Development of Visual and Tactile based Human Behavior Imitation Learning Platform

Jongbum Park, Seungwon Lee, Mingi Jung, Sunme Park, and Yujin Kim
Intelligent Robotics Research Center
Korea Electronics Technology Institute
Bucheon, South Korea
jbpark@keti.re.kr

Abstract— A novel approach to robot control, inspired by human behavior, has been introduced with the aim of effectively navigating complex and dynamically changing environments while handling a wide variety of objects. To achieve this objective, researchers have introduced the DexMV (Dexterous Manipulation from human Videos) platform [1], which harnesses human demonstration videos within a simulation environment to rapidly amass extensive datasets for learning purposes. Building upon this foundation, this study takes the DexMV platform a step further by enhancing its capabilities to incorporate tactile data obtained from human grasping actions. Moreover, significant efforts have been dedicated to bridging the gap between simulation and real-world implementation inherent in applying imitation learning outcomes to actual manipulation tasks. This paper presents a comprehensive overview of the evolution of the DexMVT (Dexterous Manipulation from human Video and Tactile sensing) platform.

Keywords— imitation learning, robot learning, human behavior imitation, tactile sensing, sim2real transfer

I. Introduction

Manufacturing tasks, such as assembling parts in industrial settings, involve repetitive work within a stable and unchanging environment. This characteristic facilitates the implementation of automation and robotics, as operational control remains relatively consistent across fixed locations. However, in the context of commercial service environments, constant changes are inherent within the surroundings, and the nature of the tasks varies. Consider the act of pouring water into a cup at a restaurant – each establishment possesses distinct environmental factors, diverse pot and cup shapes, and varying water quantities to be poured. Applying conventional manufacturing robot control methodologies to settings like restaurants necessitates reconfiguring control parameters for each unique instance, especially when pots or cups change. Given that service environments exhibit significantly greater complexity and diversity compared to manufacturing settings, this approach proves inefficient, inevitably leading to elevated robot operational costs. To address this challenge, we present an alternative approach that mimics human behavior for diverse service-related motion tasks.

II. Composition of Imitation Learning Platform

The proposed imitation learning platform, referred to as DexMVT (Dexterous Manipulation from human Video and Tactile sensing), encompasses a comprehensive set of components as illustrated in Figure 1. Building upon the foundation of the DexMV (Dexterous Manipulation from human Videos) platform [1], DexMVT integrates five key elements to facilitate effective imitation learning and enhance the manipulation capabilities of robots.

(1) Human Behavior Data Collection: The first component involves the compilation of human behavioral data. This dataset captures the diverse range of actions and motions performed by humans during manipulation tasks.

(2) 3D Pose Recognition: The second facet revolves around 3D pose recognition. This technology enables accurate identification and tracking of the spatial configurations and orientations of objects and manipulators.

(3) Robot Model Retargeting: The third aspect entails the process of robot model retargeting. Through this process, the learned human behaviors can be translated and adapted to the specific kinematics and dynamics of the robotic manipulator.

(4) Imitation Learning Training on Simulation: The fourth component focuses on the training phase. Imitation learning is carried out within a simulated environment. The robot learns to replicate human actions by observing the dataset and mapping them onto its own simulated actions.

(5) Actual Manipulator's Operation: Finally, the fifth component bridges the gap between simulation and real-world operation. The knowledge gained during simulation is applied to real manipulator tasks, enabling the robot to execute learned behaviors in a physical setting.

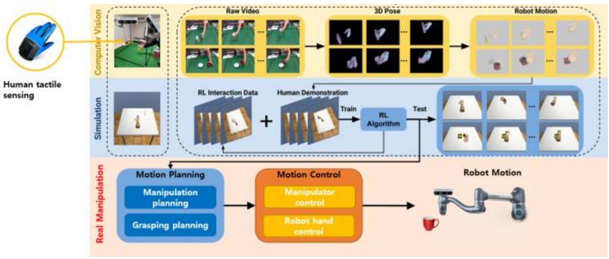

Fig. 1. DexMVT platform Overview

The DexMVT platform represents a substantial expansion of the original DexMV platform, encompassing the integration of human tactile information to facilitate the translation of acquired learning outcomes into tangible manipulation tasks. By combining tactile sensory data with visual inputs, DexMVT significantly augments the robot's capacity to comprehend and engage with its surroundings. Additionally, this platform streamlines the progression from

This work was supported by the Technology Innovation Program (20018112, Development of autonomous manipulation and gripping technology using imitation learning based on visual-tactile sensing) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

simulated learning experiences to practical manipulation activities, effectively showcasing the robot's acquired proficiencies within a functional operational setting.

A. *Collection of video-based human demonstrations*

We are working on enhancing the capabilities of commercial VR gloves by integrating tactile sensors capable of detecting pressure and shear forces. These sensors will complement the existing finger motion tracking capabilities of the gloves. Simultaneously, we are developing a progressive video-based data collection system to investigate 3D poses and the adaptation of robotic models. Our dataset, curated for imitation learning, encompasses hand posture information, including sequential finger joint angles, synchronized with specific time frames. This dataset is curated during various tasks such as object manipulation and relocation. Furthermore, the dataset includes 3D positional information of objects during task execution, in addition to hand postures.

To illustrate, we record raw videos using a camera while an individual engages in tasks involving object manipulation, such as picking up and moving a mug. By estimating hand postures and the 3D positional data of the mug within the video, we construct a dataset tailored for imitation learning. This approach is analogous to the DexMV[1] platform's methodology. The DexMV platform can be succinctly described as depicted in Figure 2. The configuration involves transferring human hand movements and object poses onto a physical engine simulator. This data is then utilized to facilitate the training of robotic hands to emulate human-like motions.

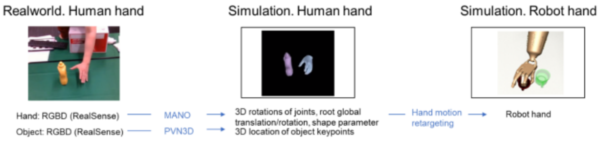

Fig. 2. DexMV Platform Overview

Fig. 3 illustrates the experimental setup used to gather human demonstration data. To bridge the gap between the robot's operational space and the real-world environment within the simulation, a camera acquisition system was devised. This system involved the installation of a framework that held two cameras in place. The enclosed space, measuring 60x60x30 cm^3, encompassed the area where objects were positioned.

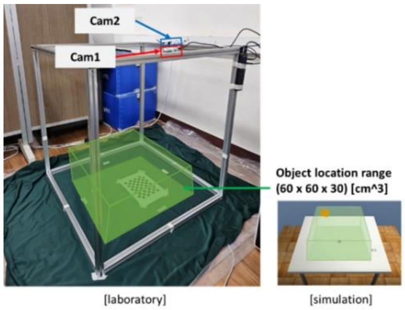

Fig. 3. Human Demo Collection Frame

B. *Pose Estimation*

The procedure for generating demo data is as follows.

(1) A task demo in which an expert directly manipulates an object (YCB object set) is taken in a cubic space (2 RealSense D435i: front, side view)

(2) The process of extracting the 3D hand pose and object pose from the captured video involves the following steps. To begin with object detection, the target object is chosen by recognizing the working environment using the weights acquired from training YOLOv5 with the YCB dataset. In order to estimate the 3D pose of the object, the DOPE++[2] algorithm is employed due to its superior performance when compared to two other algorithms: PVN3d[3] and FFB6d[4]. For the hand pose estimation, the Google Mediapipe hand estimation model[5] is utilized to extract data about the position of the wrist and the joints in the fingers.

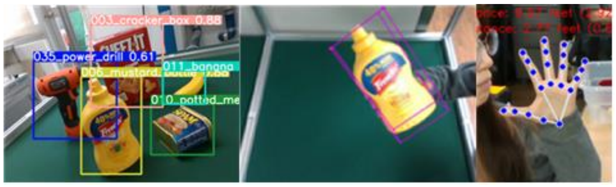

Fig. 4. Object detection(Left), Object pose estimation(Mid),. Hand pose estimation(Right)

By integrating the algorithms for object posture estimation and hand posture estimation, the conducted Object + Hand Posture Estimation test enables the achievement of pose estimation for both hands. This capability is crucial for progressing in imitation learning scenarios. Furthermore, the integrated approach allows for pose tracking even in cases of occlusion, ensuring the extraction of poses without hindrance. The simultaneous extraction of both object and hand poses is depicted in Figure 5 of the paper, illustrating the effectiveness of the proposed methodology.

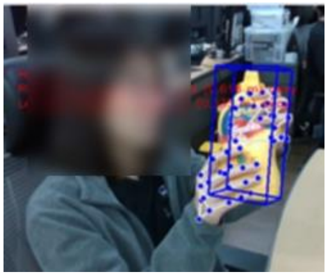

Fig. 5. Object + hand pose estimation

C. *Simulation environment modeling*

In the context of conducting imitation learning within a simulation environment, a critical process involves aligning 3D data from human hands with those of robot hands. This process, known as retargeting, necessitates the conversion of

3D data obtained from human hands and object videos. The objective is to establish a correspondence between the movements of human hands and those of robot hands. To achieve this, comprehensive models of manipulators, robot hands, and tactile sensors are crucial components.

In this study, the UR5e manipulator was chosen as the robotic arm, while the Allegro hand was selected as the robot's hand. Additionally, the tactile sensor used was the Digit sensor. The process involved creating detailed models of these components. For simulation purposes, the MuJoCo platform, based on a physics engine, was employed. This simulation environment allowed for the integration of the Digit sensor onto the tip of each finger, seamlessly incorporating it into the existing URDF model of the Allegro hand. A visual representation of this modeling process on MuJoCo, wherein the Digit sensor is attached to the Allegro hand, is illustrated in Figure 6.

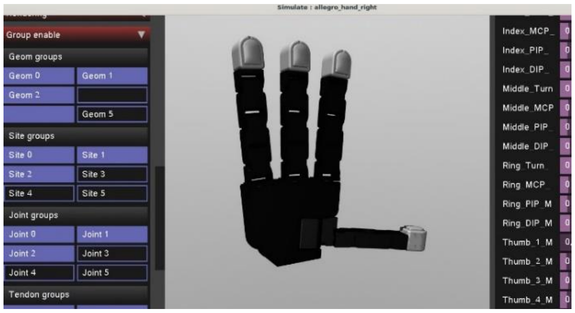

Fig. 6. MuJoCo Modeling combined with a Allegro hand and Digit sensors

An open-source simulator called TACTO was used to address the challenges of implementing tactile sensing within the MuJoCo physics engine. This simulator aimed to facilitate the simulation of a vision-based tactile sensor. The transition from the Pybullet physics engine to MuJoCo was executed to configure the environment for simulating the Digit sensor.

Furthermore, the simulator involved modeling the UR5e manipulator by incorporating information such as force specifications and the operational range for point control, as provided by the manufacturer. A visual representation in Figure 7 demonstrates the integration of the UR5e manipulator and an Allegro hand equipped with digit sensors.

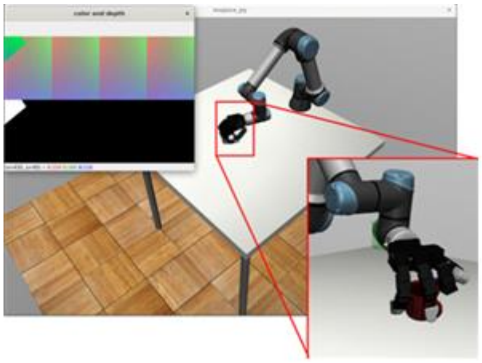

Fig. 7. Modeling of a Robot Hand with Tactile Sensors and a Manipulator

III. SIM2REAL TRANSFER

The imitation learning model developed by DexMV focuses on generating the pose trajectory of a robot hand using imitation techniques. Notably, this trajectory is generated independently of any specific dependencies on a robot manipulator model. However, directly commanding the robot controller with this pose trajectory could lead to issues such as trajectory deformation or abrupt halts due to the possibility of exceeding the maximum joint speeds of the robot.

This arises because the training of the learning network took place within a simulation environment that did not fully account for the intricate movements of an actual manipulator. To bridge this gap between simulation and reality, a solution has been proposed. A time scaling algorithm has been introduced, which applies a dampening effect to the pose trajectory derived from the learning network. This modification ensures that the resulting robot joint velocities remain within acceptable limits, preventing scenarios where the robot's movement might surpass its physical capabilities.

The algorithm's effectiveness can be observed in Figure 8, demonstrating how it mitigates potential issues arising from the disparity between simulated and real-world movements. By implementing this time scaling approach, the learned pose trajectory can be successfully transferred to practical scenarios, striking a balance between the simulated environment and the actual behavior of the robot manipulator..

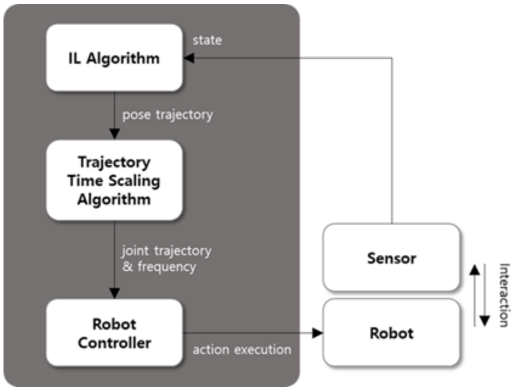

Fig. 8. Flowchart of trajectory time scaling algorithm

The equation (1) for obtaining the optimal input control period corresponding to the pose path of the learning network is as follows.[6]

$$\Delta t^{*} = \Delta t^{d} \times max\left(\frac{max(\left\|\dot{Q}_{i}^{d}(t)\right\|)}{\dot{q}_{allow,\,i}}\right) \qquad (1)$$

where,

Δt^{*} : optimal control period in the real environment

Δt^{d} : control period in the simulation environment

i : denotes the joint number of the n-dof robot

t : time step of the trajectory

$\dot{\boldsymbol{Q}}_i^d(t)$: i^{th} joint velocity corresponding to pose trajectory at time step t

$\dot{\boldsymbol{q}}_{allow,\, i}$: maximum allowable velocity of the i^{th} joint that given in advance

IV. Summary and Future Work

The DexMVT platform is constructed upon the foundation of DexMV [1], with the enhancement of incorporating tactile sensing to emulate human grasping actions. A structured approach was devised to gather exemplar data of human behaviors, facilitating the identification of hand movements and grasped objects, as well as the advancement of pose estimation techniques. Moreover, comprehensive models for manipulators, robotic hands, and tactile sensors were meticulously prepared to facilitate the translation of hand gestures—a crucial element in imitation-based learning.

As the development of the DexMVT platform advances, the research endeavors will persistently concentrate on refining tactile sensors capable of discerning pressure and shear forces in the context of human grasping actions. Furthermore, exploration into the fusion of visual and tactile sensory inputs will continue, along with the formulation of imitation learning algorithms rooted in multi-sensory information. Addressing the challenge of narrowing the gap between simulation and reality (sim2real) remains a focal point, encompassing endeavors in both real-world scenarios and virtual environments.

Acknowledgment

We are thankful to Professor Xiaolong Wang, the developer of DexMV, and Professor Nikolay A. Atanasov, the collaborating research lead at University of California San Diego, for their contributions.

References

[1] Qin, Y. et al. (2022). “DexMV: Imitation Learning for Dexterous Manipulation from Human Videos”. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13699. Springer, pp.570-587

[2] Jin M, Li J, Zhang L (2022) DOPE++: 6D pose estimation algorithm for weakly textured objects based on deep neural networks. PLoS ONE 17(6): e0269175. https://doi.org/10.1371/journal.pone.0269175

[3] Y. He, W. Sun, H. Huang, J. Liu, H. Fan and J. Sun, "PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 11629-11638.

[4] Y. He, H. Huang, H. Fan, Q. Chen and J. Sun, "FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 3002-3012.

[5] G. Amprimo, C. Ferraris, G. Masi, G. Pettiti and L. Priano, "GMH-D: Combining Google MediaPipe and RGB-Depth Cameras for Hand Motor Skills Remote Assessment," 2022 IEEE International Conference on Digital Health (ICDH), Barcelona, Spain, 2022, pp. 132-141

[6] Y. Kim, S. Lee, S. Park, and J. Park, “A method of Sim2Real transferase through trajectory time-scaling,” 2022 IEIE(The institute of Electronics and Information Engineers) syposium, Daejeon, Republic of Korea, 2022