# 로봇 협업을 위한 협업 작업 설계 및 학습에 관한 실험적 연구

김성현, 노삼열, 이동훈, 장인국
한국전자통신연구원

{kim-sh, samuel, donghun, ingook}@etri.re.kr

# An Experimental Study on the Collaborative Task Design and Learning for Robot Collaboration

Seonghyun Kim, Samyeul Noh, Donghun Lee and Ingook Jang
Electronics and Telecommunications Research Institute

요 약

본 논문은 다중 로봇이 동작하는 환경에서 다중 로봇이 협력적으로 동작하기 위한 협업 작업 설계 및 학습에 관한 연구를 다룬다. 다중 로봇 환경은 단일 로봇 환경과 다르게 로봇 간의 행동이 서로 영향을 주는 관계를 갖기 때문에, 다중 로봇 환경의 확률적 변화가 Non-stationary 의 특성을 갖는다. 이와 같은 Non-stationary 문제는 학습의 성능 저하 및 학습 모델 수렴시간 지연에 영향을 주므로, 협업 작업에 대한 복잡도와 큰 연관을 갖는다. 본 논문에서는 Pick-Push-Place 협업 작업 문제를 다루고, 이에 대한 가상환경 구현 및 학습 결과 검증에 대하여 논의한다.

## Ⅰ. 서 론

최근 강화학습은, 2 차원 게임 및 3 차원 객체 자세제어 등 다양한 환경에서 연구되어 많은 발전을 이루고 있다. 강화학습에 대한 기술적 발전은 로봇 환경에서도 강화학습을 도입하고자 하는 시도로 확장되고 있다. 로봇 환경은 제어 문제에서 난이도가 높은 문제에 해당한다[1][2]. 단일 로봇 환경에서는 Reach, Push, Place 등의 매니퓰레이션 작업을 해결하기 위한 많은 시도와 알고리즘이 제안된 반면, 다중 로봇 환경에서의 Handover, Pick-Push-Place, Collaborative Lift 등의 협업 작업에 안정적으로 동작하는 알고리즘 개발은 단일 로봇 작업에 비해 상대적으로 많은 개발이 이루어지지 않았다. 본 논문에서는 협업 작업 문제 중 하나인 Pick-Push-Place 협업 작업 문제를 다루고, 이에 대한 가상환경 구현 및 학습 결과 검증에 대하여 논의한다.

## Ⅱ. 본론

다중 로봇 환경은 단일 로봇 환경과 다르게 로봇 간의 행동이 서로 영향을 주는 관계를 갖기 때문에, 다중 로봇 환경의 확률적 변화가 매시간마다 다르게 변화하는 Non-stationary 특성을 갖는다[3]. 예를 들어, 특정 로봇의 관점에서 보았을 때, 다른 로봇의 행동이 해당 로봇의 행동 결과에 영향을 미치기 때문에, 해당 로봇의 행동으로 인한 환경 상태의 전이확률이 다른 로봇의 행에 따라 변화한다. 즉, 환경 상태 공간의 변화 범위가 넓어지기 때문에 단일 로봇 환경보다 더 많은 경험 데이터가 필요하고, 이로 인해 학습 모델의 수렴시간 지연 및 학습 성능 저하 등의 문제가 초래될 수 있다. 이와 같은 문제의 복잡도는 작업 환경의 설계와 밀접한 연관을 갖는다. 본 논문에서는 기존 다중 에이전트 학습 알고리즘이 안정적으로 동작하는 베이스라인 협업학습 환경 설계를 목적으로 Pick-Push-Place 협업 작업을 설계한다.

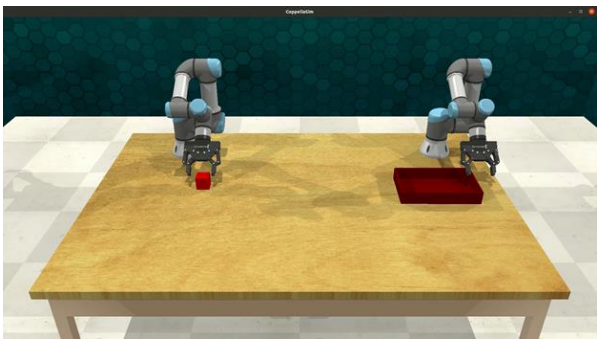

**그림 1. Pick-Push-Place 협업 작업 환경**

### 1. Pick-Push-Place 협업 작업

그림 1 은 두 개의 로봇, 큐브(빨간색), 트레이(갈색)로 구성된 Pick-Push-Place 협업 작업 환경을 나타낸다. 여기서, 각 로봇의 팔 관절 자유도는 6, 그리퍼 자유도는 1 이다. Pick-Push-Place 협업 작업에서 성공 조건은 '큐브가 트레이의 내부에 위치한다'로 정의된다. 이와

같은 협업 작업을 성공하기 위해서는 큐브와 컨테이너 사이의 먼 거리로 인해 두 로봇의 협업이 필수적이다. 즉 왼쪽의 로봇이 큐브를 잡고 트레이 방향으로 이동시켜야 하고, 마찬가지로 오른쪽의 로봇은 트레이를 큐브 방향으로 이동시켜야 하고, 마지막으로 큐브와 트레이의 위치를 잘 조율하여 큐브를 트레이 내부에 놓아야 한다.

## 2.보상함수 설계

다중 로봇 협업 작업을 강화학습 방식으로 해결하기 위해서는 보상 함수 설계가 필요하고, 관련하여 아래와 같이 다양한 보상 함수를 정의한다.

- Distance reward between instance $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$r_d^{(a,b)} = -\sqrt{\|\boldsymbol{a}_p - \boldsymbol{b}_p\|^2}$$

- Grasping reward:

$$r_g^{(i,b)} = \begin{cases} 1, & \text{for gripper } i \text{ grasping object } b \\ 0, & \text{for otherwise} \end{cases}$$

- Place reward:

$$r_p = \begin{cases} 1, & \text{for sensor on tray detecting cube} \\ 0, & \text{for otherwise} \end{cases}$$

여기서 $\boldsymbol{a_p}$ 와 $\boldsymbol{b_p}$ 는 인스턴스 $\boldsymbol{a}$ 와 $\boldsymbol{b}$ 의 위치를 의미하고, 인스턴스 $\boldsymbol{a}$ 와 $\boldsymbol{b}$ 는 로봇 $i$의 그리퍼 또는 두 물체(큐브, 트레이)가 될 수 있다. 이와 같은 보상함수들을 이용하여, 전체 보상함수는 다음과 같이 정의된다.

$$R = \sum_{i,b(i)} \left( w_d r_d^{(i,b(i))} + w_g r_g^{(i,b(i))} \right) + w_o r_d^{(C,T)} + w_p r_p \ .$$

여기서 $b(i)$는 로봇 $i$ 에 해당되는 물체를 의미한다. 즉 왼쪽 로봇 0 는 큐브 $C$, 오른쪽 로봇 1 은 트레이 $T$ 에 해당됨을 뜻한다. 그리고 $w_x$ 는 각 보상함수에 대한 가중치를 나타낸다.

## 3. MA-DDPG

다중 로봇과 같은 다중 에이전트 문제에서 발생하는 Non-stationary 를 해결하기 위한 알고리즘으로 MA-DDPG 기법이 제안되었다[4]. Actor-Critic 기반의 MA-DDPG 는 다중 에이전트의 정책을 분산적으로 실행하고 가치함수를 중앙집중적으로 업데이트한다. 가치함수를 업데이트할 때 모든 에이전트의 행동값을 활용하므로, 중앙집중 관점에서 Non-stationary 문제가 발생하지 않는다.

## Ⅲ. 실험결과

Pick-Push-Place 협업 작업에서 다중 로봇을 학습하기 위해 MA-DDPG 알고리즘을 적용하였고, 네트워크 구조와 내부 파라미터는 기존 MA-DDPG 연구와 동일하다. 학습 시 에피소드당 최대 길이는 50 스텝, 최대 에피소드는 100000 회이다. 보상함수에서 가중치 값은 $w_d = 0.5, w_g = 1.0, w_o = 2.0, w_p = 10.0$ 을 사용하였다. 즉, 성공조건의 요소를 이루는 두 물체의 Distance 와 Place 에 더 높은 가중치를 부여하여, 성공조건을 달성하기 위해 집중하도록 보상을 세팅하였다.

그림 2 는 에피소드에 따른 전체 보상을 나타낸다. 그림 2 에서 확인되듯이, 약 10000 에피소드까지 Distance 보상 증가에 따라 가파르게 성능이 증가하고. 약 50000 에피소드까지는 Grasping 보상을 학습하는 단계를 거친다. 약 65000 에피소드에서 Place 보상을 받으며 최대 성능을 얻고, 이후 탐색과정을 추가로 거치며 성능이 하락한 후 다시 성능이 증대된다.

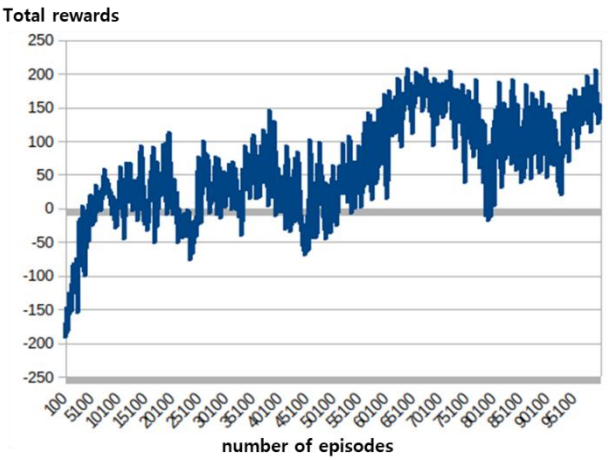


그림 2. 에피소드에 따른 전체 보상

위 결과는 다양한 가중치 조합을 통해 실험적으로 협업 작업을 성공하는 케이스를 찾은 것으로써, 동일한 가상환경 및 보상함수 요소를 사용하더라도 가중치에 따라 학습을 통해 성공을 못하는 경우가 다수 발생한다.

## Ⅳ. 결론

본 논문에서는 Pick-Push-Place 협업작업을 해결하기 위한 다중 로봇 협업 문제를 다루었다. 실험결과를 통해 가중치의 설정에 따라 MA-DDPG 의 작업 성공여부가 결정되는 것을 알 수 있다. 향후에는 본 논문에서 설계한 작업 환경 및 보상함수 세팅을 베이스라인으로 설정하여, 보다 샘플 효율적인 알고리즘을 개발하고자 한다.

## ACKNOWLEDGMENT

본 연구 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음. [22ZR1100, 자율적으로 연결· 제어· 진화하는 초연결 지능화 기술 연구].

## 참 고 문 헌

[1] M. Zhang, et al, “Work chain-based inverse kinematics of robot to imitate human motion with Kinect,” ETRI Journal, vol. 40, no. 4, pp. 511-521, Aug., 2018.

[2] M. Ilbeygi and M. R. Kangavari, “Comprehensive architecture for intelligent adaptive interface in the field of single-human multiple-robot interaction,” ETRI Journal, vol. 40, no. 4, pp. 483-498, Aug., 2018.

[3] H. Kim et al, “Avoiding collaborative paradox in multi-agent reinforcement learning,” ETRI Journal pp. 1004-1012, 43(6), 2021.

[4] R. Lowe et al, “Multiagent actor-critic for mixed cooperative-competitive environments,” in Proc. NIPS, 2017, pp. 2094-2100.