KIPS Trans. Softw. and Data Eng.
Vol.10, No.5 pp.179~186
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.5.179

Evaluation of Human Demonstration Augmented Deep Reinforcement Learning Policies via Object Manipulation with an Anthropomorphic Robot Hand

Na Hyeon Park† · Ji Heon Oh† · Ga Hyun Ryu†† · Patricio Rivera Lopez††† · Edwin Valarezo Añazco††† · Tae Seong Kim††††

ABSTRACT

Manipulation of complex objects with an anthropomorphic robot hand like a human hand is a challenge in the human-centric environment. In order to train the anthropomorphic robot hand which has a high degree of freedom (DoF), human demonstration augmented deep reinforcement learning policy optimization methods have been proposed. In this work, we first demonstrate augmentation of human demonstration in deep reinforcement learning (DRL) is effective for object manipulation by comparing the performance of the augmentation-free Natural Policy Gradient (NPG) and Demonstration Augmented NPG (DA-NPG). Then three DRL policy optimization methods, namely NPG, Trust Region Policy Optimization (TRPO), and Proximal Policy Optimization (PPO), have been evaluated with DA (i.e., DA-NPG, DA-TRPO, and DA-PPO) and without DA by manipulating six objects such as apple, banana, bottle, light bulb, camera, and hammer. The results show that DA-NPG achieved the average success rate of 99.33% whereas NPG only achieved 60%. In addition, DA-NPG succeeded grasping all six objects while DA-TRPO and DA-PPO failed to grasp some objects and showed unstable performances.

Keywords : Anthropomorphic Robot Hand, Deep Reinforcement Learning, Human Demonstration, Policy Optimization

휴먼형 로봇 손의 사물 조작 수행을 이용한 사람 데모 결합 강화학습 정책 성능 평가

박 나 현† · 오 지 헌† · 류 가 현†† · Patricio Rivera Lopez††† · Edwin Valarezo Añazco††† · 김 태 성††††

요 약

로봇이 사람과 같이 다양하고 복잡한 사물 조작을 하기 위해서는 휴먼형 로봇 손의 사물 파지 작업이 필수적이다. 자유도 (Degree of Freedom, DoF)가 높은 휴먼형(anthropomorphic) 로봇 손을 학습시키기 위하여 사람 데모(human demonstration)가 결합한 강화학습 최적화 방법이 제안되었다. 본 연구에서는 강화학습 최적화 방법에 사람 데모가 결합한 Demonstration Augmented Natural Policy Gradient (DA-NPG)와 NPG의 성능 비교를 통하여 행동 복제의 효율성을 확인하고, DA-NPG, DA-Trust Region Policy Optimization (DA-TRPO), DA-Proximal Policy Optimization (DA-PPO)의 최적화 방법의 성능 평가를 위하여 6 종의 물체에 대한 휴먼형 로봇 손의 사물 조작 작업을 수행한다. 학습 후 DA-NPG와 NPG를 비교한 결과, NPG의 물체 파지 성공률은 평균 60%, DA-NPG는 평균 99.33%로, 휴먼형 로봇 손의 사물 조작 강화학습에 행동 복제가 효율적임을 증명하였다. 또한, DA-NPG는 DA-TRPO와 유사한 성능을 보이면서 모든 물체에 대한 사물 파지에 성공하였고 가장 안정적이었다. 반면, DA-TRPO와 DA-PPO는 사물 조작에 실패한 물체가 존재하여 불안정한 성능을 보였다. 본 연구에서 제안하는 방법은 향후 실제 휴먼형 로봇에 적용하여 휴먼형 로봇 손의 사물 조작 지능 개발에 유용할 것으로 전망된다.

키워드 : 휴먼형 로봇, 강화학습, 사람 데모, 최적화 방법

1. 서 론

로봇 공학은 현대 공학 및 생산 업계에서 각광받고 있는 분야 중 하나다. 현재 로봇은 대체로 사람이 하기 힘든 정교한 작업과 사람이 다치기 쉬운 작업, 단순한 반복 노동 등에 효율적이기 때문에 다양한 산업에서 사용되고 있다.

최근 로봇의 쓰임새가 공장에서 부엌이나 카페 등 사람의 생활 반경으로 넓어지고 있다. 로봇이 사람 중심의 환경에서 물체를 재배치하거나 문의 손잡이를 여는 등의 정교한 일을 수행하기 위해서는 휴먼형(anthropomorphic) 로봇의 파지 작업이 필수적이다. 휴먼형 로봇 손은 자유도(Degree of Freedom, DoF)가 높

※ 이 논문은 2019년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2019R1A2C1003713).
※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 디지털콘텐츠원천기술개발사업의 연구결과로 수행되었음(IITP-2017-0-00655).
※ 이 논문은 2020년 한국정보처리학회 추계학술발표대회의 우수논문으로 "휴먼형 로봇 손의 사물 조작 수행을 이용한 인간 행동 복제 강화학습 정책 최적화 방법 성능 평가"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 경희대학교 전자정보융합공학과 석사과정
†† 비 회 원 : 경희대학교 전자정보융합공학과 석사과정
††† 비 회 원 : 경희대학교 전자정보융합공학과 석·박사통합과정
†††† 비 회 원 : 경희대학교 생체의공학과 및 전자정보융합공학과 교수
Manuscript Received : December 18, 2020
First Revision : January 29, 2021
Accepted : February 16, 2021
* Corresponding Author : Tae-Seong Kim(tskim@khu.ac.kr)

아 최근 로봇 사물 조작 강화학습에 Natural Policy Gradient (NPG)[1], Trust Region Policy Optimization(TRPO)[2], Proximal Policy Optimization (PPO)[3] 등 다양한 강화학습 최적화 방법이 제시되었다.

그러나 로봇 손을 학습시키더라도 샘플 복잡도와 절대적인 학습 시간이 기하급수적으로 늘어나는 문제점이 존재하여, 이를 줄이기 위해 로봇 손 관절의 자유도를 줄이는 등의 제한이 존재하였다[4]. 이러한 문제점들을 해결하기 위해 강화학습에 사람 데모 (human demonstration)를 결합시키는 행동 복제 (Behavior Cloning, BC) 방법이 제안되었다[5,6]. 사람 데모는 사람의 물체 파지 정보를 측정한 것으로, 이를 강화학습에 적용하면 문제점이었던 샘플 복잡도와 학습 시간이 현저하게 줄어든다. 또한 사람 데모가 사람이 물체를 잡는 손 모양에 대한 정보를 제공하기 때문에 로봇 손은 물체를 사람처럼 자연스럽게 잡을 수 있도록 학습된다는 등의 장점이 있어, Demonstration Augmented Policy Gradient (DA-PG) 등의 사람 데모와 결합한 강화학습 정책 최적화 방법이 제안되었다[5].

본 연구에서는 강화학습에 있어 사람 데모의 효용성 평가를 위하여 DA-NPG와 NPG의 성능을 비교하고, 강화학습의 NPG, TRPO, PPO의 정책 최적화 방법에 사람 데모를 결합한 DA-NPG, DA-TRPO 및 DA-PPO 의 각 정책 최적화 성능을 휴먼형 로봇 손의 6종 사물 조작(파지 및 재배치) 작업의 수행을 통하여 평가하였다.

2. 관련 연구

2.1 휴먼형 로봇 손

사람의 손과 동일하게 동작할 수 있는 정교한 휴먼형 로봇 손을 만들고자 하는 노력은 1900년대 초부터 꾸준히 있어왔다[7-12]. 최근 각종 센서와 소재의 발전으로 인해 더욱 세밀한 힘 조절과 자연스러운 손동작이 가능한 로봇 손이 제작되었다. Utah/MIT hand[13]부터 Shadow Dexterous Hand[14] 등의 로봇 손은 많은 actuator와 센서들을 장착한 정교한 디자인으로 사람 손의 외형과 관절, 움직임을 복제하였다. 최근에는 사람 손이 물체에 맞춰 유연하게 동작하는 것에 기반하여 기존의 단단한 소재로 제작된 로봇 손에서 벗어나 soft actuator와 유연한 재질을 사용한 SoftHand형 로봇 손이 연구되고 있다[9,15].

또한 휴먼형 로봇 손을 사람의 손처럼 동작하려면 정교한 조작을 할 수 있도록 적절한 학습 및 구동 방법이 필요하다. 최근 딥러닝 기법의 등장으로 다양하고 복잡한 작업을 수행할 수 있는 휴먼형 로봇 손의 구동이 가능해졌다[16,17].

2.2 강화학습 사물 조작 지능

휴먼형 로봇 손의 사물 조작을 학습시키는 방법으로 강화학습이 많이 활용되었다. 강화학습(Reinforcement Learning, RL)은 사람의 개입을 최소화하면서 agent가 시행착오를 통해 다양한 작업을 학습할 수 있는 특징으로 인해 자율주행 및 자동화 시스템에서 많이 사용되고 있다[18,19]. 특히, 강화학습과 딥러닝의 이점을 활용하는 심층 강화학습(Deep Reinforcement Learning, DRL)은 샘플 복잡도, 메모리 복잡도 등의 문제들로 인해 강화학습만으로 해결하지 못했던 비디오 게임과 같은 고차원적인 작업을 수행할 수 있게 하였다[20,21].

심층 강화학습으로 학습을 진행한다고 해도 여전히 학습된 로봇이 수행하는 작업의 질은 높은 수준의 kinematic 방법으로 설계된 로봇이나 수동으로 설계된 컨트롤러에 비해 현저히 떨어진다[22]. 또한 휴먼형 로봇과 같이 자유도가 높은 로봇으로 사물 조작 및 파지 등의 고차원 작업은 샘플 복잡도가 높고 결과가 안정적이지 못하는 등의 문제점이 존재하여 심층 강화학습만으로 해결하는 데에 어려움이 있다.

이러한 문제점들을 해결하기 위해서 사람이 특정 물체를 잡은 부분에 대한 정보를 제공하는 haptic maps를 사용하는 DGPN 알고리즘과 특정 물체에 대해서 다양한 로봇 손 모양을 제안하는 natural hand pose estimation 등의 사전 정보를 사용하는 연구가 제안되었다[23,24]. 제안된 방법들은 휴먼형 로봇 손의 컵, 우유갑, 캔 등의 다양한 물체 파지 작업에서 좋은 성능을 보였다.

또한 심층 강화학습에 사람 데모를 결합한 DQfD, DDPGfD 등의 방법들이 제안되었다[25-29]. 제안된 방법들은 간단한 로봇 작업과 사물 조작 및 robotic block stacking 등의 작업에서 안정적인 결과를 보였다.

3. 사람 데모 결합 강화학습 정책

본 연구에서는 휴먼형 로봇 손의 강화학습 정책인 NPG, TRPO 및 PPO 방법에 사람 데모를 결합한 DA-NPG, DA-TRPO 및 DA-PPO 정책을 구현하여 사물 조작 최적화를 진행한다. 6종의 사물 조작을 이용한 강화학습 정책의 성능 평가를 통하여 최적의 행동 복제 강화학습 정책을 도출하고자 한다.

3.1 사람 데모 측정 시스템

본 연구에서는 행동 복제를 시행하기에 앞서 각 물체마다 사람 데모를 측정했다. 사람 데모를 측정하기 위해 적외선 LED 센서를 이용하여 손과 각 관절의 위치 정보를 x, y, z (mm) 로 제공하는 Leap Motion controller[30]와 로봇 시뮬레이션 패키지인 MuJoCo (Multi-joint dynamics with Contact) 시뮬레이터[31]에서 ADROIT 로봇 손 시뮬레이션 모델을 실시간으로 연동하였다[32]. ADROIT 로봇 손은 손의 위치 정보와 각 관절의 각도를 입력받아서 움직이는데, Leap Motion controller의 손 모양과 ADROIT 로봇 손을 일치시키기 위해서 3D 관절 위치 정보를 각도로 변환한 후 각 관절의 조절 범위에 맞춰서 각도를 정규화했다. Fig. 1에 사람 데모 측정 환경을 도시하였다.

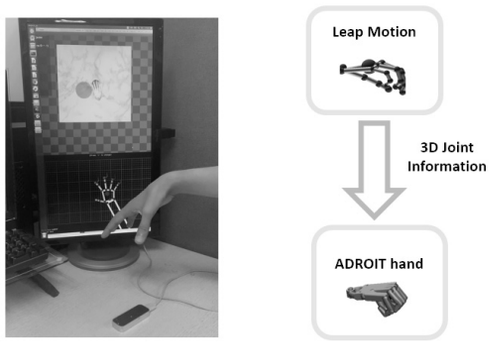

Fig. 1. (A) Human Demonstration System (B) Leap Motion Controller Provides 3D Joint Information to the ADROIT Hand in Real Time to Make the ADROIT Hand Synchronize with the Human Hand

3.2 모방 학습과 결합한 강화학습

모방 학습(Imitation Learning)은 사람 데모와 같이 주어진 전문가의 동작을 흉내 내어 학습하는 것을 말한다[33]. 본 연구에서 사용한 모방 학습 방법은 행동 복제(BC)로, 로봇 손이 사람 데모를 따라 하기 위해 지도 학습(supervised learning)을 통해 정책을 업데이트한다. 행동 복제를 본격적인 강화학습 이전에 수행하여 정책을 초기화함으로써, 샘플 복잡도와 학습 시간을 효과적으로 줄이고 로봇 손이 파지 작업과 관계없는 경험을 탐험(exploration)하느라 시간을 허비하지 않도록 한다. 그 식은 다음과 같다.

$$\underset{\theta}{maximize} \sum_{(s,a)\in\rho_D} ln\pi_\theta(a|s) \tag{1}$$

여기서 $\rho_D = (s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)}, r_t^{(i)})$ 는 사람 데모의 데이터셋을 나타낸다[5]. 하지만 행동 복제를 이용하여 정책을 초기화하는 것만으로는 로봇 손 파지와 같은 복잡한 작업을 수행하는 데에 있어서 큰 효과를 기대하기 어렵다.

3.3 사람 데모와 결합한 강화학습 최적화 알고리즘

본 연구에서 인간 행동 복제 강화학습의 전체적인 과정은 Fig. 2에, 로봇 손 사물 재배치 및 파지 작업에 사용한 물체들의 3D 모델은 Fig. 3에 도시했다.

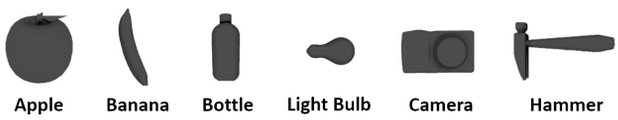

Fig. 3. 3D Models of Six Objects used in Object Manipulation using an Anthropomorphic Robot Hand

물체의 3D 모델을 이용하여 물체에 대한 사람 데모를 만들고 사람 데모로 행동 복제를 시행하여 가중치 값을 초기화해준다[5]. 로봇 손은 주어진 환경과 상호작용하여 보상을 높이는 방향으로 정책을 최적화하는데, 이 과정에서 사람 데모를 입력하여 로봇 손의 학습을 돕는다. 여기서 정책을 최적화하는 방법은 DA-NPG, DA-TRPO 및 DA-PPO 세 가지로 구분되며, 학습이 끝나면 최적화 정책이 도출된다.

1) DA-NPG

NPG는 기존의 standard gradient descent rule에 기반을 둔 Policy Gradient(PG) 방법에 Natural Gradient 기법을 접목하여 steepest descent direction으로 학습이 진행될 수 있도록 한 방법이다[1]. 이는 로봇 손이 학습하면서 그저 전보다 나은 정책이 아니라 가장 좋은 정책을 찾을 수 있도록 최적화한다. NPG의 gradient 식을 DA-NPG로 확장하면 다음과 같다[5].

$$g_{aug} = \sum_{(s,a)\in\rho_\pi} \nabla_\theta ln\pi_\theta(a|s) A^\pi(s,a) + \sum_{(s,a)\in\rho_D} \nabla_\theta ln\pi_\theta(a|s) w(s,a) \tag{1}$$

$$w(s,a) = \lambda_0 \lambda_1^k \max_{(s',a')\in\rho_\pi} A^\pi(s',a') \; \forall (s,a) \in \rho_D \tag{2}$$

Equation (1)은 정책 π의 데이터셋 ρ_π와 사람 데모의 데이터셋 ρ_D에 대한 부분으로 나눠진다. 데이터셋은 한 시점에 대한 상태(state, s)와 현 상태에서 로봇 손이 취하는 행동(action, a)의 상태-행동 쌍(s,a)을 포함하며, $A^\pi(s,a)$는 정책 π에 대한 어드밴티지 함수이다. $w(s,a)$는 사람 데모 가중치 함수이며 Equation (2)의 λ_0와 λ_1은 hyperparameter, k는 iteration을 의미한다. $\lambda_0 = 1.0$, $\lambda_1 = 0.95$로 설정함에 따라

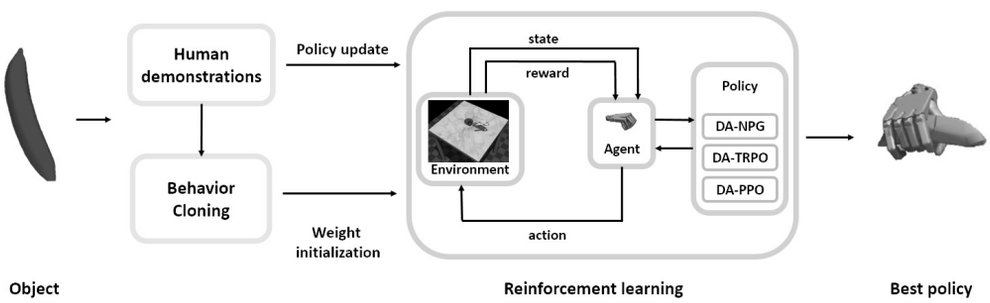

Fig. 2. Overall Process of Human Demonstration Augmented Deep Reinforcement Learning

iteration이 커질수록 $w(s,a)$의 값이 작아져 사람 데모에 대한 가중치가 줄어든다.

2) DA-TRPO

TRPO는 NPG 방법에 Trust Region이라는 constraint를 추가하여 정책의 변화 정도를 제한한 기법이다[2]. 강화학습 중 정책을 업데이트할 때 정책이 변화화면서 state visitation frequency 또한 $\rho_{\pi}(s)$에서 $\rho_{\tilde{\pi}}(s)$로 바뀌게 되는데, 이 때문에 기울기 최적화에 어려움이 생긴다. 따라서 기울기 최적화의 용이함을 위해 이 변화를 무시하고 이전 정책의 $\rho_{\pi}(s)$를 그대로 사용하기 위해서 정책이 최적화되는 정도를 Trust Region이라는 constraint로 제한한다. TRPO의 surrogate objective function은 다음과 같다.

$$\begin{aligned} &\underset{\theta}{\text{maximize}}\ \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t\right] \\ &\text{subject to}\ \hat{\mathbb{E}}_t\left[KL\left[\pi_{\theta_{old}}(\cdot\,|s_t),\pi_\theta(\cdot\,|s_t)\right]\right] \le \delta \end{aligned} \tag{3}$$

여기서 KL은 KL divergence로 업데이트하기 전의 정책과 새로운 정책 간의 변화 정도를 측정하며 δ는 Trust Region의 제한 범위를 나타낸다. DA-TRPO는 Equation (3)의 objective function의 graident에 Equation (1)을 적용한다.

3) DA-PPO

PPO는 clipped surrogate objective를 통해 TRPO의 장점은 유지하면서 TRPO의 단점인 복잡한 계산과 샘플 복잡도를 줄인 방법이다[3]. TRPO는 안정적이고 좋은 성능을 보이는 대신에 Fisher Information Matrix 등의 2차 근사 방법으로 정책을 최적화하기 때문에 계산이 복잡하다. 따라서 이러한 계산 복잡도를 감소시키기 위해서 clipping 기법을 이용하여 선형 근사 방법으로 TRPO의 Trust Region과 유사하게 제한을 둔 방법이다. PPO의 surrogate objective function은 다음과 같다.

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t \tag{4}$$

$$\underset{\theta}{\text{maximize}}\ \hat{\mathbb{E}}_t\left[min\left(r_t(\theta), clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{5}$$

여기서 ϵ은 hyperparameter로, Trust Region의 제한 범위를 지정해주는 δ와 유사한 역할을 한다. 본 연구에서는 [3]에 따라 0.2로 지정하였다. DA-PPO는 Equation (4)의 probability ratio function에 Equation (1)을 적용한다.

4. 실험 결과

4.1 DA-NPG vs. NPG의 성능 비교

기존 NPG와 사람 데모와 결합한 DA-NPG로 각각 학습시킨 로봇 손의 사물 파지 작업 결과를 비교한다. Fig. 4는 6종 사물 조작 작업에 대해 학습이 완료된 로봇 손이 사물을 잡는 손 모양을 보여준다. NPG로 학습시킨 로봇 손이 물체를 잡는 모습은 사람이 잡는 모습과 달리 관절이 부자연스럽고 물병과 망치 파지 작업에 실패하였지만, DA-NPG로 학습시킨 로봇 손은 사람과 같이 안정적으로 물체를 잡으며 모든 물체 파지 작업에 성공하였다.

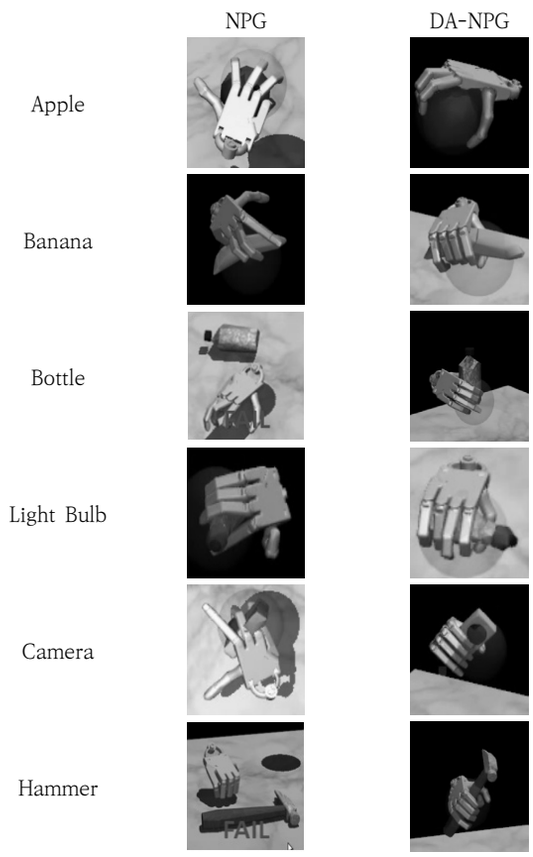

Fig. 4. Results of the Object Manipulation using NPG and DA-NPG

Fig. 5의 그래프는 DA-NPG와 NPG의 사물 조작 강화학습 보상 그래프를 보여준다. 그래프를 보면 DA-NPG와 NPG의 성능은 사과와 카메라에서는 유사하고 바나나와 전구는 DA-NPG가 우세한 경향을 보이며, 물병과 망치는 NPG는 전혀 학습되지 않으면서 DA-NPG만이 사물 파지 및 재배치 작업에 성공한다. 학습 전반적으로 DA-NPG가 NPG에 비해 성능이 뛰어나며, 파지에 성공하기까지 학습에 필요한 시간도 적게 걸린다.

Table 1은 NPG와 DA-NPG로 학습시킨 로봇 손을 각 물체에 대하여 100번 사물 조작 시도한 것에 대하여 물체 파지 작업 성공 횟수를 나타낸다. NPG의 평균 성공률은 60%, DA-NPG의 평균 성공률은 99.33%로 DA-NPG가 성공률이 월등히 높은 것을 확인할 수 있다. Fig. 5에서 보상 그래프가 전혀 올라가지 않았던 물병과 망치는 Table 1의 파지 시뮬레이션 또한 단 한 번도 성공하지 못하였다. 따라서 사람 데모는 로봇 손의 사람과 같은 자연스러운 손동작을 가능하게 하고, 최적화 정책의 최종 보상 및 파지 성공률을 향상하며 학습 시간을 줄인다는 것을 알 수 있다.

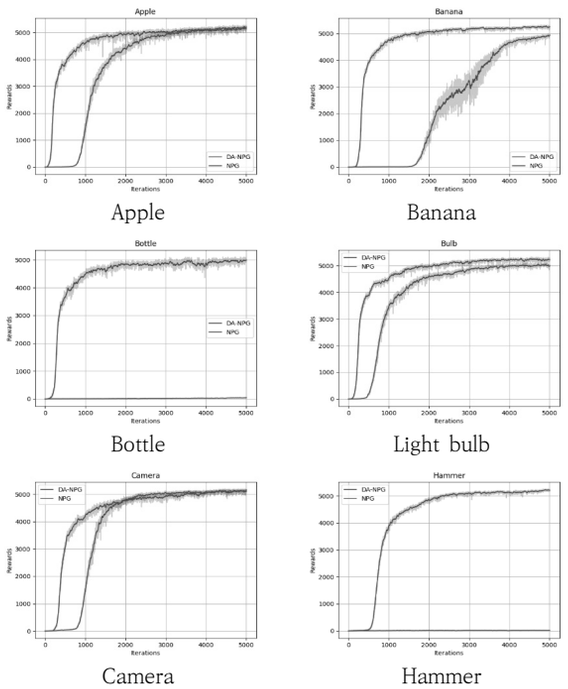

Fig. 5. DRL Reward Plots for Six Objects using DA-NPG(green) and NPG(red)

Table 1. Success Rate of Grasping Six Objects 100 Trials using NPG and DA-NPG

	NPG	DA-NPG
Apple	99	99
Banana	98	100
Bottle	0	99
Light bulb	82	99
Camera	81	99
Hammer	0	100
Average (%)	60	99.33

4.2 정책 최적화 방법에 따른 성능 비교

DA-NPG, DA-TRPO 및 DA-PPO로 각각 학습시킨 로봇 손의 사물 파지 작업 결과를 비교한다. Fig. 6은 6종 사물 조작 작업에 대해 DA-NPG, DA-TRPO 및 DA-PPO로 학습시킨 로봇 손의 파지 수행 결과를 보여준다. DA-NPG, DA-TRPO, 그리고 DA-PPO는 모두 사람 데모가 추가되어 학습이 진행되기 때문에, 이 방법들로 학습시킨 로봇 손이 물체를 잡는 손 모양은 Fig. 4의 NPG로 학습시킨 로봇 손의 손 모양보다 비교적 더 자연스러운 것을 Fig. 6의 파지 수행 결과를 통해 확인할 수 있었다.

Fig. 7은 정책 최적화 방법에 따른 6종 사물 조작 작업의 보상 그래프를 도시했다. Fig. 7의 보상 그래프를 보면 전반적인 물체들에 대해 DA-TRPO의 성능이 가장 높고 안정적이며 Fig. 6과 같이 사물을 잡는 손 모양도 자연스럽지만, 카메라 조작 작업에 실패하였고, DA-NPG는 DA-TRPO와 비슷한 성능을 보이면서 모든 사물 조작에 성공하였다. 반면에

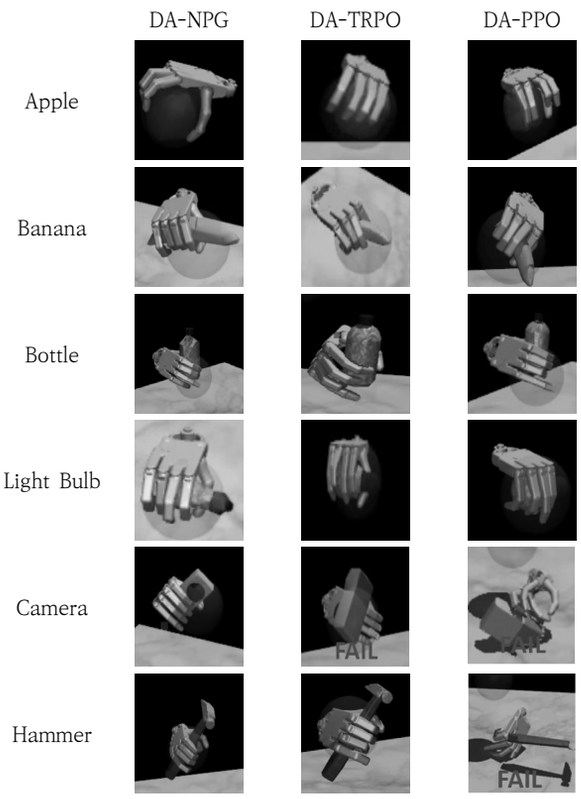

Fig. 6. Results of the Object Manipulation using DA-NPG, DA-TRPO, and DA-PPO

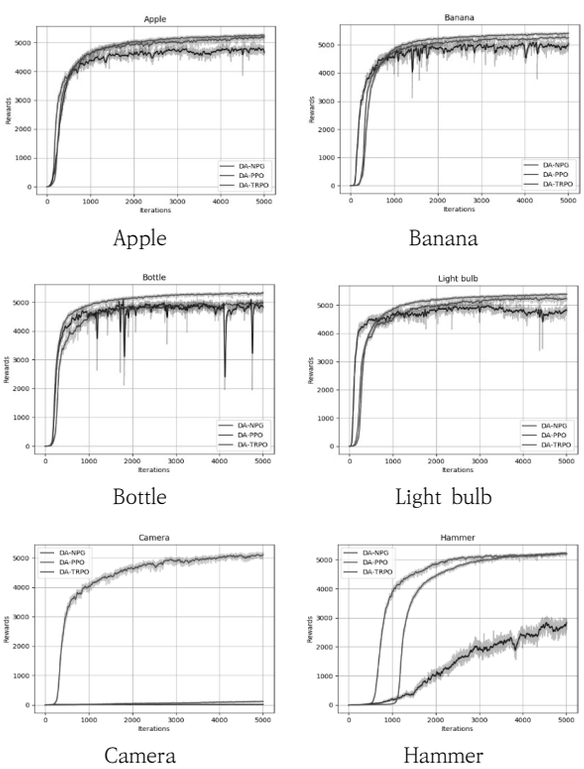

Fig. 7. DRL Reward Plots for Six Objects using DA-NPG(green), DA-TRPO(red), and DA-PPO(blue)

Table 2. Success Rate of Grasping Six Objects 100 Trials using DA-NPG, DA-TRPO, and DA-PPO

	DA-NPG	DA-TRPO	DA-PPO
Apple	99	100	97
Banana	100	100	92
Bottle	99	99	97
Light bulb	99	99	93
Camera	99	0	0
Hammer	100	100	67
Average(%)	99.33	83	74.33

DA-PPO는 바나나, 병, 전구 등의 작업에 대해서 비교적 빠르게 학습되는 경향을 보였으나, 카메라와 망치 작업에 실패하였고, DA-NPG와 DA-TRPO에 비하여 정책이 업데이트되는 과정에서 받는 보상의 그래프 경향이 가장 불안정한 것을 확인할 수 있다. 따라서 로봇 사물 조작 작업에 대한 정책 최적화 알고리즘 중 성능이 가장 안정적인 것은 DA-NPG이다.

Table 2는 DA-NPG, DA-TRPO, 그리고 DA-PPO 의 6종 물체에 대한 파지 작업 성공률을 나타내었다. DA-NPG와 DA-TRPO가 전반적인 물체 파지 작업에 대하여 높은 성공률을 보였지만, DA-TRPO는 카메라 파지 작업을 전혀 수행하지 못하여 평균 성공률이 83%가 된 것을 확인할 수 있다. 또한 DA-PPO의 성공률을 보면 DA-NPG, DA-TRPO에 비하여 전반적으로 성공률이 저조하고 카메라 파지 작업을 전혀 수행하지 못했으며, 망치 파지 작업 또한 DA-NPG와 DA-TRPO에 비해 현저히 낮은 성공률을 보여 평균 성공률 또한 74.33%로 가장 낮은 것을 확인할 수 있다.

4.3 정책 최적화 방법들의 학습시간 비교

Table 3은 NPG, DA-NPG, DA-TRPO, DA-PPO의 강화학습의 학습시간을 작성한 것이다. 모든 최적화 방법에서 성공한 물체인 사과, 바나나, 그리고 전구를 학습시켰을 때, 한 번 학습하는 데 걸리는 시간 평균(초), 보상이 3000에 처음으로 도달할 때까지 걸린 시간(시), 전체 학습 시간(시)을 각각 최적화 방법별로 평균을 취했다. 그 결과, 한 번 학습하는데 걸리는 시간과 총 학습 시간이 가장 짧은 것은 DA-TRPO, 보상이 3000에 도달할 때까지 걸린 시간이 가장 짧은 것은 DA-PPO인 것을 확인할 수 있다.

5. 고 찰

세 가지 정책 최적화 방법들의 실험 결과에 대한 고찰을 진행한다. Table 3을 보면 사람 데모를 사용하지 않은 NPG가 사람 데모를 사용한 다른 알고리즘들보다 전체적으로 학습 시간이 길게 측정되었다. 특히 보상 3000에 도달할 때까지 걸리는 시간은 약 4.5~4.9시간 정도의 차이를 보이는데 이는 사람 데모를 기반으로 행동 복제를 통해 정책을 초기화해주고, 최적화 방법에 사람 데모를 결합하여 지속해서 정보를 제공함으로써 로봇 손이 학습할 방향을 제시해주었기 때문이다.

Table 3. Training Time of Optimization Algorithms

Time / Algorithm	Average per iteration (sec)	Training time for up to reward 3000 (hour)	Overall (hour)
NPG	22.25	5.52	30.90
DA-NPG	18.08	1.06	25.12
DA-TRPO	14.59	1.06	20.26
DA-PPO	14.69	0.66	20.41

또한, DA-NPG보다 DA-TRPO와 DA-PPO가 학습하는 데에 시간이 덜 걸리는 것을 확인할 수 있는데, 이는 DA-TRPO와 DA-PPO가 정책을 최적화할 때 각각 Trust Region이라는 constraint와 clipping surrogate objective를 통해 정책의 변화 정도를 제한하는 방식을 통해 기울기 최적화를 쉽게 수행했기 때문이다.

DA-PPO가 보상 3000에 도달할 때까지 걸리는 시간은 평균 0.6시간으로 다른 방법들에 비해 덜 걸리는데, 이는 clipping 기법을 통한 1차 근사로 최적화를 진행하여 DA-TRPO보다 계산 복잡도를 낮췄기 때문이다. 하지만 한 번 학습하는 데 걸리는 학습 시간이나 전체 학습 시간은 DA-TRPO와 유사한 것을 Table 3에서 볼 수 있는데, 이는 Fig. 7과 같이 DA-PPO가 불안정하게 정책을 최적화하면서 시간 지연이 발생한 것이다.

성능상 가장 안정적인 최적화 방법은 DA-NPG로 결과를 도출하였으나, 학습 시간과 성능을 동시에 고려할 때 Table 3의 결과에 따라 DA-NPG와 DA-TRPO가 총 학습 시간이 약 5시간 정도 차이가 나므로, 적당한 성능과 비교적 학습 시간이 빠른 DA-TRPO를 타협점으로써 선택할 수 있다.

6. 결 론

본 연구에서는 DA-NPG와 NPG의 학습 성능을 비교하여 사람 데모의 효용성을 입증하고, 행동 복제 방법과 DA-NPG, DA-TRPO, DA-PPO를 이용한 강화학습으로 학습시킨 로봇 손이 수행하는 6종 물체에 대한 사물 조작 작업의 학습 성능을 평가하였다.

학습 결과, DA-NPG는 NPG보다 성능 수치가 높은 것과 더불어 물체를 잡는 손 모양 또한 더욱 자연스러웠고 DA-NPG는 DA-TRPO와 유사한 성능을 보이면서, 일부 사물에 대한 작업에 실패한 DA-TRPO와 DA-PPO와 달리 모든 물체에 대한 작업에 성공하였기 때문에 그 성능이 가장 안정적이었다. 또한, NPG에 비하여 사람 데모를 사용한 DA-NPG, DA-TRPO 및 DA-PPO의 학습 시간이 전체적으로 짧다는 것을 확인하였다.

References

[1] S. Kakade, “Natural Policy Gradient,” *Neural Information Processing systems (NIPS)*, 14:1531-1538. 2001.

[2] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. "Trust Region Policy Optimization," *Proceedings of the 32nd International Conference on Machine Learning*, PMLR 2015, 37: 1889-1897.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," arXiv:1707.06347v2 [cs.LG].

[4] S. Gu, E. Holly, T. Lillicrap, S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, pp.3389-3396, 2017.

[5] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine. "Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations," arXiv:1709.10087v2 [cs.LG]. 2018.

[6] A. Gupta, C. Eppner, S. Levine, and P. Abbeel, "Learning dexterous manipulation for a soft robotic hand from human demonstrations," *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, pp.3786-3793, 2016.

[7] C. Piazza, G. Grioli, M. G. Catalano, and A. Bicchi, "A century of robotic hands," *Annual Review of Control, Robotics, and Autonomous Systems*, Vol.2, pp.1-32, 2019.

[8] Zhou, Jianshu, et al., "A soft-robotic approach to anthropomorphic robotic hand dexterity," *IEEE Access*, Vol.7, pp.101483-101495, 2019.

[9] C. Piazza, et al., "The SoftHand Pro-H: a hybrid body-controlled, electrically powered hand prosthesis for daily living and working," *IEEE Robotics & Automation Magazine*, Vol.24, No.4, pp.87-101, 2017.

[10] A. Kargov, et al., "Development of an anthropomorphic hand for a mobile assistive robot," *IEEE In 9th International Conference on Rehabilitation Robotics*, New York, 2005. pp.182-186.

[11] N. Correll, et al., "Analysis and observations from the first Amazon Picking Challenge," *IEEE Transactions on Automation Science and Engineering*, Vol.15, No.1, pp.172-188, 2018.

[12] Kontoudis GP, Liarokapis MV, Zisimatos AG, Mavrogiannis CI, Kyriakopoulos KJ. "Opensource, anthropomorphic, underactuated robot hands with a selectively lockable differential mechanism:towards affordable prostheses," In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.5857-5862. New York: IEEE. 2015.

[13] Jacobsen S, Iversen E, Knutti D, Johnson R, Biggers K. 1986. "Design of the Utah/M.I.T. dextrous hand." In *1986 IEEE. International Conference on Robotics and Automation*, Vol.3, pp.1520-1532, New York: IEEE.

[14] Shadow Robot Co. 2018. Shadow Dexterous Hand. Shadow Robot Company. https://www.shadowrobot.com/products/dexterous-hand

[15] A. Firouzeh, and J. Paik, "Grasp mode and compliance control of an underactuated origami gripper using adjustable stiffness joints," *IEEE/ASME Transactions on Mechatronics*, Vol.22, No.5, pp.2165-2173, 2017. doi: 10.1109/TMECH.2017.2732827.

[16] Billard, Aude, and Danica Kragic, "Trends and challenges in robot manipulation," *Science*, 364.6446, 2019.

[17] Chao, Ya, Xingchen Chen, and Nanfeng Xiao,"Deep learning-based grasp-detection method for a five-fingered industrial robot hand," *IET Computer Vision*, Vol.13, No.1, pp.61-70, 2018.

[18] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," *IEEE International Conference on Robotics and Automation, 2004. Proceedings*. ICRA '04. 2004, New Orleans, LA, USA, 2004, pp. 2619-2624 Vol.3, doi: 10.1109/ROBOT.2004.1307456.

[19] A.Y. Ng, et al., "Autonomous Inverted Helicopter Flight via Reinforcement Learning," In: Ang M.H., Khatib O. (eds) Experimental Robotics IX. Springer Tracts in Advanced Robotics, Vol.21. Springer, Berlin, Heidelberg. 2006. https://doi.org/10.1007/11552246_35.

[20] V. Mnih, et al., "Human-level control through deep reinforcement learning," *Nature*, Vol.518, No.7540, pp.529-533, 2015. https://doi.org/10.1038/nature14236.

[21] D. Silver, et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, Vol.529, No.7587, pp.484-489, 2016. https://doi.org/10.1038/nature16961

[22] Nicolas Heess, Dhruva TB, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, S. M. Ali Eslami, Martin A. Riedmiller, and David Silver, "Emergence of Locomotion Behaviours in Rich Environments," 2017. CoRR abs/1707.02286. arXiv:1707.02286

[23] E. Valarezo Añazco, et al., "Natural object manipulation using anthropomorphic robotic hand through deep reinforcement learning and deep grasping probability network," *Applied Intelligence*, Vol.51, No.2, pp.1041-1055, 2021. https://doi.org/10.1007/s10489-020-01870-6

[24] Edwin Valarezo Añazco, Patricio Rivera Lopez, Hyemin Park, Nahyeon Park, Jiheon Oh, Sangmin Lee, Kyungmin Byun, and Tae-Seong Kim. "Human-like Object Grasping and Relocation for an Anthropomorphic Robotic Hand with Natural Hand Pose Priors in Deep Reinforcement Learning," In *Proceedings of the 2019 2nd International Conference on Robot Systems and Applications (ICRSA 2019)*. Association for Computing Machinery, New York, NY, USA, 46-50. DOI:https://doi.org/10.1145/3378891.3378900

[25] Gao, Yang, et al.,"Reinforcement learning from imperfect demonstrations," arXiv preprint arXiv:1802.05313. 2018.

[26] Vecerik, Mel, et al.,"Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards," arXiv preprint arXiv:1707.08817. 2017.
[27] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming Exploration in Reinforcement Learning with Demonstrations," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, 2018, pp.6292-6299, doi: 10.1109/ICRA.2018.8463162.
[28] Hester, Todd, et al.,"Learning from demonstrations for real world reinforcement learning," 2017.
[29] Osa, Takayuki, Jan Peters, and Gerhard Neumann, "Hierarchical reinforcement learning of multiple grasping strategies with human instructions," *Advanced Robotics*, Vol.32, No.18, pp.955-968, 2018.
[30] Leap Motion [Internet], https://www.ultraleap.com/
[31] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura, 2012, pp.5026-5033.
[32] V. Kumar, Z. Xu, and E. Todorov, "Fast, strong and compliant pneumatic actuation for dexterous tendon-driven hands," *2013 IEEE International Conference on Robotics and Automation*, Karlsruhe, pp.1512-1519, 2013.
[33] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne, "Imitation Learning: A Survey of Learning Methods," *ACM Computing Surveys*, Vol.50, No.2, Article 21, pp.1-35, 2017.

박 나 현
https://orcid.org/0000-0003-4349-498X
e-mail : nhpark@khu.ac.kr
2019년 경희대학교 생체의공학과(학사)
2019년 ~ 현 재 경희대학교 전자정보융합공학과 석사과정
관심분야 : 인공지능, 로봇지능, 강화학습, 머신러닝, 컴퓨터 비전

오 지 헌
https://orcid.org/0000-0003-0488-116X
e-mail : dhwlgjs3@khu.ac.kr
2020년 경희대학교 생체의공학과(학사)
2020년 ~ 현 재 경희대학교 전자정보융합공학과 석사과정
관심분야 : 인공지능, 로봇지능, 강화학습, 머신러닝, 컴퓨터 비전

류 가 현
https://orcid.org/0000-0002-4609-4394
e-mail : yugacandy@khu.ac.kr
2020년 경희대학교 생체의공학과(학사)
2020년 ~ 현 재 경희대학교 전자정보융합공학과 석사과정
관심분야 : 인공지능, 로봇지능, 강화학습, 머신러닝

Patricio Rivera Lopez
https://orcid.org/0000-0001-6440-5478
e-mail : patoalejor@khu.ac.kr
2015년 Univ. of the Armed-Forces-ESPE, Electronics, Automation and Control Engineering(학사)
2016년 ~ 현 재 경희대학교 전자정보융합공학과 석·박사통합과정
관심분야 : Signal & Depth image Processing, Reinforcement Learning, Autonomous Robotic Systems.

Edwin Valarezo Añazco
https://orcid.org/0000-0003-0077-8528
e-mail : edgivala@khu.ac.kr
2013년 ESPOL, Engineering(학사)
2016년 ~ 현 재 경희대학교 전자정보융합공학과 석·박사통합과정
관심분야 : Deep Learning, Human Activity Recognition, Hand Gesture Recognition, Robotic Vision, Reinforcement Learning, Autonomous Object Manipulation.

김 태 성
https://orcid.org/0000-0001-7118-1708
e-mail : tskim@khu.ac.kr
1991년 Univ. of Southern California, Biomedical Engineering(학사)
1993년 Univ. of Southern California, Biomedical Engineering(석사)
1998년 Univ. of Southern California, Electrical Engineering(석사)
1999년 Univ. of Southern California, Biomedical Engineering(박사)
2013년 ~ 현 재 경희대학교 생체의공학과 및 전자정보융합공학과 교수
관심분야 : 기계학습, 패턴인식, 인공지능, 뇌공학