東 昱 *, 沈 貴 寶 *

(Dong Wook Lee and Kwee Bo Sim)

요 약

자율분산로봇시스템에서 개개의 로봇은 스스로 주위의 환경과 자신의 상태를 스스로 판단하여 행동하고, 필요에 따라서는 다른 로봇과 협조를 통하여 어떤 주어진 일을 수행할 수 있어야 한다. 따라서 개개의 로봇은 동적으로 변화하는 환경에 잘 적응할 수 있는 학습과 진화능력을 갖는 것이 필수적이다. 이를 위하여 본 논문에서는 지연된 보상능력이 있는 강화학습과 분산유전알고리즘을 이용한 새로운 자율이동로봇의 행동학습 및 진화방법을 제안한다. 지연 보상능력이 있는 강화학습은 로봇이 취한 행동에 대하여 즉각적인 보상을 가할 수 없는 경우에도 학습이 가능한 방법이다. 또한 개개의 로봇이 통신을 통하여 염색체를 교환하는 분산유전알고리즘은 각기 다른 환경에서 학습한 우수한 염색체로부터 자신의 능력을 향상시킨다. 특히 본 논문에서는 진화의 성능을 향상시키기 위하여 강화학습의 특성을 이용한 선택 교배방법을 채택하였다. 제안된 방법은 협조탐색 문제에 적용하여 컴퓨터 시뮬레이션을 통하여 그 유효성을 검증한다.

Abstract

In distributed autonomous robotic systems, each robot must behaves by itself according to the its states and environments, and if necessary, must cooperates with other robots in order to carry out a given task. Therefore it is essential that each robot has both learning and evolution ability to adapt the dynamic environments. In this paper, the new learning and evolution method based on reinforcement learning having delayed reward ability and distributed genetic algorithms is proposed for behavior learning and evolution of collective autonomous mobile robots. Reinforcement learning having delayed reward is still useful even though when there is no immediate reward. And by distributed genetic algorithm exchanging the chromosome acquired under different environments by communication each robot can improve its behavior ability. Specially, in order to improve the performance of evolution, selective crossover using the characteristic of reinforcement learning is adopted in this paper. we verify the effectiveness of the proposed method by applying it to cooperative search problem.

* 正會員, 中央大學校 制御計測工學科
(Dept. of Control and Instrumentation Engineering, Chung-Ang Univ.)

※ 이 연구는 1996년도 한국과학재단 연구비지원에 의한 결과임(과제번호 : 96-0102-13-01-3)
接受日子: 1997年5月20日, 수정완료일: 1997年7月28日

Ⅰ. 서 론

최근 로봇 응용분야의 확장으로 다수의 로봇으로 구성된 시스템에 대한 연구가 많이 이루어지고 있다. 특히 자연계 생물의 특징인 자율 분산성을 가지는 로봇시스템에 관한 연구가 관심을 모으고 있다. 이러한 자율분산 로봇시스템은 중앙관리형 시스템에 비하여 다음의 몇 가지 특징을 갖는다. 첫째로 각각의 로봇은 주변의 환경이나 물체, 다른 로봇의 행동 등을 인식하여 자신의 행동을 독립적으로 결정하며, 또한 주어진 작업을 잘 수행하기 위하여 다른 로봇과 협동할 수 있다. 둘째로 자율분산로봇시스템은 강건성(robustness)과 유연성(flexibility)을 가지고 있다. 몇 대의 로봇이 고장나더라도 시스템의 정상적인 동작에 영향을 주지 않으며, 주어진 일에 대하여 오직 로봇의 행동 규칙만 바꾸어 줌으로써 여러 가지 작업에 적용할 수 있다. 셋째로 시스템의 크기가 커지더라도 개개의 로봇의 자신의 주변상황에 따라 자신의 일을 판단하여 결정하므로 시스템의 복잡도가 증가하지 않는다.

다수의 자율이동로봇에 의한 효과는 크게는 군 전체의 거동을 나타내는 군행동과 작게는 시스템내의 작업을 수행하는 협조작업으로서 나타나게 된다. 일반적으로 군 전체의 이동이나 배치 등과 같은 군행동은 센싱기능에 의해서 충분히 구현할 수 있다. 그러나 협조작업의 경우 센싱에 의하여 다른 로봇의 행동을 예측하여 작업을 수행하기 위하여는 고도의 추론능력이 필요하다. 이러한 경우 통신에 의해 자신의 상태 및 정보를 교환함으로써 쉽게 협조작업을 수행할 수가 있다.

자율분산로봇시스템에서 개개의 로봇은 사실상 다른 모든 로봇의 정보를 알 필요가 없으며 자신이 처한 환경만 인식하여 행동하면 된다. 그러나 일반적으로 동적으로 환경이 변화하는 시스템에서 로봇 스스로 협조를 위한 최선의 행동을 결정하는 것은 매우 어렵다. 최근 자연계의 생물체의 구조 및 거동을 인공적으로 연구하는 인공생명의 방법이 이와 같은 예측이 불가능하고 복잡한 문제를 해결하는데 새로운 해결책으로 기대되고 있다. 이러한 접근방식의 결과물로서 Brooks가 제안한 행동기반 로봇[1]과 Mataric이 제안한 복수 로봇의 상호작용에 의한 군행동의 실현[2] 등을 들 수가 있다. 인공생명에 대한 연구는 현재 신경망, 퍼지시스템, 진화알고리즘 및 면역시스템과 이들의 융합에 의한 방법으로 많이 연구되고 있으며 계속적으로 새로운 연구방법도 등장할 것으로 예상된다.

저자들이 생각하고 있는 인공생명에 의한 로봇은 기존의 지능로봇에 더하여 다음과 같은 세 가지의 특징을 갖는다[3]. 첫째로 문제의 수행에 있어서 사전에 짜여진 완벽한 계획보다는 예측하지 못한 문제가 발생하였을 경우 즉각적인 대처와 참여에 의해 적응 및 학습을 해나가는 능력이 있다. 따라서 행동계획은 보다 자연스럽고 유동적이게 하여 환경조건으로부터 발현될 수 있도록 한다. 이를 위하여 로봇 설계자는 완벽한 사전계획보다는 로봇 스스로가 문제를 해결할 수 있는 구조를 만들어주는 것이 필요하다. 두 번째로 개체간 또는 환경과의 상호작용에 의해 창발적인 행동이 나타난다. 세 번째로 자연계에서 개체는 다른 개체에 대한 관찰과 모방(흉내)을 통하여 학습을 한다. 이것은 개체가 부가적인 지식을 얻는 매우 실제적인 방법으로 로봇에게 적용하면 로봇이 자신의 프로그램을 학습하는데 좋은 결과를 얻을 수 있는 바탕이 될 수 있을 것이다. 그러나 대부분의 로봇은 다른 로봇이 무엇을 하는지, 또 어디로 가는지에 대한 것을 감지할 수 있는 충분히 발전된 인지능력을 가지고 있지 못하다. 이러한 한계점 때문에 관찰과 모방에 의한 학습을 실현하기 어렵다.

이러한 관점에서, 본 논문에서는 자율분산로봇시스템에서 자율적으로 행동하며 시스템의 목적을 달성하는 로봇을 실현하기 위하여 사전에 짜여진 완벽한 계획이 아닌 시스템에 적응할 수 있는 구조를 설계하여 주었다. 로봇은 주어진 환경에서 자신의 행동을 학습하기 위하여 강화학습을 이용하였고 진화를 위하여 분산유전알고리즘을 도입하였다. 주어진 일에 대하여 이와 같은 목적을 달성하기 위하여 각각의 로봇은 기본적으로 주변의 환경을 인식할 수 있는 센싱능력과 서로 통신을 할 수 있는 능력을 가지고 있다. 로봇은 센싱에 의하여 올바른 행동을 학습하고 통신을 통하여 다른 로봇과 정보를 교환함으로써 행동전략을 진화시킨다.

강화학습은 환경에 대한 사전지식이 없는 경우 강화신호에 의하여 행동을 학습시켜 나가는 방법[4]으로 Sutton의 TD method에 의한 Actor-critic 구조[5]와 Watkins의 Q-learning[6] 등이 있다. 이 방법들은 현재에 즉각적인 보상(강화신호)이 없는 경우 강화신호를 예측하여 학습하는데 반하여, 본 논문에서는 강화신호의 예측대신 강화신호를 받은 순간 이전의 행동에 대하여 Q-값을 수정하는 지연보상이 있는 Q-학습법을 제안하였다. 한편 진화를 위하여 개개의 로봇이 하나의

염색체를 가지며 통신을 통하여 선택 및 교배를 하는 분산유전알고리즘을 도입하였다[7]. 단순 유전 알고리즘에서는 적합도의 평가, 선택, 교배 및 돌연변이의 과정이 일괄적으로 이루어진다. 그러나 분산유전알고리즘은 이러한 연산이 각 개체에 대하여 분산적으로 이루어진다. 이것은 진화 알고리즘의 연속세대 모델에 해당하는 것으로 각각의 개체(본 논문에서는 로봇이 됨)는 적합도의 평가 능력과 선택, 염색체의 교배 및 돌연변이의 기능을 갖추어야 한다. 이러한 진화의 과정에 의해 로봇은 다른 환경에서 획득된 우수한 로봇의 염색체를 받아들여 자신의 수행능력을 향상시킨다. 이것은 다수의 로봇에 의한 협조 학습(진화)의 과정으로 볼 수 있다. 참고적으로 문헌 [8] 에서 제시한 분산실행 가능한 유전알고리즘은 염색체로 상태천이 함수를 사용하기 때문에 동적으로 변화하는 환경에 대한 적응 능력에 한계가 있으며, 로봇의 학습을 오로지 진화에만 의존하고 기존의 교배방법을 그대로 사용하기 때문에 우수한 로봇을 찾아낼 수는 있지만 역으로 성능이 매우 떨어지는 로봇도 동시에 발생하여 전체적으로 로봇군의 성능을 개선하는 데에는 불리하다. 따라서 본 논문에서는 실제적으로 적용할 수 있는 로봇 모델의 제시와 함께 강화학습의 특성을 이용한 새로운 교배방법을 제안하여 진화의 효율을 향상시켰다. 교배를 위한 로봇의 염색체는 현재까지 학습된 정보이고, 로봇은 자신보다 우수한 로봇과 마주쳤을 경우 지역적 통신을 이용하여 염색체를 받아온다. 이와같이 통신을 이용한 로봇의 진화는 로봇 자신이 직접 경험하지 못한 상태에 대한 정보도 획득할 수 있다. 이러한 진화방법은 생물체가 관찰과 모방에 의해 학습하는 것과 같은 효과를 갖는다.

제안된 방법은 협조탐색 문제에 적용하여 컴퓨터 시뮬레이션을 통하여 그 유효성을 검증한다.

Ⅱ. 자율이동로봇의 센싱 및 통신 시스템

자율분산로봇시스템에서 협조행동을 통하여 복잡하고 어려운 문제를 수행하기 위해서는 통신을 사용하는 것은 필수적이다. 일반적으로 통신은 어떠한 로봇과도 통신이 가능한 전역적 통신과 주변의 로봇과만 통신이 가능한 지역적 통신으로 나눌 수 있다. 전역적 통신은 로봇의 수가 적을 경우 효과적이지만 로봇의 수가 증가함에 따라 통신자원의 한계와 다룰 수 있는 정보량이 증가하며 통신간섭이나 부적절한 정보의 전달의 문제가 발생하기 쉽기 때문에 복잡한 일에 적용하기 어렵다. 따라서 본 논문에서는 각각의 로봇이 국소적으로 정보를 전달하는 지역적 통신시스템을 채택함으로써 정보의 범람뿐만 아니라 통신의 복잡도도 막는다.

본 논문에서는 적외선 센서에 의해 센싱 및 통신을 수행한다. 따라서 로봇은 다른 로봇이나 장애물로부터의 거리를 측정할 수 있고 적외선 펄스열로서 정보를 전달할 수 있다. 로봇은 다른 로봇과 마주칠 경우 로봇간의 통신이 행하여진다(지역적 통신). 또한 로봇은 각 방향에 식별센서(예를 들면 칼라센서)를 가지고 있어서 감지된 물체의 거리뿐만 아니라 종류도 구별할 수 있다. 각각의 로봇은 사인보드 모델의 형태로 주변에 자신의 정보를 전파하고 만일 자신보다 적합도가 높은 로봇을 만났을 경우 정보전달 모델의 방법으로 통신한다. 그림 1은 본 논문에서 사용한 자율이동로봇을 위한 센싱과 통신 시스템을 나타낸다[9].

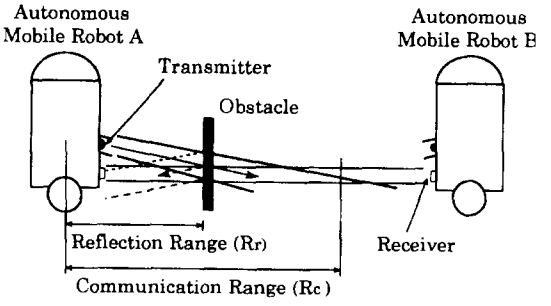

그림 1. 자율이동로봇의 센싱 및 통신 시스템
Fig. 1. Sensing and local communication system of autonomous mobile robot.

Ⅲ. 행동학습 및 진화시스템의 구조

그림 2는 본 논문에서 제안한 자율분산로봇시스템의 행동학습 및 진화를 위한 시스템의 개념도를 나타낸다. 각각의 로봇은 다수의 상태-행동 규칙을 테이블의 형태로 가지고 있으며, 테이블의 값은 행동 결과에 의해 주어진 보상이나 벌칙에 따라 제안한 Q-학습의 방법(4장 내용)으로 갱신해 나간다. 만일 로봇이 자신보다 우수한 로봇을 만났을 경우 이 로봇은 통신을 통해 상대방의 행동규칙을 획득하고, 유전 알고리즘에 의해 자신의 행동규칙을 진화해 나간다. 이와 같이 학습 및 진화능력에 의해 로봇은 주어진 환경에 적응하여 주어진 목적을 수행한다. 다음의 4장과 5장에서는 이를 실현하기 위한

구체적인 방법에 대해서 서술한다.

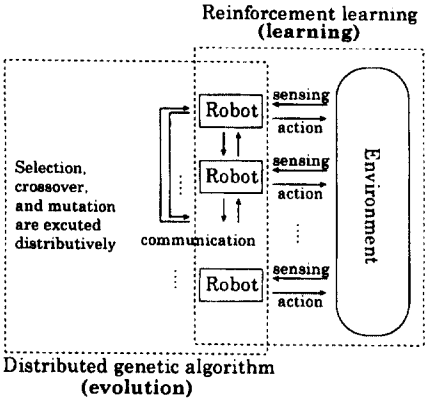

그림 2. 자율분산로봇시스템의 행동학습 및 진화의 개념도

Fig. 2. Conceptual diagram of behavior learning and evolution for DARS.

Ⅳ. 강화학습 기반의 행동학습

1. 강화학습

자신과 환경과의 상호관계와 이에 따른 강화신호를 통하여 자신의 행동을 개선해 나감으로써 보상을 최대로 받도록 하는 것이 강화학습의 목적이다[4]. 이러한 강화학습법은 환경에 대한 정확한 사전의 지식이 없이 학습 및 적응성을 보장하기 때문에 로봇의 학습에 많이 적용되고 있다.

한편 Q-학습은 확률적 동적 계획법에 기반을 둔 자유 모델(model free) 강화 학습법의 하나로서 개발되었다. 이 방법은 마코브 환경(markovian environment) 하에서 학습능력을 가진 로봇이 최적으로 행동할 수 있도록 해준다.

이산된 환경의 상태 집합을 S, 행동 집합을 A라고 놓았을 때 기본적인 Q-학습 알고리즘[6]은 다음과 같다.

[Q-학습 알고리즘]

1. 모든 상태 s와 행동 a에 대하여 $Q(s,a)$를 임의의 값(일반적으로 0)으로 초기화한다.
2. 현재의 상태 s를 인식한다.
3. 상태-행동 규칙에 따라 행동 a를 선택한다.
4. 주어진 환경에서 행동 a를 수행하고, 다음상태를 s' 즉각적인 보상을 r로 놓는다.
5. s, a, s', 그리고 r로부터 상태-행동 규칙을 갱신한다.

$$Q_{t+1}(s,a)=(1-\alpha_t)Q_t(s,a)+\alpha_t(r+\gamma \max_{a'\in A} Q_t(s',a')) \quad (1)$$

이때 α_t는 학습률이고, γ는 0과 1사이의 고정된 감쇠계수이다.

6. 2로 되돌아간다.

본 논문에서는 로봇이 학습하는 방법으로 기본적으로 Q-학습을 사용한다. 다음의 **4.2~4.4**절에서 설정한 문제에 대하여 **4.5**절의 지연보상이 있는 Q-학습법을 사용함으로서 로봇의 학습을 실현한다.

2. 협조행동을 위한 문제 설정

여러 대의 자율이동로봇으로 구성된 자율분산로봇시스템을 이용하여 할 수 있는 일은 산업현장에서의 자재운반, 해저나 우주에서의 자원채취, 발전소·화학 플랜트 등 극한환경에서 배관 검사 및 수리 등 매우 다양하며 앞으로는 기술의 발전과 더불어 응용분야도 점점 확대될 전망이다.

그러나 본 논문에서는 제안한 시스템의 유효성을 검증하기 위하여 시스템의 목적을 여러 대의 로봇에 의한 물체의 협조탐색으로 정하였다. 따라서 주어진 환경에서 로봇은 장애물이나 로봇 서로간의 충돌을 회피하며 물체를 획득해야 한다.

3. 로봇의 상태

로봇은 다른 장애물과 로봇, 그리고 획득할 물체를 구분할 수 있는 식별 센서 및 통신과 거리 측정을 위한 적외선 센서를 각각 여덟 개씩 가지고 있다. 이때 정면의 센서번호를 0번(S_0)으로 하고 반시계 방향으로 번호가 증가한다고 했을 때, 센서의 입력에 따른 로봇의 상태를 나타내면 표 1과 같다. 로봇이 가질 수 있는 총 상태는 $81(3^4)$상태이다.

표 1. 로봇의 상태

Table 1. Robot's states.

센서	앞쪽 센서 (S_0)	오른쪽 센서 (S_1,S_2,S_3)	왼쪽 센서 (S_5,S_6,S_7)	뒤쪽 센서 (S_4)
상태	없음 물체 장애물	없음 물체 장애물	없음 물체 장애물	없음 물체 장애물

표 1에서 장애물은 획득하고자 하는 물체를 제외한 나머지, 즉 고정된 장애물과 다른 로봇을 포함하여 말한

다. 또한 오른쪽과 왼쪽의 센서에 물체와 장애물이 있을 경우, 예를 들면 S_2에 물체가 감지되고 S_3에 장애물이 감지된 경우, 상태는 물체가 있는 상태로 간주한다. 그 이유는 그 다음의 행동이 좌 또는 우회전일 경우 이 회전은 물체가 감지된 센서방향 쪽으로 회전을 하도록 설정하여 물체를 향할 수 있기 때문이다.

4. 로봇의 행동

로봇이 취할 수 있는 기본적인 행동은 다음과 같다.

- 좌로 회전
- 우로 회전
- 뒤로 회전
- 목표물(물체, 장애물, 로봇)을 향해 전진. 단, 목표물이 없을 경우에는 직진한다.

5. 제안한 Q-학습 법

본 시스템에서 로봇의 행동에 대한 보상이나 벌칙은 주로 이전 행동의 영향에 의하여 받은 것이다. 또한 현재 받은 보상이나 벌칙은 바로 다음에 이어지는 상태에는 별로 영향을 미치지 않는다. 따라서 학습은 보상이나 벌칙을 받은 시점에서 과거의 행동에 대하여 수행하는 지연보상이 있는 Q-학습법을 사용하였다. 즉, 현재의 행동결과는 과거행동의 영향을 받았다는 가정 하에 감쇠계수 $\gamma(0<\gamma<1)$를 사용하여 보상이나 벌칙을 받은 순간, 현재로부터 과거의 행동에 대하여 Q-값을 갱신한다. 실제적으로는 k값이 커짐에 따라 γ의 값은 0에 지수적(γ^k)으로 가까워지기 때문에 근사적으로 과거의 한정된 스텝에 대하여만 계산하여도 된다. 강화신호를 받은 시점에서 Q-값의 학습이 이루어지므로 일단 현재의 행동에 대하여 아무런 보상이 없다면 Q-값은 갱신되지 않는다. 그러나 보상이나 벌칙을 받지 못하는 행동을 계속 할 경우 학습이 수행되지 않기 때문에 연속된 일정한 시간동안 보상이나 벌칙이 없으면 그 기간의 행동들에 대하여 작은 벌칙을 주어 다른 방향으로 학습이 계속되도록 할 수 있다.

다음은 본 논문에서 제안한 지연보상이 있는 Q-학습법이다.

[지연 보상이 있는 Q-학습법]

1. 모든 상태 s와 행동 a의 $Q(s,a)$에 대하여 0보다 크고 1이하의 임의의 값으로 초기화한다(초기값이 0일 경우 아래의 (2)식에서 행동선택이 안됨).
2. 현재의 상태 s를 인식한다.
3. 다음의 행동 선택확률 $P(a)$에 의해 행동 a를 선택한다.

$$P(a)=\frac{Q(a,s)^{\frac{1}{T}}}{\sum_{a'\in A}Q(a',s)^{\frac{1}{T}}} \tag{2}$$

단, T는 탐사시간을 제어하는 온도변수로 학습이 진행될수록 감소하여 확률적인 선택의 요소를 줄이는 역할을 한다.

4. 주어진 환경에서 행동 a를 수행한다.
또한 지연된 보상 r_d를 계산한다.(지연된 보상은 즉시 계산되지 않을 수도 있다.)
5. s, a, r_d로부터 Q-값을 갱신한다.

$$Q_{t+1-k}(s,a)=(1-\gamma^k\alpha_t)Q_{t-k}(s,a)+\gamma^k\alpha_t(0.5r_d+0.5) \tag{3}$$

이때 k는 0에서 t의 값을 가지며 α_t는 학습률, γ는 감쇠계수이다.
r_d가 즉시 계산되지 못할 지라도, 일단 r_d가 계산되면 Q-값은 이전의 행동에 대하여 모두 갱신된다.

6. 2로 되돌아간다.

제안한 지연보상이 있는 Q-학습법은 계속적으로 보상을 주지 못하는 경우 학습이 되지 않는 점을 보완하기 위하여 보상이 있는 시점에서 과거의 행동에 대하여 학습을 하는 방법이다. 이때 (2)식과 같은 행동선택 확률을 사용하여 초기의 Q값을 0보다 큰 임의의 값으로 하였으며 Q값의 갱신식도 (3)식과 같이 변형하였다. 또한 (2)식에서 온도계수 T를 도입하여 학습이 진행될수록 행동선택의 확률요소를 줄일 수 있도록 하였다. 즉, T값이 작아질수록 한 상태 s에 대하여 각 행동의 $P(a)$의 값의 차이가 더욱 커지게 되므로 Q값이 가장 큰 행동을 취할 확률이 점점 커진다. 즉 초기에는 여러 가지 행동이 발생할 확률을 높여 학습의 효과를 높이며 학습이 진행된 후에는 이미 학습된 결과를 이용하도록 하는 방법을 사용하는 것이다.

기본적인 Q-학습법의 수렴성에 대하여는 이미 참고문헌 [6] 등에서 증명이 되어 있으며, 본 논문에서 제안한 지연보상이 있는 강화학습법도 마찬가지로 방법으로 수렴성이 보장될 수 있다.

V. 분산유전알고리즘에 의한 행동진화

1. 분산유전알고리즘

분산유전알고리즘은 유전 알고리즘의 변형된 것으로서 실행방법에 따라 다음의 세 가지 형태가 있다.

첫째는 개체군을 여러 개의 군으로 나누어 다른 컴퓨터에서 진화하면서 각각의 개체군을 통합하는 방법[10]이고, 둘째는 하나의 개체(agent)가 하나의 염색체가 되며 각 개체간 통신에 의한 방법 등으로 일괄적이 아닌 분산적으로 진화하는 방법[7][8], 셋째는 염색체를 여러 개의 부분으로 나누어 개체(agent)에 할당하고 임무 수행 후 다시 개체를 합쳐 하나의 염색체로 재구성하는 방법[11][12]이다.

본 논문에서는 두 번째의 방법을 사용하여 로봇의 진화를 실현하였다. 이 방법은 진화의 대상인 염색체가 하나의 로봇이 됨으로서 여러 대의 로봇으로 구성되어있는 자율분산로봇시스템에 실제적으로 적용하여 각각의 개체인 로봇이 시스템의 목적(예를 들면 협조행동을 통한 작업의 완수)에 맞도록 진화를 시킬 수 있는 장점이 있다.

단순 유전 알고리즘에서는 적합도의 평가, 선택, 교배 및 돌연변이의 과정이 일괄적으로 이루어진다. 그러나 분산유전알고리즘은 이러한 연산이 각 개체에 대하여 분산적으로 이루어진다. 이것은 진화 알고리즘의 연속세대 모델에 해당하는 것으로 각각의 개체(본 논문에서는 로봇이 됨)는 적합도의 평가 능력과 선택, 염색체의 교배 및 돌연변이의 기능을 갖추어야 한다. 이러한 진화의 과정에 의해 로봇은 다른 환경에서 획득된 우수한 로봇의 염색체를 받아들여 자신의 수행능력을 향상시킨다. 이것은 다수의 로봇에 의한 협조 학습(진화)의 과정으로 볼 수 있다.

◆ 염색체(Chromosome)

로봇이 진화의 대상으로 하는 것은 자신이 가지고 있는 염색체이다. 본 논문에서는 로봇이 현재까지 학습한 데이터인 Q-테이블의 값을 염색체로 하였다. 이 Q-테이블의 값은 로봇이 환경에 대응하여 학습한 결과로서 로봇마다 각자 학습한 다른 값을 가지고 있으며 진화의 대상으로 하기에 적당하다. 따라서 염색체는 실수치의 연속으로 구성되어 있는 Q-값이 된다.

◆ 선택(Selection)

환경에 대하여 평가도 받아보지 못한 로봇이 선택되는 것을 방지하기 위하여 교배 후 최소한 일정한 시간(T_{eval} : 평가시간)이 지난 로봇에 대하여 선택 될 수 있는 자격을 부여한다. 만약 어떤 로봇이 자신보다 우수한 로봇을 만나면 그 로봇을 선택하여 유전자를 받아오고 자신의 유전자와 교배를 하여 새로운 유전자를 만들어 낸다. 물론 선택의 과정은 로봇의 지역적 통신에 의해 이루어진다.

2. 교배(crossover)

염색체가 Q-테이블이기 때문에 하나의 상태와 그 상태에서 취할 수 있는 행동의 집합을 하나의 유전자(gene)로 하였다. 따라서 유전자의 총 수는 로봇이 가질 수 있는 총 상태의 수인 81개가 된다.

일반적으로 두 개의 부모개체를 선택하여 교배하면 새로 생기는 자식의 개체는 두 부모의 특성을 함께 가지게 된다. 이때 두 부모개체의 형질은 새로 생겨난 자식들에게 유전됨으로서 일단 선택이 된 부모의 염색체는 소실되지 않고 두 개의 자식에게 나누어 유전된다. 그러나 본 논문에서 사용하는 분산유전알고리즘은 하나의 로봇은 선택에 의해 가져온 염색체와 자신의 염색체를 합쳐 새로운 하나의 염색체를 재생산하여 자신의 염색체로 치환하므로 두 부모의 염색체 중 절반은 소실된다. 따라서 이러한 교배에 의하여 우수한 개체가 소실될 가능성도 존재한다. 이러한 점을 보완하기 위하여 본 논문에서는 개선된 교배방법을 제안하였다.

교배의 방법은 기본적으로 일정교배(uniform crossover)와 유사하다. 그러나 강화학습의 특성을 살리기 위하여 임의로 발생한 0과 1의 마스크를 사용하여 교배를 하는 기존의 방법과는 달리 각 유전자는 현재까지 학습한 횟수를 저장하고 있어서 이 횟수에 비례하여 두 로봇의 유전자 중 하나의 유전자를 택한다. 결국 학습이 많이 된 유전자에 대하여 선택될 확률을 높여서 좋은 유전자의 소실을 막을 수 있도록 하였다.

부모개체의 두 염색체를 x(유전자)와 q(학습된 횟수)의 쌍으로 표현하면 로봇 1과 로봇 2의 염색체는 (4)식과 같이 나타낼 수 있다.

$$(\vec{x}^1, \vec{q}^1) = ((x_1^1, \cdots, x_n^1), (q_1^1, \cdots, q_n^1))$$
$$(\vec{x}^2, \vec{q}^2) = ((x_1^2, \cdots, x_n^2), (q_1^2, \cdots, q_n^2)) \quad (4)$$

단, n은 총 유전자의 개수

이때 새로운 교배 방법에 의하여 생성되는 염색체는 (5)식과 같이 나타낼 수 있다.

$$(\vec{x}, \vec{q}) = ((x_1^{s_1}, \cdots, x_n^{s_n}), (q_1^{s_1}, \cdots, q_n^{s_n})) \quad (5)$$

단, $s_i = \begin{cases} 1 & p_i < \dfrac{q_i^1}{q_i^1 + q_i^2} \\ 2 & else \end{cases}$,

$i = 0 \cdots n$,

p_i = 0과 1사이의 임의의 난수이다.

즉, 유전자의 학습된 횟수(q)를 고려하여 부모 1과 2의 염색체를 유전 받는다.

3. 적합도 함수(Fitness function)

적합도 함수는 진화의 방향을 결정하는 가장 중요한 파라메타이다. 실제적으로 이 적합도 함수에 의하여 로봇들이 원하는 행동이나 협조행동을 하도록 진화해 간다. 뿐만 아니라, 적합도의 값은 서로 다른 로봇을 선택하는 기준이 된다. 본 논문에서는 협조탐색의 문제로서 충돌을 피하면서 많은 물체를 획득하는 것을 목표로 하고 있으므로, 물체를 획득하였을 경우 적합도가 상승하고, 장애물이나 로봇에 충돌하였을 경우 적합도가 떨어진다. 여기에서 로봇의 적합도는 최종 T_{eval}시간 동안 받은 보상이나 벌칙에 의해 (1)식과 같이 표현할 수 있다. 여기서 T_{eval}시간은 교배 등에 의해 염색체가 바뀐 후 최소한의 평가를 받는 시간이며 모든 로봇이 동등한 조건에서 평가를 받을 수 있도록 과거 T_{eval}시간 동안 계산된 적합도를 가지고 선택을 위한 판단을 할 수 있도록 하였다. 또한 (6)식의 세 번째 항은 취하는 행동에 대하여 소비되는 에너지의 양이 다른 경우 도입할 수 있다.

$$fitness = \alpha \times \#\text{ of reward} - \beta \times \#\text{ of penalty} - \gamma \times \text{energy consumption} \quad (6)$$
(during last T_{eval} times)

단, α, β, γ는 중요도를 나타내는 비례상수이다.

적합도 함수에 의하여 로봇의 진화의 추이를 변화시킬 수 있다. 예를 들어 (6)식에서 α의 값을 상대적으로 크게 하면 충돌이나 에너지 소비를 작게 하는 개체보다는 물체획득에 뛰어난 개체가 많이 생겨난다. 그러나 초기에 학습의 영향이 클 때에는 진화의 방향도 세 가지 항목에 대하여 비슷하게 이루어지지만 시간이 지남에 따라 학습의 영향이 줄어들게 되므로(온도계수 T의 감소에 의해) 점점 진화의 영향이 증대된다.

Ⅵ. 시뮬레이션 결과

1. 협조탐색을 위한 초기 환경

본 논문에서는 자율이동로봇군의 행동학습과 진화를 위하여 로봇군의 협조탐색 문제의 하나인 다수 로봇에 의한 물체획득 문제로 설정하였으며 모의 실험을 위하여 다음과 같은 환경 조건을 설정하였다.

(1) 로봇의 수 : 25대
물체의 수 : 500개
장애물의 수 : 100대
1회의 수행시간 : 1000단위시간
평가시간(T_{eval}) : 500단위시간
작업공간 : 20×20m(로봇의 크기는 5×5cm)
(1단위시간 동안 로봇은 회전 또는 2.5cm 전진할 수 있다)
통신반경 : 75cm(참고문헌 [9]에서 제안된 방법으로 산출)
센싱반경 : 32.5cm(통신반경의 절반)

(2) 물체 및 장애물은 작업공간내에 골고루 퍼져 있다.

(3) 우선, 모든 로봇은 작업공간에서 다른 로봇과의 거리를 충분히 유지하도록 흩어진 후 작업을 수행한다.

(4) 물체의 정확한 위치는 8개의 센서에 의해 45도 범위 단위로 대략적으로만 알 수 있으므로 로봇의 회전각도를 45도의 배수로 한다.

매회의 수행 때마다 물체의 수를 다시 회복시키고 학습과 진화를 계속하였다. 그림 3은 자율이동로봇군에 의한 물체의 협조탐색을 나타낸 그림이다.

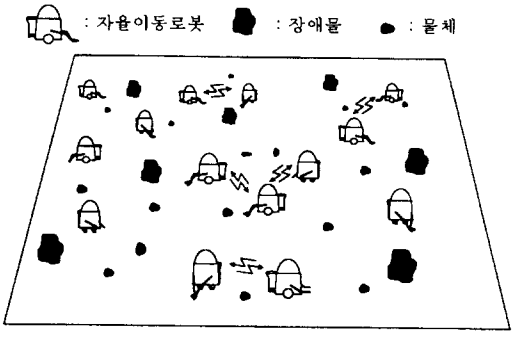

그림 3. 자율이동로봇군에 의한 협조탐색
Fig. 3. Cooperative search of collective autonomous mobile robots.

제안된 방법들의 유효성의 검증을 위하여 모의 실험은 ① 학습과 진화를 하지 않을 경우 ② 학습만을 수행할 경우 ③ 진화와 학습을 동시에 수행할 경우에 대하여 수행하여 결과를 비교하였으며, 적합도 함수에서 ⓐ α의 비중을 크게 한 경우와 ⓑ β의 비중을 크게 한 경우에 대하여 진화의 추이를 비교하였다.

2. 시뮬레이션 결과

그림 4는 시행회수에 따른 매회(1회 : 1000단위시간) 적합도의 총합을 나타낸 그림이다. 학습과 진화를 하지 않은 경우 로봇은 자신이 취할 수 있는 행동에 대하여 같은 비율로 임의의 행동을 취하므로 우연히 물체 앞에 도달하였을 경우만 물체를 획득한다. 반면 학습을 수행한 경우는 시간이 지남에 따라 물체를 획득 및 장애물 회피의 행동을 꾸준히 학습하고 있다. 또한 학습과 진화를 동시에 수행한 경우는 학습만 수행한 경우에 비하여 좋은 성능을 보여주고 있다. 이것은 진화를 통하여 로봇이 학습하지 못한 상태에 대한 정보도 가질 수 있게 됨으로 성능의 향상을 가져온 것으로 볼 수 있다. 학습과 진화를 동시에 시행한 경우 모든 로봇이 획득한 물체는 초기 시행때는 200개에서 400개 까지 꾸준히 향상되며 충돌횟수는 초기에는 125회 내외에서 50회 정도로 감소하였다.

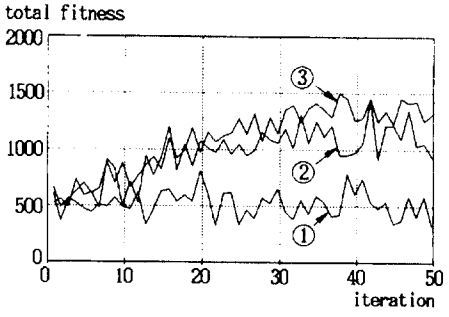

그림 4. 시행회수에 따른 적합도 합의 변화(α, β=5)
Fig. 4. Relationship between total fitness variation and iteration numbers(α, β=5).

그림 5와 6에서 획득한 물체에 대하여 적합도 산출시 α의 비중을 크게 하였을 경우(ⓐ : α=20, β=1) ⓑ의 경우에 비해 장애물 회피보다는 물체의 획득에 우수한 더 능력을 나타냈으며, β의 비중을 크게 하였을 경우(ⓑ : α=1, β=20)는 그 반대의 경우가 나타났다. 이러한 차이는 적합도가 다른 로봇의 선택의 기준이 되기 때문에 적합도의 특성에 맞는 로봇이 주로 선택된 결과이다. 이때 뚜렷한 진화의 영향은 학습이 어느 정도 진행된 후에 나타난다.

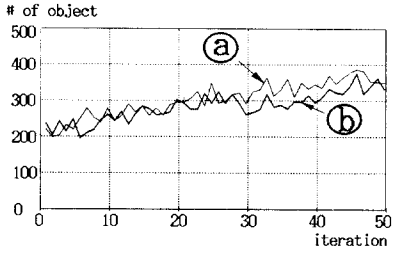

그림 5. 시행회수에 따른 획득한 물체의 수
Fig. 5. Relationship between object obtain numbers and iteration numbers.

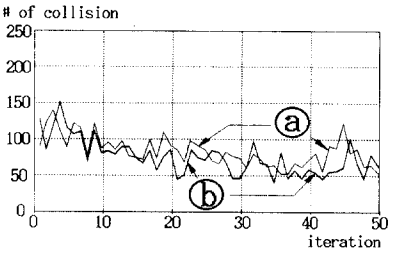

그림 6. 시행회수에 따른 충돌회수
Fig. 6. Relationship between collision times and iteration numbers.

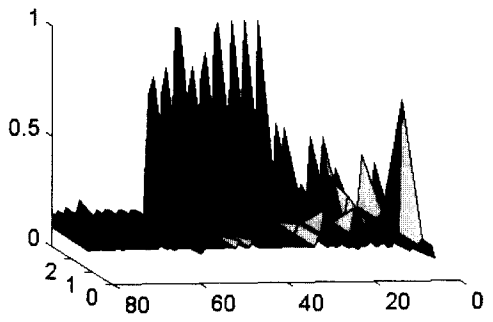

그림 7. 최종 얻어진 Q 값
Fig. 7. Finally obtained Q-value.

그림 7은 50회 수행 후 진화된 로봇 중 임의의 한 로봇의 학습된 Q값을 나타낸 그림이다. 그림에서 알 수 있듯이 주어진 상태에서 유익한 행동의 Q값이 매우 크게 성장해 있음을 볼 수 있다. 또한 상태 54에서 80은 뒤쪽 센서(S_4)에 물체가 감지되는 경우로 이러한 경우는 앞에 있는 물체를 취하지 않고 180도 회전 할 경우

에만 발생하는 상태이다. 따라서 점점 학습이 진행됨에 따라 그와 같은 경우는 거의 발생하지 않기 때문에 학습이 되지 않았음을 보여주고 있다.

Ⅶ. 결 론

본 논문에서는 다수의 로봇으로 구성된 자율분산로봇 시스템에서 로봇의 행동학습 및 진화를 위하여 강화학습과 분산유전알고리즘을 도입한 방법을 제안하였다. 각각의 로봇은 주변을 인식하여 자신의 행동을 결정하며, 이때 지연보상이 있는 Q-학습법을 제안하여 적용하였고, 지역적 통신시스템을 이용하여 시스템의 목적에 맞도록 진화해 나가는 방법을 사용하였다. 또한 시뮬레이션 결과로부터 학습과 진화의 유효성을 검증하였다.

다수의 로봇으로 구성된 시스템에서 동적인 환경의 변화를 고려하여 로봇의 행동 규칙을 정하는 것은 쉽지 않다. 따라서 최근 많은 연구자들은 고전적인 인공지능 접근방식 대신에 인공생명 접근방식을 택하고 있다. 특히 강화학습법을 포함한 신경회로망, 유전 알고리즘, 퍼지 시스템 등과 이들의 융합에 관심을 가지고 있다. 본 논문에서는 로봇에게 완전한 프로그램을 만들어 주는 대신 동적으로 변화하는 환경에 대하여 유연하게 대처할 수 있는 행동이 발현되고 진화해 나갈 수 있는 제어구조를 만들어 줌으로써 협조행동을 실현하였다. 제안한 방법의 유효성을 검증하기 위하여 본 논문에서는 비교적 간단한 문제에 적용하였지만 앞으로는 자율이동로봇 및 마이크로 로봇의 기술 발달과 더불어 적용할 수 있는 분야는 계속 늘어날 전망이며 제안한 방법은 이를 실현할 수 있는 기본전략으로 사용될 수 있을 것으로 기대한다.

참 고 문 헌

[1] R.A. Brooks, "Behavior Humanoid Robotics," *Proc. of Int. Conf. on IROS*, pp. 1-8, 1996.

[2] M.J. Mataric, "Designing Emergent Behaviors : From Local Interactions to Collective Intelligence," *Proc. of 2nd Int. Conf. on Simulation of Adaptive Behavior*, pp. 432-441, 1993.

[3] 심귀보, "인공생명을 갖는 지능 로봇시스템의 실현," *대한전자공학회지 인공생명 기술특집*, vol. 24, no. 3, pp. 70-82, 1997. 3.

[4] L.P. Kaelbling, "On Reinforcement Learning for Robotics," *Proc. of Int. Conf. on IROS*, pp. 1319-1320, 1996.

[5] R.S. Sutton, "Learning to Predict by the Methods of Temporal Differences," Machine Learning, vol 8, pp. 9-44, 1992.

[6] C.J.C.H Watkins, P. Dayan, Technical Note : "Q-Learning," *Machine Learning*, vol. 8, pp. 279-292, 1992.

[7] 이동욱, 심귀보, "분산유전알고리즘을 이용한 자율이동로봇군의 행동진화," *제5회 인공지능, 신경망 및 퍼지시스템 종합학술대회(JCEANF '96) 논문집*, pp. 127-130, 1996. 10

[8] E. Horiuchi, K. Tani, "Behavior Learning of Group of Mobile Robots with a Distributable Genetic Algorithm," *J of RSJ(Japanese)*, vol. 11, no. 8, pp. 1212-1219, 1993.

[9] 이동욱, 심귀보, "자율이동로봇군의 협조행동을 위한 통신시스템의 개발," *대한전자공학회 논문지*, vol. 34-S, no. 3, pp. 33-45, 1997. 3

[10] A. Loraschi et. al., "Distributed Genetic Algorithms with An Application to Portfolio Selection Problems," *Proc. of Int. Conf. Artificial Neural Nets and Genetic Algorithms*, pp. 384-387, 1995.

[11] T. Ueyama, T. Fukuda, "Cooperative Search Using Genetic Algorithm Based on Local Information -Path Planning for Structure Configuration of Cellular Robot-," *Proc. of Int. Conf. on IROS*, pp. 1110-1115, 1993.

[12] T. Fukuda, T. Ueyama, *Cellular Robotics and Micro Robotic System*, World Scientific, 1994.

저 자 소 개

李 東 昱(正會員) 第 34 卷 S編 第 3 號 參照
현재 중앙대학교 제어계측공학과 석사과정

沈 貴 寶(正會員) 第 34 卷 S編 第 3 號 參照
현재 중앙대학교 제어계측공학과 부교수