원격조작 시연으로 학습한 시각기반 관절 경로 예측과 Kinova Gen3 Lite 조작

이경호, 김성연, 이민서, 조경훈
인천대학교 정보통신공학과
e-mail : {kno022603, 202201569, minseo6384, ckh0923}@inu.ac.kr

Vision-Conditioned Joint Trajectory Learning from Teleoperation for Kinova Gen3 Lite Manipulation

Kyoungho Lee, Seongyeon Kim, Minseo Lee, Kyunghoon Cho
Incheon National University

Abstract

We propose a vision-conditioned manipulation method that directly predicts joint-space targets from a single RGB image. Teleoperation with an HTC Vive on a Kinova Gen3 Lite provides demonstrations; gripper close/open events yield grasp and place joint states (q_g, q_p). A ResNet-based model regresses (q_g, q_p) using only MSE loss. At inference, an Intel RealSense D435i supplies RGB (depth unused); the predicted joints are streamed under simple velocity/acceleration limits, avoiding the image→EE pose→IK pipeline and its discontinuities. We evaluate a two-color pick-and-place: the system grasps the selected block and places it in the matching cup. Real-robot results show stable execution and consistent color-to-target matching with RGB-only input. Contributions are: (i) a minimal teleop-to-joint labeling pipeline, (ii) a simple vision-to-joint mapping without IK, and (iii) real-world execution with lightweight constraints.

* 이경호, 김성연은 공동 제1저자로서 이 논문에 동일하게 기여하였음.

I. 서론

로봇 조작 과제는 물체 형태·배치 변화, 가림, 비정형 접촉 등으로 불확실성이 크고, 전통적 파이프라인(탐지→포즈추정→경로계획→IK)은 단계별 오차가 누적되어 실행 불안정·계획 실패로 이어지기 쉽다. 본 연구는 이러한 병목을 줄이기 위해 단일 RGB 영상에 조건화된 관절공간 경로를 직접 예측하는 방식을 채택한다[1]. 즉, 시작 상태에서 잡기 관절 상태를 거쳐 놓기 관절 상태로 이어지는 연속 관절 시퀀스를 곧바로 산출하여 IK의 다해·특이점 비연속성 문제와 EE 포즈 오차의 비선형 증폭을 피한다.

시각 조건부 정책의 유효성은 대규모 실제 데이터·모델 연구에서도 확인되고 있다(QT-Opt[2], RT-1[3], Transporter Networks[4], Diffusion Policy[5] 등). 우리는 이 흐름을 관절공간 직접 매핑으로 더 단순화한다. 특히 원격조작(teleoperation)은 사람이 장면을 보고 선택한 움직임을 감독 신호로 제공하므로, 모델이 "이 장면이면 이런 관절 움직임"이라는 시각-행동 대응을 학습하게 한다. 본 연구에서는 원격조작 기반 시연으로 데이터를 수집하되, 추론 시에는 오직 영상만으로 관절 경로가 결정되므로 "시각

기반" 접근이라 볼 수 있다.

종합하면, 본 연구의 필요성은 (1) 단계적 파이프라인의 오차 누적·IK 비연속성을 회피해 실행 안정성을 높이고, (2) 영상 조건부 관절 시퀀스로 장면 변화에 대한 적응력을 확보하며, (3) 원격조작 시연으로 현실조건에서 데이터를 효율적으로 수집하는 데 있다. 이를 위해 우리는 ResNet 기반 표현에서 관절 경로(시작–잡기–놓기)를 직접 회귀하는 간결한 모델, 그리퍼 이벤트 기반 자동 라벨·동기화 절차, 그리고 속도·가속도 제약을 반영한 시간화 후처리로 Kinova Gen3 Lite에서 즉시 실행 가능한 파이프라인을 제시한다.

II. 본론

2.1 텔레오퍼레이션으로 데이터 수집

본 연구의 데이터는 HTC-Vive 컨트롤러를 이용한 원격조작(teleoperation)으로 수집하였다. 사용자는 컨트롤러를 손에 쥐고 실제 물체를 집어 지정된 위치에 놓는 일련의 동작을 수행하며, 이때 영상 프레임(RGB), 컨트롤러 포즈, 로봇 관절 각도, 그리퍼 상태(open/close)가 공통 시간 기준(ROS time)으로 동기화 되어 기록된다. 학습 표적이 관절 시퀀스이므로, 카메라–로봇 오차는 크게 중요하지 않다. 라벨링은 그리퍼 닫힘/열림 이벤트를 경계로 하여 잡기 관절 상태(q_g)와 놓기 관절 상태(q_p)를 자동 추출하고, 에피소드의 시작→q_g→ q_p로 이어지는 연속 관절 시퀀스를 시연 궤적으로 저장한다. 시나리오는 평탄한 작업대 위 다양한 색상·모양의 소형 물체를 대상으로 구성하였으며, 사용자가 대상 물체를 선택해 집기–이동–배치를 수행하도록 유도하였다. 데이터 로그에는 (i) 프레임 타임스탬프, (ii) 로봇 관절 벡터와 그리퍼 명령, (iii) 컨트롤러 포즈가 포함되며, 최종적으로 각 에피소드는 단일 RGB 이미지(또는 선택적으로 소수의 프레임)와 그에 대응하는 대표 관절 시퀀스(시작→ q_g → q_p), 그리고 그리퍼 트리거 시점으로 요약된다.

2.2 네트워크 학습

모델은 단일 RGB 이미지를 입력으로 받아 관절공간 대표 시퀀스 중 잡기 관절 상태와 놓기 관절 상태만을 직접 회귀한다. 백본으로 ResNet을 사용하였고, 다중 퍼셉트론 헤드가 각 관절 차원(DoF)에 대해 $[\hat{q}_g, \hat{q}_p] \in \mathbb{R}^{2\times DoF}$을 출력한다. 표적(target)은 원격조작 로그에서 추출한 절대 관절 상태$[\hat{q}_g, \hat{q}_p]$이며, 학습 안정성을 위해 각 DoF에 min–max(작업범위 기반) 스케일링을 적용해 네트워크 출력과 동일한 스케일에서 비교한다. 손실함수는 평균제곱오차(MSE) 단일 항으로 정의하며, 추가 정규화(연속성, 관절 한계 패널티 등)는 사용하지 않았다:

$$\mathcal{L}_{MSE} = \frac{1}{2DoF}\sum_{k\in\{g,p\}}\sum_{d=1}^{DoF}\left(\hat{q}_{k,d} - q_{k,d}\right)^2.$$

최적화는 Adam(초기 학습률 1×10^{-3}, weight decay 선택), 배치 크기 32, 에폭 100으로 수행했고, 학습 초기에는 커리큘럼으로 $\hat{q}_g$만 회귀한 뒤 $\hat{q}_p$를 추가하는 단계적 학습을 적용해 수렴을 돕고 최종적으로 두 상태를 동시 회귀한다. 추론시에는 속도 가속도 제약을 반영한 채로 예측된 $\hat{q}_g, \hat{q}_p$를 고정 제어 주기로 수행하였다.

Ⅲ. 실험 및 결과

본 연구의 실험은 Kinova Gen3 Lite 매니퓰레이터와 Intel RealSense D435i 카메라로 구성된 단일 팔 시스템에서 수행하였다(그림 1). 카메라는 고정된 위치에 장착하여 RGB 프레임만 사용하였고, 깊이(Depth) 정보는 활용하지 않았다. 작업대는 평탄한 테이블이며, 두 가지 색상의 블록(예: 녹색·베이지)과 그에 색상 매칭되는 컵 2개를 배치하였다. 임무는 "지정된(선택된) 블록을 집어 동일 색상의 컵 내부에 정확히 배치"하는 pick–and–place로 정의한다.

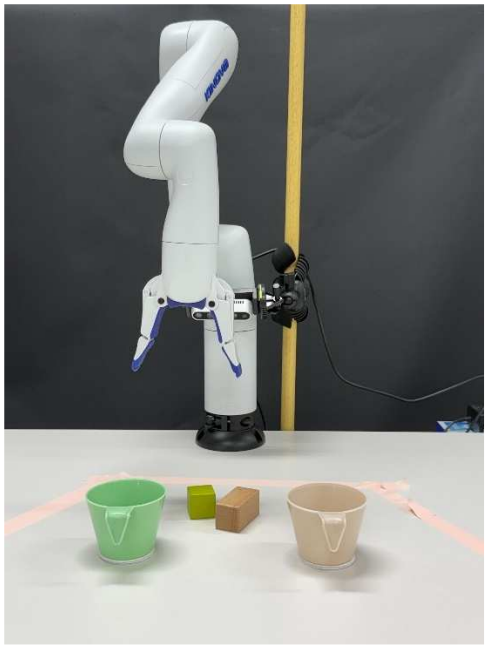

그림 1. 실험 환경 구성: Kinova Gen3 Lite,

RealSense D435i(RGB만 사용),
두 색상의 블록과 매칭 컵 2개.

그림 2는 HTC Vive 컨트롤러를 이용해 시연 데이터를 수집하는 과정을 보여준다. 수집 시 불필요한 손떨림과 무의미한 진동을 줄이기 위해 그리퍼의 방향을 항상 하향(테이블 법선 음의 방향)으로 유지하도록 제약을 두었으며, 이는 집기 전·후의 자세 변동을 최소화하고 라벨(q_g, q_p)의 분산을 줄이는 데 도움을 주었다. 사용자는 컨트롤러의 트래킹 포즈에서 유도된 이동 방향 정보로 로봇을 조작하고, 그리퍼 개폐(open/close)는 컨트롤러의 별도 버튼으로 명확히 트리거하였다. 모든 신호(영상 프레임, 컨트롤러 포즈, 관절 각도, 그리퍼 이벤트)는 공통 타임스탬프 하에 로깅되어, 이후 그리퍼 폐합/개방 순간을 기준으로 q_g, q_p를 자동 추출한다.

그림 2. 텔레오퍼레이션 기반 데이터 수집 장면. Vive 컨트롤러로 이동 방향을 입력하고, 별도 버튼으로 그리퍼 개폐를 트리거한다. 수집 과정에서는 그리퍼가 하향을 유지하도록 제약을 두어 시연의 일관성과 라벨 품질을 확보하였다.

로봇 제어는 관절 공간에서 이루어지며, 네트워크가 예측한 잡기 관절 상태 q_g와 놓기 관절 상태 q_p를 고정 제어 주기로 스트리밍하여 실행하였다. 별도의 최적 시간화는 수행하지 않았고, 속도·가속도 제약만 컨트롤러 단계에서 반영하였다.

그림 3은 실제 테스트한 에피소드의 시퀀스를 보여준다. 타깃 물체는 베이지 블록으로 설정하였다.
(a) 그리퍼가 블록 상면에 정렬되어 접근하고, q_g에서 그리퍼 폐합으로 집기를 수행한다.
(b) 집기 후, 네트워크가 예측한 q_p방향으로 매칭 색상의 컵(베이지 컵)으로 이동한다.
(c) 컵 상단 림 바로 위에서 정지하여 자세를 안정화한다(컨트롤러의 속도·가속도 제한 하에서 자연스러운 감속).
(d) q_p에 도달하면 그리퍼 개방으로 물체를 컵 내부에 안전하게 배치한다.

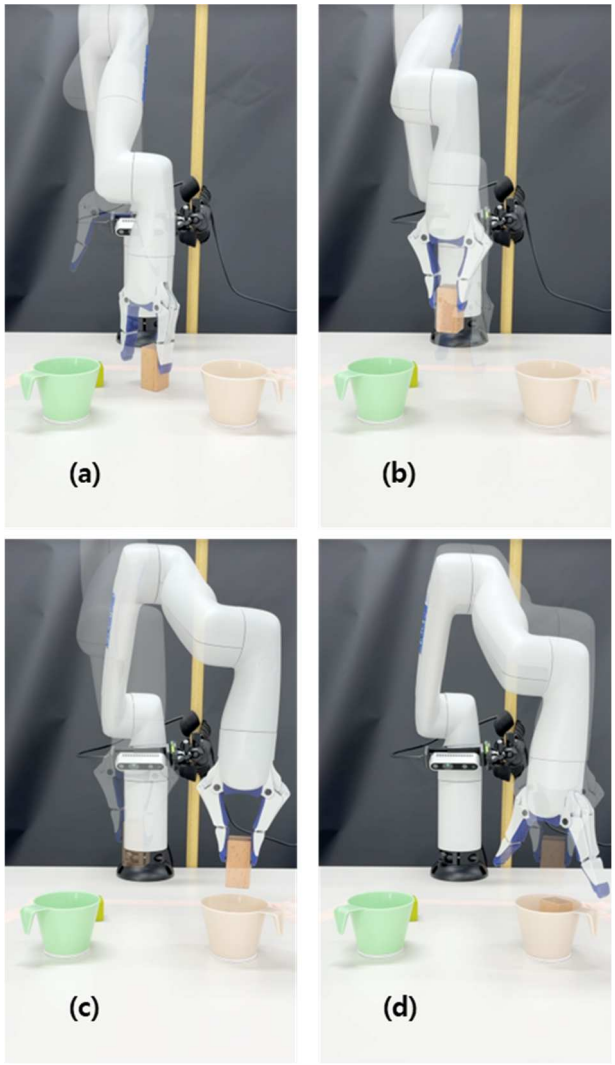

그림 3. 단일 RGB 기반 관절 경로 예측을 이용한 pick–and–place 시퀀스: (a) 베이지 블록 집기, (b) 베이지 컵 방향 이동, (c) 컵 림 상부 정지, (d) 컵 내부 배치.

동일 프로토콜을 녹색 블록–녹색 컵 조합에도 적용한 결과, 정성적으로 동일한 성공 동작을 확인하였다. 이는 제안된 네트워크가 단순한 구조임에도 불구하고, 단일 RGB 입력만으로 색상·배치 문맥에 조건화 된 관절 상태를 안정적으로 산출해 실용 가능한 pick–and–place 동작을 수행함을 시사한다. 실패 사례는 주로 컵 림 반사 및 부분 가림 상황에서 드물게 관찰되었으며, 이는 조명 증강 강화 또는 깊이 정보 결합(RGB–D)으로 개선될 여지가 있다.

Ⅳ. 결론 및 향후 연구 방향

본 연구는 단일 RGB 영상에 조건화된 관절공간 표적(q_g, q_p)을 직접 예측해 IK를 거치지 않고 Kinova Gen3 Lite에서 즉시 실행 가능한 pick-and-place를 구현했음을 보였다; 단순한 MSE 학습만으로도 색상 매칭 과제에서 정성적 안정성을 확인했으나, RGB 단일 입력·두 지점 표적·간이 속도·가속도 제약 등으로 복잡한 경유지 계획, 가림·반사 대응, 시간 최적성 측면의 한계가 남는다. 향후에는 (i) 다중 스텝 등 더 복잡한 임무로 확장하고, (ii) Diffusion Policy와 Transformer[6]계열 모델을 도입해 시퀀스·경로 분포·불확실도를 모델링하며, (iii) RGB-D/멀티뷰 등을 결합해 견고성을 높이는 방향으로 연구를 진행할 예정이다.

사사

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업의 지원을 받아 수행된 연구임(IITP-2025-RS-2023-00259678)

참고문헌

[1] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-End Training of Deep Visuomotor Policies," *J. Mach. Learn. Res.*, vol. 17, no. 39, pp. 1334–1373, 2016.

[2] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation," in *Proc. Conf. Robot Learn. (CoRL)*, PMLR, vol. 87, pp. 651–673, 2018.

[3] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, *et al.*, "RT-1: Robotics Transformer for Real-World Control at Scale," *arXiv:2212.06817*, 2022.

[4] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, and J. Lee, "Transporter Networks: Rearranging the Visual World for Robotic Manipulation," in *Proc. Conf. Robot Learn. (CoRL)*, PMLR, vol. 155, pp. 726–747, 2021.

[5] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," *Int. J. Rob. Res.*, early access, 2024.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.