Journal of Korea Robotics Society (2020) 15(4):404-412
https://doi.org/10.7746/jkros.2020.15.4.404 ISSN: 1975-6291 / eISSN: 2287-3961

테이블 균형맞춤 작업이 가능한 Q-학습 기반 협력로봇 개발

Cooperative Robot for Table Balancing Using Q-learning

김 예 원[1] · 강 보 영†
Yewon Kim[1], Bo-Yeong Kang†

Abstract: Typically everyday human life tasks involve at least two people moving objects such as tables and beds, and the balancing of such object changes based on one person's action. However, many studies in previous work performed their tasks solely on robots without factoring human cooperation. Therefore, in this paper, we propose cooperative robot for table balancing using Q-learning that enables cooperative work between human and robot. The human's action is recognized in order to balance the table by the proposed robot whose camera takes the image of the table's state, and it performs the table-balancing action according to the recognized human action without high performance equipment. The classification of human action uses a deep learning technology, specifically AlexNet, and has an accuracy of 96.9% over 10-fold cross-validation. The experiment of Q-learning was carried out over 2,000 episodes with 200 trials. The overall results of the proposed Q-learning show that the Q function stably converged at this number of episodes. This stable convergence determined Q-learning policies for the robot actions. Video of the robotic cooperation with human over the table balancing task using the proposed Q-Learning can be found at http://ibot.knu.ac.kr/videocooperation.html.

Keywords: Cooperative Robot, Reinforcement Learning, Q-learning, Image Processing, Classification, AI (Artificial Intelligence), NAO Robot

1. 서 론

바쁜 일상 속에서 개인의 휴식이나 자기 개발을 위한 시간은 매우 중요해졌고, 이를 위해 집안일과 같은 필수적인 일을 기계 또는 로봇이 대신하는 경향이 늘고 있다. 집안일을 대신 수행하는 기계인 세탁기, 건조기, 식기세척기는 많은 가정에서 볼 수 있으며, 인공지능(artificial intelligence, AI)이 적용된 로봇청소기 또한 등장하였다. 로봇에 강화학습을 적용한 선행연구[1-13]에 나타나는 기술은 사람을 대신하여 과제를 수행하지만 대부분 사람과 협력 없이 로봇 스스로 작업을 수행하는 연구이다. 그러나 테이블, 침대와 같은 물체 운반 작업은 두 사람의 협력이 필요하다. 물체 운반 작업에 필요한 두 사람의 협력 중 하나는 물체 운반 중 물체 손상 방지를 위해 물체의 균형을 지면과 평행하게 유지하는 것이다. 그러므로 사람을 대신하여 로봇이 물체를 옮기는 기술에는 물체를 옮기는 동안 물체의 손상 방지를 위해 물체의 균형을 맞추는 기술이 필요하다. 두 사람이 물체를 들고 있을 때, 한 사람의 동작에 의해 테이블의 균형이 달라진다. 따라서 본 논문에서는 사람과 로봇의 협력 작업 수행을 위한 테이블 균형맞춤 작업이 가능한 Q-학습 기반 협력로봇 개발 기술을 제안한다. 본 논문에서 제안하는 협력로봇은 다음과 같이 동작한다. 예를 들어, 두 사람이 테이블을 들고 있을 때, 한 사람의 동작에 의해 테이블의 균형이 달라진다. 테이블의 균형을 다시 맞추기 위해서는 테이블을 움직인 사람의 동작과 같은 동작을 다른 사람이 수행하면 된다. 로봇과 사람이 테이블을 들고 있을 때도 마찬가지로 테이블의 균형을 맞추기 위해서는 사람의 동작을 로봇이 수행하는 것이 필요하다. 로봇이 사람 동작을 수행하기 위한 필요요소의 첫 번째는 로봇의 사람 동작 인식, 두 번째는 협력로봇의 사람 동작 수행이다. 이를 위해 딥러닝(deep learning) 기술 중

Received : Sep. 7. 2020; Revised : Oct. 14. 2020; Accepted : Oct. 15. 2020
※ This project was supported by the National Research Foundation of Korea Funded by the Korean Government under grant NRF-2019R1A2C1011270
1. Master Student, Mechanical Engineering, Kyungpook National University, Daegu, Korea (yewonkim.knu@gmail.com)
† Professor, Corresponding author: Mechanical Engineering, Kyungpook National University, Daegu, Korea (kby09@knu.ac.kr)

하나인 AlexNet[14]을 이용하여 사람의 동작을 이미지 분류(classification)로 인식하고, 강화학습(reinforcement learning)의 Q-학습을 이용해 로봇 동작을 결정한다. AlexNet을 이용한 사람 동작 인식은 로봇의 카메라를 통해 보는 테이블의 상태를 통하여 구분한다. 훈련된 AlexNet과 훈련된 Q-학습의 정책을 이용해 제안한 협력로봇은 실시간으로 사람의 동작을 인식하고 테이블 맞춤을 위한 동작을 수행한다. 이로써 본 논문에서 제안하는 협력로봇은 사람과 테이블 맞춤 작업을 할 수 있게 된다.

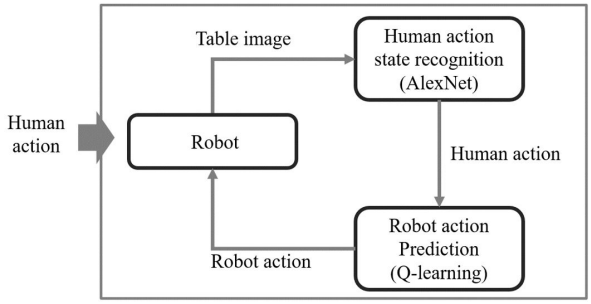
2. 선행 연구

로봇을 강화학습으로 구동한 예로는 아틀라스(Atlas)[1]가 있다. 아틀라스는 휴머노이드 로봇으로 보스턴 다이나믹스에서 개발되는 로봇 중 하나이며, 아틀라스가 점프와 달리기를 하는 데 강화학습이 사용되었다. 그 결과 2017년 11월에 아틀라스가 징검다리형 구조물을 건너고 높은 곳에서 백 덤블링을 하는 영상을 공개하며 로봇 인공지능 기술의 발전을 보여주었다. 아틀라스 이외에 실제 로봇의 특정 행동 구현에 강화학습을 적용한 또다른 연구는 길찾기 로봇[2], 일어나는 로봇[3], 균형잡기 및 문열기 로봇[4]이 있다. Shuhuan Wen et al.[2]은 Q-학습과 EKF-SLAM (extended Kalman filter for simultaneous localization and mapping)을 이용해 NAO 로봇이 자율적으로 길을 찾는 연구를 하였다. M. Danel[3]은 NAO 로봇이 누워있을 때 강화학습을 이용해 일어나는 연구를 하였다. Freek Stulp et al.[4]은 로봇이 오른쪽 팔을 들어 올리고 내리며 균형을 잡는 과제와 문열기 과제에 강화학습을 적용하여 수행하였다.

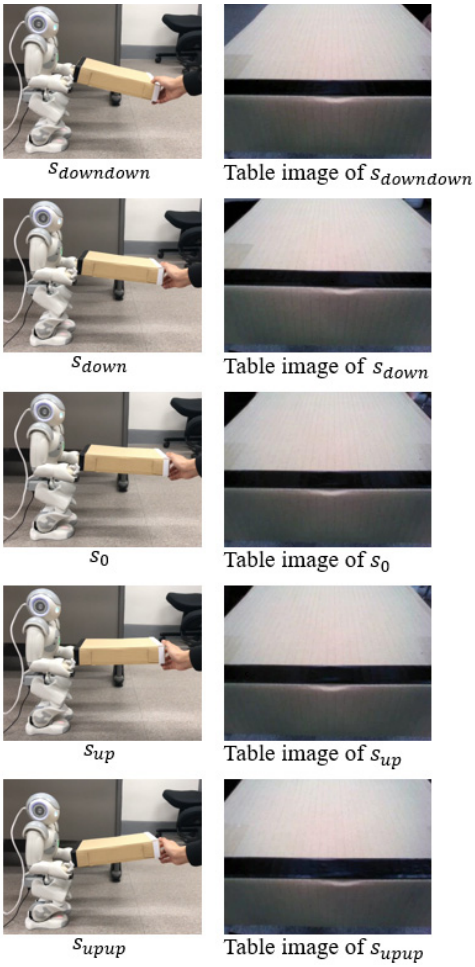
강화학습을 적용한 가장 유명한 연구는 알파고[5]이다. David Silver et al.[5]은 바둑 AI 기술에 주로 사용하는 탐색 트리(tree search)에 딥러닝과 강화학습을 접목하여 성능을 높였다. 알파고가 바둑(Go)을 한 것처럼 게임을 함께하기 위한 연구로 탁구 로봇[6], 볼링 로봇[7]이 있다. Katharina Mülling et al.[6]이 연구한 탁구 로봇은 강화학습을 통해 탁구하는 방법을 익혔다. Debnath와 Nassour[7]는 NAO 로봇이 볼링을 하는 데 강화학습을 사용하여 연구를 하였다. 바둑, 볼링, 탁구 등과 같이 사람과 대결 가능한 게임을 로봇이 배우는 연구 이외에도 실제로 로봇만을 사용하여 축구하는 경기인 RoboCup Standard Platform League (SPL)을 위한 연구가 있다. 이 경기에 참여하는 로봇은 스스로 상황을 판단하여 축구를 하며 축구 경기 수행 기술에 주로 강화학습이 사용되었다. Okan As et al.[8]은 모방 학습을 통해 로봇이 프로그래밍한 축구 전문가의 동작을 모방하여 축구 동작을 구동하는 연구를 하였으며, 모방 학습은 CNN (convolutional neural network)과 강화학습을 통해 구현하였다. Kenzo Lobos-Tsunekawa et al.[9]은 휴머노이드 로봇을 위한 map-less visual navigation system 구현을 통해 맵 정보 없이 NAO 로봇 자체 카메라 이미지를 이용한 이동 경로 생성 기술에 강화학습을 적용하였다. 청소, 빨래, 설거지와 같은 집안일이나 도구를 옮기고 사용하는 일, 문을 여는 일 등과 같은 과제를 수행하는 로봇을 구현하는 연구[10-13] 또한 진행되어왔다. Sergey Levine et al.[10]은 카메라 인식을 통해 1100가지 종류의 물체를 7 자유도 팔 로봇이 최적화된 방식으로 들어올리는 과제에 강화학습을 사용하였다. Pin-Chu Yang et al.[11]은 휴머노이드 로봇이 카메라 이미지를 통해 수건을 인식하고 수건을 접는 과제(folding task)를 수행하는데 강화학습을 적용하였으며, 수건을 접는 과제와 비슷한 과제인 책을 덮는 과제도 수행함으로써 과제 실행의 완성도를 높였다. Chang Wang et al.[12]은 로봇이 쓰레기통을 사용하는 과제를 수행하였다. 로봇의 카메라 이미지에서 컬러 영상 분할(color image segmentation)을 수행하여 대상을 인식한 후 쓰레기통의 문을 여는 과제를 강화학습을 이용해 수행하였다. Suay와 Chernva[13]는 로봇이 작은 구체를 옮기는 과제를 수행하였고, 로봇 카메라를 이용하여 물체를 인식하고 강화학습을 통해 왼쪽 또는 오른쪽 컵으로 물체를 이동하였다.

로봇 구동을 위한 자연스러운 동작 생성에 강화학습을 적용한 연구로는 사람 동작 모방[15], 지형에 따른 동물 동작 생성[16] 이 있다. Josh Merel et al.[15]은 강화학습만을 이용해 배우는 모방 동작의 부자연스러움을 극복하고자 사람 동작을 모방하는 학습에 대한 연구를 진행하였다. 모션 캡쳐로 만든 동작과 생성기에서 만든 동작을 비교하는 방식으로 강화학습이 적용되었다. X. B. Peng et al.[16]은 동물들이 지형에서 움직이는 동작을 구현하기위해 지형에 따른 동작을 생성하는 기술에 강화학습을 사용하는 연구가 진행되었다. 이와 같은 로봇들에 적용된 강화학습 기술로는 Q-학습을 시작으로 여기에 신경망(neural network)을 접목한 DQN[17], 3차원 연속 동작에 사용하기 좋은 DDPG[18], 결과 값을 일정 비율로 업데이트하던 방식에서 결과 값의 신뢰 영역 이내로 근사화한 값을 업데이트하는 TRPO[19]와 PPO[20] 등이 있다. 위의 선행연구의 대부분은 로봇 단독으로 과제를 수행하는 기술이며, 로봇이 사람과 함께 수행하는 과제에 대한 연구는 게임이나 운동을 함께하는 연구이다.

사람과 로봇이 협력하여 과제를 수행하는 연구로서 테이블 균형맞춤 과제 수행 연구[21,22]가 있다. 두 연구 모두 사람의 동작에 맞춰 테이블 균형맞춤 로봇 동작을 수행하였다. Anand Thobbi et al.[21]은 로봇의 테이블 균형 맞춤 수행을 위해 두 컨트롤러를 사용하였다. 두 컨트롤러는 EKF (extended kalman filter)를 이용한 사람 동작 예측 컨트롤러와 Q-학습을 이용한 테이블 균형 맞춤을 위한 로봇 동작 출력 컨트롤러이며, 두 컨트롤러 모두 테이블의 좌표 정보를 입력값으로 이용하여 결과를 출력하였다. Ye Gu et al.[22]은 관절 좌표를 이용해 가우시안



[Fig. 1] Proposed Q-learning based table balancing workflow



State	$s_{downdown}$	s_{down}	s_0	s_{up}	s_{upup}
Area	33,647	35,165	36,622	39,636	41,276

[Fig. 2] 5 human action states with NAO's camera image of 5 states. And table upper areas in the table images

혼합 모델(gaussian mixture model)과 가우시안 혼합 회귀(gaussian mixture regression), Q-학습을 통해 로봇이 사람 동작을 모방하였다. 두 논문 모두 the Vicon MX 모션 캡처 시스템[23]에서 광학용 마커 라벨링과 다수의 고해상도 카메라를 이용하여 테이블 좌표 정보와 관절 정보를 획득하였다. 하지만 테이블의 위치정보 수집에 다수의 고해상도 카메라와 광학용 마커 라벨링이 요구되는 the Vicon MX 모션 캡처 시스템의 사용은 고성능 외부 장치의 필요성과 마커 레이블링을 위한 사람의 개입 및 반자동 시스템의 한계를 나타낸다. 본 논문의 연구에서는 고성능 외부장치의 도움 없이 로봇에 장착된 일반 성능의 카메라와 딥러닝 비전 기술, 강화학습을 이용하여 테이블 균형맞춤 과제를 자동으로 학습하고 수행하고자 한다.

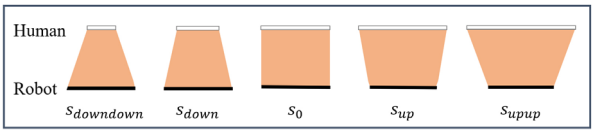
로봇이 테이블의 균형을 유지하기 위해서는 테이블을 움직인 사람의 동작을 살펴보고 로봇이 균형을 맞추기 위해 적합한 동작을 수행하면 된다. 제안한 기술은 먼저, 딥러닝 기술인 AlexNet을 이용해 사람의 동작을 인식한다. 이때, 사람의 동작 인식을 위해 로봇의 카메라로 촬영한 테이블 이미지를 사용한다. 사람의 동작에 따라 카메라에 촬영되는 테이블의 이미지가 달라지므로 테이블 이미지를 이용해 사람의 동작을 인식할 수 있다. 그런 후 강화학습인 Q-학습 정책에 따라 로봇 동작을 결정한다. 훈련된 AlexNet과 Q-학습 정책에 기반하여 제안한 협력로봇은 실시간으로 사람의 동작을 인식하고 테이블 맞춤을 위한 동작을 수행하여 사람과 테이블 균형 맞춤 작업을 할 수 있게 된다.

본 논문의 구성은 다음과 같다. Section Ⅲ 에서는 제안한 기술의 방법에 대해서 설명한다. Section Ⅳ 에서는 제안한 기술을 사용하여 얻은 실험의 결과에 대해 서술되어 있다. Section Ⅴ에서 본 논문의 결론을 맺고 향후 연구에 대해 고찰한다.

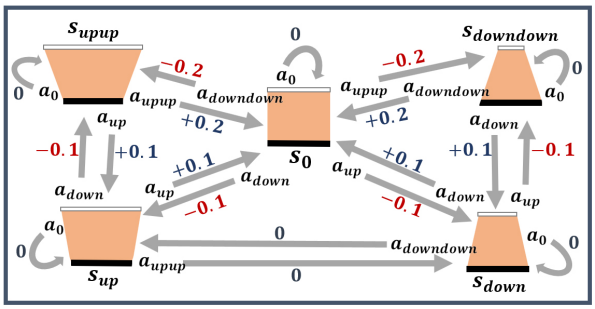
3. 제안한 Q-학습기반 균형맞춤로봇

테이블 균형맞춤 작업이 가능한 Q-학습 기반 협력로봇 개발 기술의 전체 동작 흐름은 [Fig. 1]과 같다; 제안한 기술은 사람의 동작을 인식하기 위한 사람 동작 상태 인식(Human action state recognition) 컴포넌트와 로봇의 테이블 균형맞춤 동작을 예측하여 출력하는 테이블 맞춤을 위한 로봇 동작 예측(Robot action prediction) 컴포넌트로 구성된다.

사람 동작 상태 인식(Human action state recognition) 컴포넌트에서 로봇은 사람 동작을 인식하며, 이를 위해 로봇의 카메라로 촬영한 테이블 이미지를 이용한다. 사람과 로봇이 테이블을 잡고 있을 때 사람의 동작에 의해 테이블의 균형이 달라지고, 테이블 균형 변화에 따라 로봇 시야의 테이블 형태도 달라진다. 따라서 로봇의 카메라로 촬영한 테이블 이미지를 이용해 사람의 동작을 인식할 수 있다. 로봇의 카메라로 촬영한 테이블 이미지를 딥러닝 기술인 AlexNet에 입력하여 사람의 동작을 이미지 분류로 인식한다. 사람 동작 인식 기술로 AlexNet을 선택한 이유는 병렬구조 CNN의 초기 모델이기 때문이다. 제안한 기술은 구현 초기 단계이므로 다양한 발전 가능성을 위해 가장 기본적인 모델을 사용하여 구현하였으며, 자동으로



[Fig. 3] 5 human action states from table image by NAO's camera



[Fig. 4] Proposed Q-learning model of the table balancing task

사람 동작 상태를 분류하기 위해 딥러닝 기술을 채택하였다.

테이블 맞춤을 위한 로봇 동작 예측(Robot action prediction) 컴포넌트에서는 사람 동작 상태 인식(Human action state recognition) 컴포넌트에서 확인한 사람의 동작을 이용해 로봇의 테이블 균형맞춤을 위한 동작을 출력한다. 테이블 균형맞춤을 위한 로봇 동작 출력 기술로 강화학습의 Q-학습을 채택한다. Q-학습 프로세스에 따라 테이블 균형맞춤 과제 수행에 적절한 로봇 동작을 학습한 후 각 상태에 따른 로봇 동작 정책을 결정한다. 결정된 Q-학습 정책에 따라 테이블 균형맞춤을 위한 로봇 동작이 출력되며, 로봇은 Q-학습에서 출력된 동작을 구동하여 테이블 균형맞춤 작업을 한다. 본 논문에서 사용한 로봇은 Softbank의 NAO로봇[24]이며, 사용한 테이블은 [Fig. 2]와 같이 직사각형 모양의 테이블로 가로 31 cm, 세로 23 cm, 높이 6 cm, 무게 0.1 kg이다.

3.1 사람 동작 상태 인식(Human action state recognition)

본 절에서는 사람의 동작 상태를 인식을 위해 로봇의 카메라로 촬영한 테이블 이미지를 사용하는 과정을 설명한다. 사람과 로봇이 테이블을 들고 있을 때 사람의 동작에 따라 테이블의 균형이 달라진다. 테이블의 균형 변화는 로봇 시야의 테이블 형태 변화를 가져오므로 로봇의 카메라로 촬영하는 테이블 이미지의 변화로 사람의 동작을 인식할 수 있다. 본 컴포넌트에서 사용할 테이블 이미지는 NAO 로봇에 장착된 두 개의 카메라 중 하부 카메라(bottom camera)로 촬영한 이미지를 사용하는데, 하부 카메라의 시야가 대각선 아래 방향이므로 테이블 윗면 이미지 촬영이 가능하기 때문이다.

테이블 이미지에 따라 분류할 사람의 동작은 5가지로 구분하였으며, 이에 대한 사람의 동작과 테이블 이미지는 [Fig. 3]에 나타나있다. 사람의 동작 분류는 테이블 올리기, 내리기, 유지하기가 있으며, 올리고 내린 테이블 기울기 정도에 따라 다음과 같이 정의한다.

- 테이블 많이 내리기($s_{downdown}$): $s_{downdown}$ < -4°
- 조금 내리기(s_{down}): -4° ≤ s_{down} ≤ -1°
- 유지하기(s_0): -1°<s_0<1
- 조금 올리기(s_{up}): 1° ≤ s_{up} ≤ 4°
- 많이 올리기(s_{upup}): 4°<s_{upup}

사람이 테이블을 올리거나 내린 정도에 따라 테이블 이미지의 형태는 달라지며, 이는 [Fig. 2]의 테이블이미지와 표에 나타나있다. 표의 테이블 영역은 영상 분석 프로그램인 ImageJ를 통해 측정하였다. 예를 들어, 사람이 테이블을 올린 상태인 s_{upup}과 s_{up}을 비교하면 s_{upup}은 이미지 내에서 테이블이 차지하는 영역이 41,276 픽셀이며 이는 s_{up}인 39,636픽셀에 비해 높은 수치이다. 반대로 사람이 테이블을 내린 상태인 $s_{downdown}$과 s_{down}을 비교하면 $s_{downdown}$은 이미지 내에서 테이블이 차지하는 영역이 33,647픽셀로 s_{down}인 35,165픽셀보다 적다.

본 논문에서는 딥러닝 중 CNN (convolutional neural network) 기술인 AlexNet을 통해 테이블 상태를 분류하였으며, 테이블 상태 분류를 통해 사람의 동작을 인식한다. AlexNet의 훈련(training)을 위한 데이터 셋은 로봇의 하부 카메라로 사람의 동작에 의한 테이블 이미지를 촬영하여 구성하였으며, 구성된 데이터 셋을 이용해 로봇을 훈련하였다.

3.2 로봇 동작 예측(Robot action prediction)

본 절에서는 강화학습을 사용하여 테이블 균형맞춤 작업을 위한 로봇 동작을 출력한다. 로봇 동작 출력을 위해 사용한 강화학습 기술은 Q-학습이다. 로봇 동작 예측 컴포넌트는 Q-학습을 이용해 사람 동작 상태 인식 컴포넌트에서 분류한 사람의 동작에 따라 상황에 적합한 로봇 동작을 출력한다.

Q-학습은 강화학습 방법으로 식 (1)과 같이 상태(s)와 로봇 동작(a)에 대한 Q 함수(동작가치함수)를 이용해 각 상태에서 최적의 로봇 동작을 구현하는 기술이다. 각 상태에서 구동할 로봇 동작은 정책(π)에 따라 선택한다. 정책은 각 상태에서 구동할 로봇 동작을 의미하며, 본 논문에서는 탐욕 정책(greedy policy)을 사용하며 다음 식 (1)과 같이 나타낸다.

$$\pi(s) = \mathrm{argmax}_{a \in A} Q(s,a) \tag{1}$$

식 (1)에서 π는 정책, s는 상태, a는 로봇 동작, A는 모든 로봇 동작의 집합을 의미한다. 모든 로봇 동작의 집합인 A는 모

든 로봇 동작인 $a_{downdown}$, a_{down}, a_0, a_{up}, a_{upup}을 모두 포함한다. $\mathrm{argmax}_{a\in A}Q(s,a)$는 해당 상태($s$)에서 Q 함수의 값이 가장 높은 로봇 동작(a)을 출력한다. 즉, 탐욕정책은 각 상태에서 Q값이 가장 높은 로봇 동작을 정책으로 사용한다. 단, 최대 Q값을 가진 로봇 동작 출력 시 다수의 로봇 동작이 출력될 경우 다수의 로봇 동작 중 하나의 로봇 동작을 랜덤하게 선택한다.

테이블 균형 맞춤을 위한 제안된 Q-학습 모델은 [Fig. 3], [Fig. 4]와 같다. [Fig. 3]은 사람 동작에 따른 5가지 테이블 상태를 표현하며, [Fig. 4]는 각 테이블 상태에서 다음 테이블 상태로 가는 로봇 동작에 따른 보상의 관계도이다. 테이블의 균형맞춤 상태에 근접하면 양수 보상, 멀어지면 음수 보상이 주어진다. 로봇 동작 선택에 사용되는 Q값은 로봇 동작에 따라 업데이트되며 Q 함수의 업데이트 식은 식 (2)와 같다.

$$Q(s_t,a_t)=(1-\alpha)Q(s_t,a_t)+\alpha(r_t+\gamma\max_{a\in A}Q(s',a)) \quad (2)$$

식 (2)에서 r_t는 [Fig. 4]에 따라 현재 상태에서 다음 상태로 가는 로봇 동작에 대한 보상을 의미하며, s_t, a_t, r_t의 t는 실행 횟수(episode)을 의미한다. s'은 다음 상태이며 s_t일 때 로봇이 a_t를 구동한 후의 상태이다. s'과 s_{t+1}의 차이는 동작의 주체이다. s_t와 s_{t+1}은 사람의 동작 이후의 상태이며 제안한 기술의 입력값이고, s'은 로봇 동작 구동 이후의 상태이며 제안한 기술의 결과이다. s_t에서 a_t 구동 후 s'가 되었을 때 사람의 동작에 의해 s_{t+1}으로 상태가 바뀐다. α는 학습률, γ는 다음 상태의 Q값에 대한 영향을 감소하는 감가율이며, α, γ 모두 0에서 1 사이의 값을 사용한다(본 모델에서는 α는 0.1, γ는 0.9를 사용하였다). $\max_{a\in A}Q(s',a)$는 다음 상태의 로봇 동작 중 최대 Q값을 의미한다. 다음 상태의 최대 Q값 사용은 미래 영향을 고려한다는 의미이며, 감가율 γ을 사용하여 현재 상태 업데이트에 미래 영향을 감쇠한다.

Q 함수 업데이트 및 로봇 학습 과정을 식 (2)를 사용하여 예로 들면 다음과 같다; α는 0.1, γ는 0.9를 사용한다고 가정하자. 그리고 $t=1$일 때 사람 동작 상태가 s_{down}이고, 식 (1)인 탐욕 정책에 따라 로봇 동작 a_{down}을 출력했다고 가정하자($t=1$이므로 Q 함수는 초기 상태이며, 로봇 동작 a_{down}은 랜덤 선택에 의해 출력되었다). 가정에 따라 $t=1$일 때 $Q(s_1,a_1)$은 $Q(s_{down},a_{down})$이며 s_{down}의 a_{down}에 대한 Q 함수 업데이트가 진행된다. $t=1$일 때 Q 함수는 초기화되어 있으므로 $Q(s_{down},a_{down})$는 0이다. 그리고 [Fig. 4]에 따라 현재 상태인 s_{down}에서 a_{down}을 구동하면 보상(r_1)으로 0.1을 받으며 다음 상태(s')는 s_0가 된다. $\max_{a\in A}Q(s',a)$는 s_0의 로봇 동작 중 가장 높은 Q값을 의미하며 마찬가지로 초기화되어 있으므로 0을 갖는다. 따라서 식 (2)를 이용해 Q 함수를 업데이트하면 식 (3)과 같이 $Q(s_1,a_1)$인 $Q(s_{down},a_{down})$은 0에서 0.01로 업데이트된다.

$$\begin{aligned}Q(s_1,a_1)&=Q(s_{down},a_{down})\\&=(1-\alpha)Q(s_{down},a_{down})\\&\quad+\alpha(r_t+\gamma\max_{a\in A}Q(s_0,a))\\&=(1-0.1)\times0+0.1\times(0.1+0.9\times0)\\&=0.01\end{aligned} \quad (3)$$

예시에서 $t=1$ 일 때 s_{down}에서 a_{down}의 출력은 랜덤 선택에 의한 정책이었으나, 해당 업데이트 이후 사람 동작에 의해 s_{down}에 도달하면 식 (1)에 따라 a_{down}이 정책으로 출력된다. Q-학습 과정 동안 로봇은 각 상태에서 구동한 로봇 동작에 따라 보상을 얻으며, 보상의 영향에 따른 Q 함수의 업데이트로 정책은 점차 달라진다. 따라서 Q-학습은 정책이 로봇 동작에 따른 모델의 영향을 지속적으로 받게 하며, 결과적으로 로봇이 각 상태에서 적합한 로봇 동작을 출력하는 안정된 최종 정책을 얻게 된다.

[Table 1] Confusion matrix

Predict \ Actual	True	False
True	TP (True Positive)	FP (False Positive)
False	FN (False Negative)	TN (True Negative)

[Table 2] 10-fold cross validation of data set

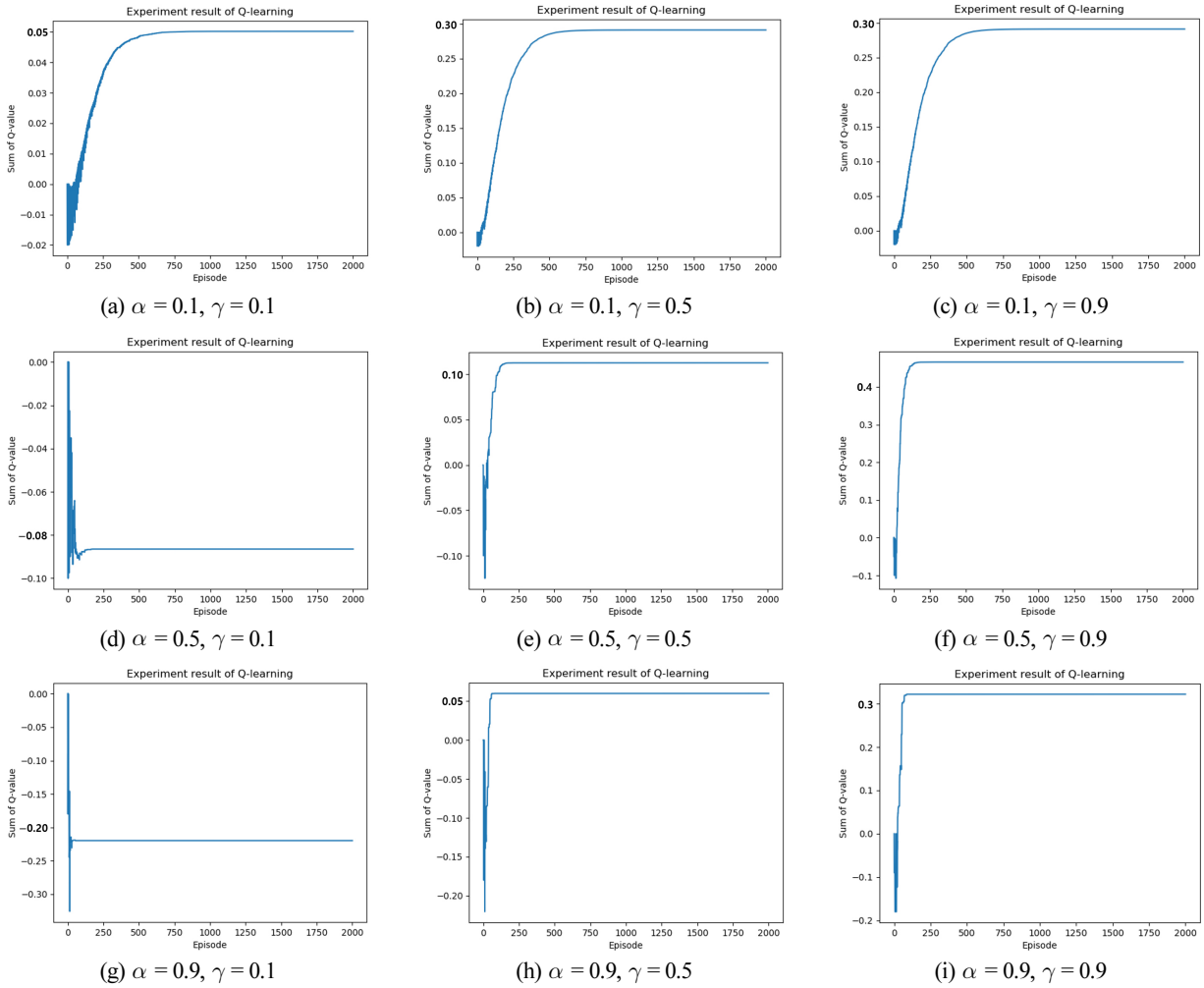
Type of data set	Average accuracy	Best accuracy
Train	94.7%	96.9%
Test	93.9%	95%

4. 실험 결과

본 절에서는 제안한 Q-학습 기반의 테이블 균형맞춤 로봇 기술의 성능을 검증하기 위하여, 사람 동작 상태 인식을 위한 데이터 셋 구축 및 인식, Q-학습 결과 및 기술 구현 프로토타입 시연에 대해 설명한다.

먼저, 사람 동작 상태 인식을 위해 구축한 데이터 셋은 각 상태별 500개로, 총 2500개의 이미지로 구성하였다. 사람 동작 상태 인식 컴포넌트 훈련을 위한 데이터 셋은 [Fig. 2]와 같이 5가지 사람 동작 상태를 태깅(tagging)한 테이블 상태 이미지로 총 2500장으로 구성된다. 2500장 중 2250장은 훈련 데이터 셋, 250장은 테스트 데이터 셋으로 사용하여 10겹 교차검증(10-fold cross-validation)을 진행하였다. 정확도 계산식[25]은 식 (4)와 같다.

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \quad (4)$$



(a) $\alpha = 0.1, \gamma = 0.1$ (b) $\alpha = 0.1, \gamma = 0.5$ (c) $\alpha = 0.1, \gamma = 0.9$

(d) $\alpha = 0.5, \gamma = 0.1$ (e) $\alpha = 0.5, \gamma = 0.5$ (f) $\alpha = 0.5, \gamma = 0.9$

(g) $\alpha = 0.9, \gamma = 0.1$ (h) $\alpha = 0.9, \gamma = 0.5$ (i) $\alpha = 0.9, \gamma = 0.9$

[Fig. 5] Sum of Q-value graph with parameter changes ($\alpha = 0.1, 0.5, 0.9$, $\gamma = 0.1, 0.5, 0.9$)

[Table 1]은 데이터셋 인식의 분류를 나타낸다. 따라서 TP와 TN은 정확한 인식 횟수이며, FP와 FN은 오류적 인식 횟수이다. 식 (4)는 정확한 인식 횟수에서 전체 인식 횟수를 나누어 전체 대비 정답율을 계산한다. 10겹 교차검증을 통한 인식 성능의 결과는 [Table 2]와 같으며 실험 결과, 훈련 데이터 셋의 사람 동작 상태 인식 평균 정확도는 94.7%, 최대 정확도는 96.9%이며, 테스트 데이터셋의 사람 동작 상태 인식 평균 정확도는 93.9%, 최대 정확도는 95%로 우수한 성능을 보였다.

Q-학습에서 최종 정책은 Q 함수 업데이트에도 Q값이 수렴하여 상태별 로봇 동작을 결정하는 정책이 결정됨을 의미한다. 최종 정책 결정의 확인을 위한 Q 함수 수렴 계산식은 식 (5)와 같다.

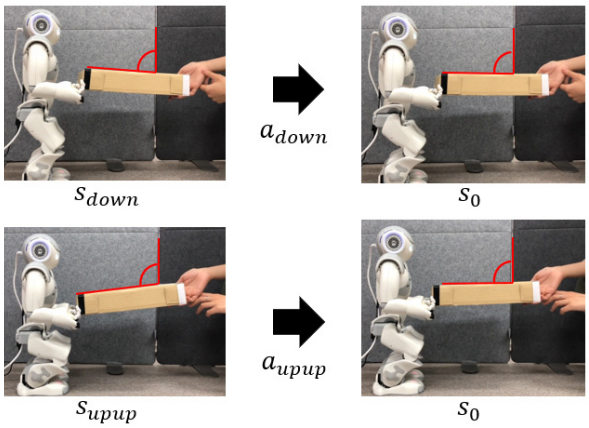
$$\text{Sum of } Q_t = \Sigma_{n=1}^{5} \Sigma_{m=1}^{5} Q_t(s_n, a_m) \tag{5}$$

식 (5)에서 Q_t는 t일 때의 Q 함수를 의미하며, 상태와 동작의 종류가 각 5가지이므로 n과 m을 사용하면 t일 때의 모든 Q값을 계산할 수 있다.

로봇 동작 예측 컴포넌트에서 제안한 모델의 안정된 최종 정책 수렴을 위해 Q-학습의 파라미터(학습율 α, 감가율 γ)를 각 0.1에서 0.9까지 0.1단위로 변화하며 실험한 결과의 일부는 [Fig. 5]와 같으며, 각 파라미터 변화당 100번씩, 실행횟수 2000회로 실험하였다. [Fig. 5]의 그래프들은 학습율 α와 감가율 γ가 각각 0.1, 0.5, 0.9일 때의 그래프이다. [Fig. 5]에서 학습율 α가 높아질수록 수렴 속도의 증가 양상을 보이며, 감가율 γ가 높아질수록 양수 업데이트 양상이 높아짐을 확인할 수 있다.

학습율 α는 식 (2)에서 새로 학습한 Q값의 업데이트 비율을 의미한다. 따라서 학습율 α가 높아짐에 따라 새로 학습한 Q값의 업데이트 비율이 높아진다.

예를 들어, [Fig. 5(c)], [Fig. 5(f)], [Fig. 5(i)]는 학습율 α가 0.1, 0.5, 0.9로 변화하며 감가율 γ가 0.9일 때를 나타낸다. 학습율 α가 0.1인 [Fig. 5(c)]에 비해 학습율 α가 높은 [Fig. 5(f)], [Fig. 5(i)]의 수렴이 더 빠르다. 이는 새로 학습한 Q값의 업데이트 사용 비율 증가로 업데이트 변화량 또한 증가하기 때문이라고 분석된다.



State	$s_{down} \rightarrow s_0$	$s_{upup} \rightarrow s_0$
Angle variation	88.8° → 90.3°	98.2° → 90.2°
Angular variation	+1.5°	-8.0°

[Fig. 6] Scenario implementation and table angles by ImageJ (Scenario experiment video link; http://ibot.knu.ac.kr/videocooperation.html)

하지만, 업데이트 변화량의 급격한 변화는 Q 함수의 불안정한 업데이트를 초래할 수 있다. [Fig. 5(d)], [Fig. 5(g)]를 보면 Q 함수의 수렴 이전까지 불안정한 업데이트가 관찰되며, [Fig. 5(f)], [Fig. 5(i)]보다 학습율 α가 0.1인 [Fig. 5(c)]에서 완만한 수렴 곡선이 관찰됨에 따라 학습율 α가 낮을 수록 Q 함수가 안정적으로 업데이트되는 것으로 해석할 수 있다. 이러한 현상은 [Fig. 5(a)], [Fig. 5(d)], [Fig. 5(g)]와 [Fig. 5(b)], [Fig. 5(e)], [Fig. 5(h)]에서도 동일하게 관찰할 수 있다.

감가율 γ는 식 (2)에서 미래의 영향력을 의미하며, 감가율 γ가 증가할 수록 미래 영향력이 증가한다. 예를 들어, 감가율 γ가 0.9와 0.1을 비교하면 감가율 γ가 0.9일 때 미래 상태의 영향력은 0.9, 0.81, 0.729, 0.6561로 점차 줄어들지만 감가율 γ가 0.1일 때 미래 상태의 영향력은 0.1. 0.01, 0.001, 0.0001로 급격하게 감소한다. 이는 감가율 γ가 높을 수록 먼 미래의 영향도 함께 받으며, 감가율 γ가 낮을 수록 가까운 미래만 영향을 미치는 것을 의미한다. 감가율 γ가 높을 수록 미래의 영향을 많이 받으므로, 업데이트 형태가 일정해지며 본 모델의 경우 Q 함수가 양수로 수렴한다. 이는 [Fig. 5]에서 관찰할 수 있다. [Fig. 5]에서 감가율 γ가 0.1일 때 [Fig. 5(a)]만 양수로 수렴하고 [Fig. 5(d)], [Fig. 5(g)]는 각기 다른 형태로 Q 함수가 수렴하는 반면, 감가율 γ가 0.5, 0.9인 그래프는 모두 양수 수렴을 나타낸다. 이러한 결과는 감가율 γ가 높을 수록 미래의 영향력이 커지므로 바로 앞의 미래에만 좌우되지 않고 안정적인 업데이트가 이루어 지는 것으로 분석된다. 이러한 일련의 실험 결과에 기반하여 본 모델은 안정적인 업데이트를 위해 학습율 α는 0.1, 감가율 γ는 0.9를 적용하였다. 로봇 동작 예측 컴포넌트에서 학습율 α를 0.1, 감가율 γ를 0.9로 적용한 모델의 최종 정책의 결정여부는 [Fig. 5(c)]를 통해 알 수 있다. [Fig. 5(c)]에서 실행 횟수 500번 이전에는 Q값의 합이 증가하다가 실행 횟수 500번 이후부터 Q값의 합의 증가량이 안정적으로 변화하면서 0.5에 가깝게 수렴하였다.

계산한 최적 정책에 기반하여 테이블 균형맞춤 작업 기술을 구현하여 프로토타입을 시연한 결과와 이미지 분석 프로그램인 ImageJ를 통해 붉은 선으로 표시된 테이블 각도를 측정한 결과는 [Fig. 6]과 같다. [Fig. 6]은 NAO의 구동 프로그램인 Choregraphe와 파이썬을 통해 로봇에 제안한 기술을 적용한 후 테이블 균형 맞춤 과제를 위한 로봇 동작을 구동하였다. 기울기가 90°일 때 테이블과 지면이 평행하며, 90°이하일 때 테이블은 로봇 방향보다 사람 방향이 더 낮고(s_{down}, $s_{downdown}$), 90°이상일 때 테이블은 로봇 방향보다 사람 방향이 더 높다(s_{up}, s_{upup}). [Fig. 6]에서 첫번째 이미지의 테이블 각도는 88.8°로 s_{down}이며 로봇이 a_{down} 구동 후 90.3°가 되어 s_0에 도달하였으며, 아래 이미지의 테이블 각도는 98.2°로 s_{upup}이며 로봇이 a_{upup} 구동 후 90.2°가 되어 s_0에 도달하였다. NAO 로봇이 실시간으로 촬영한 테이블 이미지를 이용하도록 본 기술을 구현하였으며, 기술 구현을 통한 로봇 구동 동영상은 해당 링크 (http://ibot.knu.ac.kr/videocooperation.html)를 통해 확인할 수 있다.

5. 결론 및 고찰

본 연구는 사람과 로봇이 협업하는 기술인 Q-학습기반 테이블 균형맞춤 기술을 제안하였다. 제안한 기술은 고성능 외부장치없이 로봇에 장착된 일반성능의 카메라만을 이용해 로봇의 테이블 균형맞춤 기술을 구현하였다. 실험 결과, 사람 동작 상태 인식 컴포넌트에서 테이블 분류의 10겹 교차검증 테스트셋 정확도는 93.9%를 보였으며, 로봇 동작 예측 컴포넌트에서는 Q-학습의 파라미터(학습율 α, 감가율 γ) 변화 실험 결과로 최적파라미터인 학습율 α 0.1과 감가율 γ 0.9를 얻고 제안한 모델에 적용하여 각 상태에서 안정된 로봇 동작을 구동하는 안정된 최적 정책을 얻을 수 있었다. 이에 따라 사람 동작 상태 인식 컴포넌트와 로봇 동작 예측 컴포넌트는 실시간으로 사람의 동작을 인식한 후 테이블 균형 맞춤을 위한 로봇 동작을 구동하여, 제안된 기술의 성능을 성공적으로 검증하였다. 하지만 제안된 기술은 1개의 테이블에 맞춰진 기술이며 제한된 사람 동작 상태와 로봇 동작을 이용한다는 한계를 가진다. 따라서 앞으로 제안한 기술에서 테이블 형태를 다양화와 사람 동작 상태와 로봇 동작의 범위를 확장, 학습 기술의 고도화를 진행함으로써 보다 섬세한 사람 로봇간 균형맞춤 기술을 구현하여 실생활에 적용할 수 있도록 연구를 확장하고자 한다.

References

[1] What's new, Atlas?, [Online], https://www.youtube.com/watch?v=fRj34o4hN4I, Accessed: Mar. 20, 2020.

[2] S. Wen, X. Chen, C. Ma, H. K. Lam, and S. Hua, "The Q-learning obstacle avoidance algorithm based on EKF-SLAM for NAO autonomous walking under unknown environments," *Robotics and Autonomous Systems*, vol. 72, pp. 29-36, Oct., 2015, DOI: 10.1016/j.robot.2015.04.003.

[3] M. Danel, "Reinforcement learning for humanoid robot control," *POSTER 2017, Prague, Czech Republic*, 2017, [Online], http://poseidon2.feld.cvut.cz/conf/poster/proceedings/Poster_2017/Section_IC/IC_021_Danel.pdf.

[4] F. Stulp, J. Buchli, E. Theodorou, and S. Schaal, "Reinforcement learning of full-body humanoid motor skills," *2020 10th IEEE-RAS International Conference on Humanoid Robots*, Nashville, TN, USA, pp. 405-410, 2010, DOI: 10.1109/ICHR.2010.5686320.

[5] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484-489, 2016, DOI: 10.1038/nature16961.

[6] K. Mülling, J. Kober, O. Kroemer, and J. Peters, "Learning to select and generalize striking movements in robot table tennis," *The International Journal of Robotics Research*, vol. 32, no. 3, pp. 263-279, 2013, DOI: 10.1177/0278364912472380.

[7] S. Debnath and J. Nassour, "Extending cortical-basal inspired reinforcement learning model with success-failure experience," *4th IEEE International Conference on Development and Learning and on Epigenetic Robotics*, Genoa, Italy, pp. 293-298, 2014, DOI: 10.1109/DEVLRN.2014.6982996.

[8] O. Aşık, B. Görer, and H. L. Akın, "End-to-End Deep Imitation Learning: Robot Soccer Case Study," *arXiv preprint arXiv:1807.09205*, 2018, [Online], https://arxiv.org/abs/1807.09205.

[9] K. Lobos-Tsunekawa, F. Leiva, and J. Ruiz-Del-Solar, "Visual navigation for biped humanoid robots using deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3247-3254, Oct., 2018, DOI: 10.1109/LRA.2018.2851148.

[10] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421-436, 2018, DOI: 10.1177/0278364917710318.

[11] P.-C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, "Repeatable Folding Task by Humanoid Robot Worker Using Deep Learning," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 397-403, Apr., 2017, DOI: 10.1109/LRA.2016.2633383.

[12] C. Wang, K. V. Hindriks, and R. Babuska, "Active learning of affordances for robot use of household objects," *2014 IEEE-RAS International Conference on Humanoid Robots*, Madrid, Spain, pp. 566-572, 2014, DOI: 10.1109/HUMANOIDS.2014.7041419.

[13] H. B. Suay and S. Chernova, "Effect of human guidance and state space size on Interactive Reinforcement Learning," *2011 IEEE International Symposium on Robot and Human Interactive Communication*, Atlanta, GA, USA, pp. 1-6, 2011, DOI: 10.1109/ROMAN.2011.6005223.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, pp. 1097-1105, 2012, [Online], http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networ.

[15] J. Merel, Y. Tassa, D. TB, S. Srinivasan, J. Lemmon, Z. Wang, G. Wayne, and N. Heess, "Learning human behaviors from motion capture by adversarial imitation," *arXiv preprint arXiv:1707.02201*, 2017, [Online], https://arxiv.org/abs/1707.02201.

[16] X. B. Peng, G. Berseth, and M. Van De Panne, "Terrain-adaptive locomotion skills using deep reinforcement learning," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1-12, 2016, DOI: 10.1145/2897824.2925881.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," *arXiv preprint arXiv:1312.5602*, Dec., 2013, [Online], https://arxiv.org/abs/1312.5602.

[18] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015, [Online], https://arxiv.org/abs/1509.02971.

[19] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust Region Policy Optimization," *32nd International Conference on Machine Learning*, pp. 1889-1897, 2015, [Online], http://proceedings.mlr.press/v37/schulman15.html.

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, 2017, [Online], https://arxiv.org/abs/1707.06347.

[21] A. Thobbi, Y. Gu, and W. Sheng, "Using human motion estimation for human-robot cooperative manipulation," *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Francisco, CA, USA, pp. 2873-2878, 2011, DOI: 10.1109/iros.2011.6094904.

[22] Y. Gu, A. Thobbi, and W. Sheng, "Human-robot collaborative manipulation through imitation and reinforcement learning," *2011 IEEE International Conference on Information and Automation*, Shenzhen, China, pp. 151-156, 2011, DOI: 10.1109/ICINFA.2011.5948979.

[23] Vicon, [Online], https://www.vicon.com, Accessed: Sep. 3, 2020.

[24] SoftBank Robotics, [Online], https://www.softbankrobotics.com, Accessed: Sep. 3, 2020.

[25] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, Dec., 2011, [Online], http://hdl.handle.net/2328/27165.

김 예 원

2017 경북대학교 신소재공학부 금속신소재 전공(학사)

2019~현재 경북대학교 기계공학과(석사)

관심분야: 강화학습, 인공지능, 협력 로봇, 딥러닝

강 보 영

2004 경북대학교 컴퓨터공학과박사

2006 KAIST ICC/서울대학교 박사후 연구원

2009 서울대학교 치의학전문 대학원 연구조교수

2018~현재 경북대학교 정교수

관심분야: 딥러닝, 강화학습, 인공지능, 협력 로봇