Robotics: Science and Systems 2025 – Workshop on Continual Robot Learning from Humans

A Generalizable Gripper Interface for Hardware-Agnostic Policy Learning in Robotic Manipulation

Author Names Omitted for Anonymous Review.

***Abstract*—Leveraging on the recent advances on the Universal Manipulation Interface (UMI) solution [1] for cost-effective and easy human demonstration acquisition for robot learning, the paper at hand introduces a generalizable hand-held gripper implementation that broadens its usage to any type of robotic gripper. Proposed solution consists on a hardware implementation that directly captures gripper state together with an interface that ensures synchronization by embedding into video acquisition. Besides tackling the issues from the default vision-based approach for extracting gripper width, i.e. occlusions and additional computation load, this approach allows to consider different gripper operation modes such that any gripper configuration can be integrated, which has been exemplified in the paper for switch-actuated ones. For this purpose, a mechanical design has been also proposed to quickly change between different grippers adaptations using the standard UMI design. Solution performance results are presented together with its application on a manufacturing use case introducing switch-actuated grippers.**

***Index Terms*—Hand-held grippers, Learning from Demonstration**

I. Introduction

Process automation relies on endowing robots the ability to execute tasks reliably and with the same level of performance as skilled human operators. Leveraging on the recent advances in Artificial Intelligence (AI), Learning from Demonstration (LfD) aims at transferring operator performance to robot execution from a set of demonstration containing relevant data on the task at hand [6]. Classical approaches considered guiding the robot platform directly to perform these task demonstrations, but recent solutions have shifted towards cost-effective hand-held grippers to enable a natural acquisition process without any background in robotics. Moreover, not relying on a physical robot detaches demonstration from an specific platform such that data is valid for several with minimal setup, i.e. enables cross-embodiment. Among all the works in this direction, the Universal Manipulation Interface (UMI) [1] has arose as a promising device to integrate the simultaneous acquisition of different types of data that could lead to an improved policy training, e.g. tactile information [2]. But broadening its adoption calls for an effort on solving some of its issues that keep it from being reliable and cover a variety of manipulation scenarios. Some works focused on proposing an adaptation to substitute the default visual-based localization system, which required an initialization procedure that could ill-posed the acquisition of gripper trajectories and had poor performance in homogeneous scenes, e.g. in [7] through a marker-based approach using external self-calibrated cameras. More recently, FastUMI [3] eliminated the need for intrinsic mirror-based depth from UMI through the artificial generation of depth maps, and also specific adaptations on robot deployment for the specific parallel gripper model to maintain visual consistency through a mechanical contraption. Nevertheless, there are still two main caveats that still need to be addressed:

1) Only parallel grippers are considered in the literature since they can perform in a wide range of manipulation tasks. However, they might not be optimal or even suitable for many, specially on industrial settings where tasks generally involve the usage of tools, e.g. a screwdriver.
2) In all UMI variants, gripper width is generated through the detection of ArUco markers [5], which makes it prone to illumination changes or occlusions in manipulating some elements, e.g. cables, on top of increasing computation load.

This paper tackles both issues through the development of a interface upon FastUMI design to generalize over any type of robotic gripper, enabling a fully hardware-agnostic policy learning. Approach presents (i) an integrated solution to capture synchronized gripper positions through audio signals, which (ii) enables introducing any gripper configuration, supported by a mechanical adaptation to easily change between different adaptations. Figure II summarizes the main characteristics of proposed solution, which are detailed in Section II. Section III presents results on solution performance and its application on the dismantling of a desktop computer using a vacuum gripper and Section IV lays out main conclusions and future work.

II. Generalizable Gripper Interface for UMI

The core innovation on the proposed solution lies in replacing ArUco-based gripper width estimation with direct hand-held trigger position feedback, embedded as an additional analog signal into the demonstration data captured by the on-board GoPro. This approach takes inspiration from the one proposed in ManiWAV [4], where microphones are used as *tactile* sensing devices such that their feedback can be naturally synchronized with captured video. In this way we

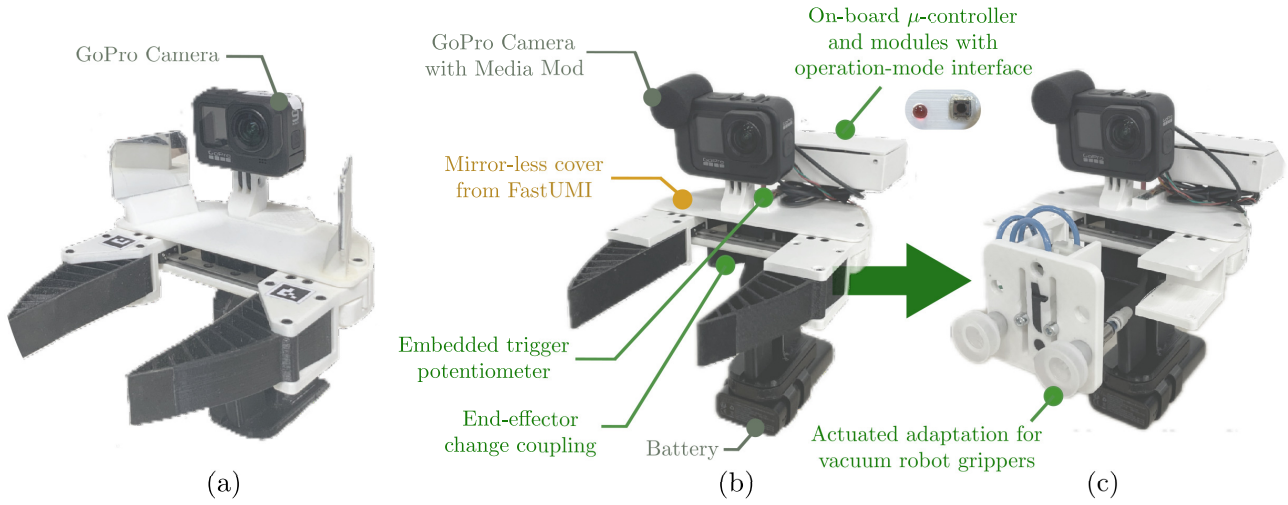

Figure 1. Comparison between original UMI device (a) and proposed adaptation (b), exemplifying quick end-effector change feature for a vacuum robotic gripper (c). All the contributions on the UMI device from this paper are highlighted in green.

achieve a low-latency and more reliable measure of gripper state that is robust to occlusions and lighting variability, and other factors such as wear and tear. Moreover, this approach allows to naturally extend beyond parallel grippers to a wide spectrum of end-effectors that might not present continuous behaviours. In this category lie those tools that operate in discrete on/off states, such as a vacuum or a magnetic grippers, namely switch-actuated, which have been considered in this paper to exemplify the flexibility of the generalizable interface.

A. Hardware Integration and Trigger Signal Conditioning

Gripper position feedback has been integrated using minimal hardware modifications and keeping the GoPro as the central data capturing hub. The mechanical transmission of the original tool remains a rack-pinion-rack assembly, ensuring symmetric and coordinated motion of the gripper jaws. Therefore, the position of the trigger has been bounded to a linear potentiometer which is connected to an analog-to-audio signal conditioning stage, such that it can be embedded as a sound-input into the GoPro, synchronized with captured video. The conversion takes the voltage proportional to trigger potentiometer position in real time using 12-bit ADC readings from an added on-board μ-controller, which then turns it into a 440 Hz sine-wave carrier whose amplitude maps the measured position, such that the conditioned signal is transmitted through a DAC into a standard 3.5 mm TRS jack connected to the on-board GoPro, using its dedicated multimedia housing.

However, enabling the trigger position signal capturing disables external sound input, which could serve other purposes within the demonstration acquisition as in [4]. Therefore, leveraging into stereo sound input characteristic, transmitted sound signal has been constructed as a dual-channel data bus: right one for captured environmental sound and left one for trigger position. This involves introducing an additional microphone to be also connected with the μ-controller, as shown in Fig. 2a. Hence, this solution preserves external audio information alongside gripper position signal, guaranteeing the precise synchronization of both with captured video.

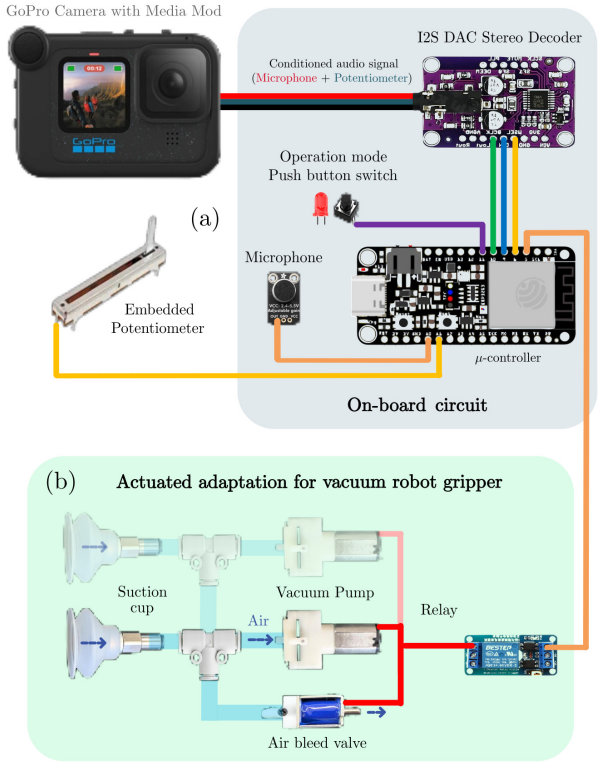

Figure 2. Connection diagrams for acquiring trigger position through embedded potentiometer and its processing, including environmental audio capture (a) and for the actuated adaptation for vacuum gripper (b). Power connections are omitted for clarity.

B. Adaptation for Switch-Actuated Gripper

To enable attaching a wide variety of end-effectors, we also developed a quick change coupling at the upper part of the trigger. This mechanism relies on a mechanical connection designed to firmly secure the gripper adaptation through a blocking mechanism activated through a side-to-side leverage. Hence, the only additional operation that needs to be performed on the original parallel actuation is to remove the

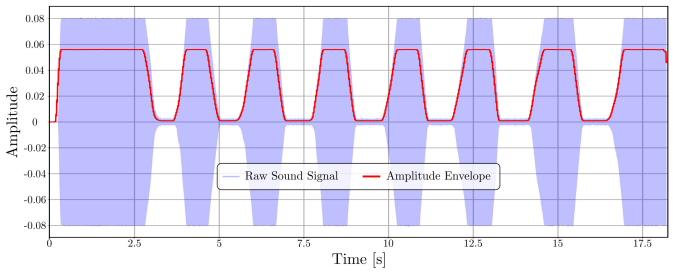

Figure 3. Sinusoidal audio signal containing trigger position from embedded potentiometer and corresponding amplitude envelope used to reconstruct it on demonstration post-processing.

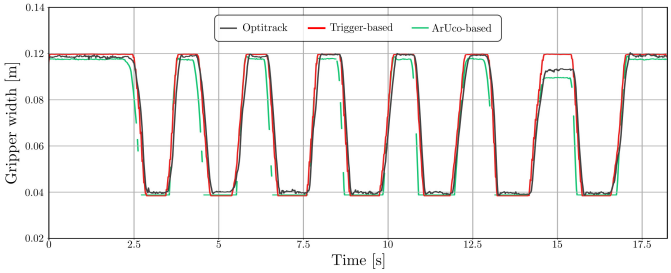

Figure 4. Comparison of gripper position acquisition through proposed embedded potentiometer and ArUco-based detection against Optitrack ground-truth. Gaps in the ArUco-based graph represent missing samples.

flexible fingers from the adapters. Note that this means that the rigid part will be still moving under trigger actuation, but its impact on the posterior policy training can be tackled through customary image masking on data preprocessing.

As aforementioned, in this paper we take as an example the switch-actuated grippers due to their popularity in manufacturing contexts, showing how the gripper position from trigger feedback mechanism can be easily adapted to this configuration. Once the specific gripper adaptation is mechanically coupled to the hand-held device, the actuation mean needs to be wired to the μ-controller. For switch-actuated grippers they can be emulated through a relay connected to the particular actuation mechanism, as it is shown in Fig. 2b for a vacuum gripper, where both a vacuum pump and a bleed valve need to be simultaneously activated. With regard to trigger position audio signal, using the same signal conditioning pipeline used for the parallel gripper, a simple user-defined thresholding is used to encode the gripper's state such that the amplitude renders on/off states. Additionally, aiming at having user-friendly device, a push button coupled with a LED has been added to the hand-held device, such that user can switch on-board between the continuous (LED off) or discrete (LED on) modes. This has been also depicted in Figs. 1 and 2a. By considering both operation modes under a common abstraction layer, proposed adaptations fully decouple gripper-specific hardware from the diffusion-policy learning pipeline improving flexibility and deployment efficiency.

C. Decoding and calibration

Before using the trigger position data on the UMI policy training pipeline, a signal decoding step needs to be performed. The amplitude from the sinusoidal wave is extracted using RMS, normalized with the corresponding range and stored alongside its respective timestamp, which can be seen in Fig. 3. Additionally, due to the sound sampling rate being higher (440 Hz) than the video sampling (60 Hz), an interpolation of the extracted amplitude values is needed over the timestamps of the video frames. The result is a single, uniformly formatted time series for gripper states that pairs every video frame with a corresponding actuator-state value. This synchronized, normalized dataset is what is used in the policy training process.

Dataset size	Post-processing Time [mm:ss]		Time reduction (%)
	ArUco-based	Trigger-based	
50	13:13	06:19	52.2%
100	26:52	11:55	55.7%
150	39:51	18:38	53.5%

Table I
POST-PROCESSING TIMES FOR ARUCO AND TRIGGER-BASED SAMPLES ON DIFFERENT DATASET SIZES. ALL EXPERIMENTS WERE CONDUCTED ON A MACHINE EQUIPPED WITH AN INTEL® CORE™ I7-10700K CPU RUNNING AT 3.80 GHZ, 31 GIB OF RAM, AND AN NVIDIA GEFORCE RTX 3090 GPU. THE OPERATING SYSTEM USED WAS UBUNTU 22.04.5 LTS.

III. PERFORMANCE AND USE CASE

First, the accuracy of proposed approach is compared to the default gripper position extraction from ArUco detection against the ground-truth given by an Optitrack system [1]. Obtained results, presented in Fig. 4, show that proposed method based on the trigger embedded potentiometer outperforms ArUco-based one, providing a continuous signal than even presents a faster response than Optitrack one. As expected, ArUco detection fails to provide some samples under fast movements and illumination changes. Although interpolation can solve this issue, as it can be seen on the results, this process induces a delay around 100 [ms] with respect to ground-truth, which might cut downs demonstration quality for training due to data desynchronization as UMI authors outline.

Detecting ArUco markers also introduces an additional computation load at post-processing that influences the scalability of the model training phase from acquired demonstrations. Table I summarizes the post-processing times for both approaches on different dataset sizes, which shows that a mean reduction of approx 53% is achieved using proposed approach.

Finally, to assess proposed solution on a real manufacturing scenario involving switch-grippers, the partial dismantling of a desktop computer has been used as a use case: first removing the side lid, which needs to be performed using an actuated adaptation for a vacuum gripper and then unplugging an internal cable and hard-drive by changing to the default parallel gripper configuration. The available video [2] shows the complete demonstration acquisition process and Fig. 5 depicts both phases together with the evolution of trigger position

[1]Natural Point, Optitrack: Motion Capture System - optitrack.com/
[2]Use Case video: shorturl.at/JaeFp

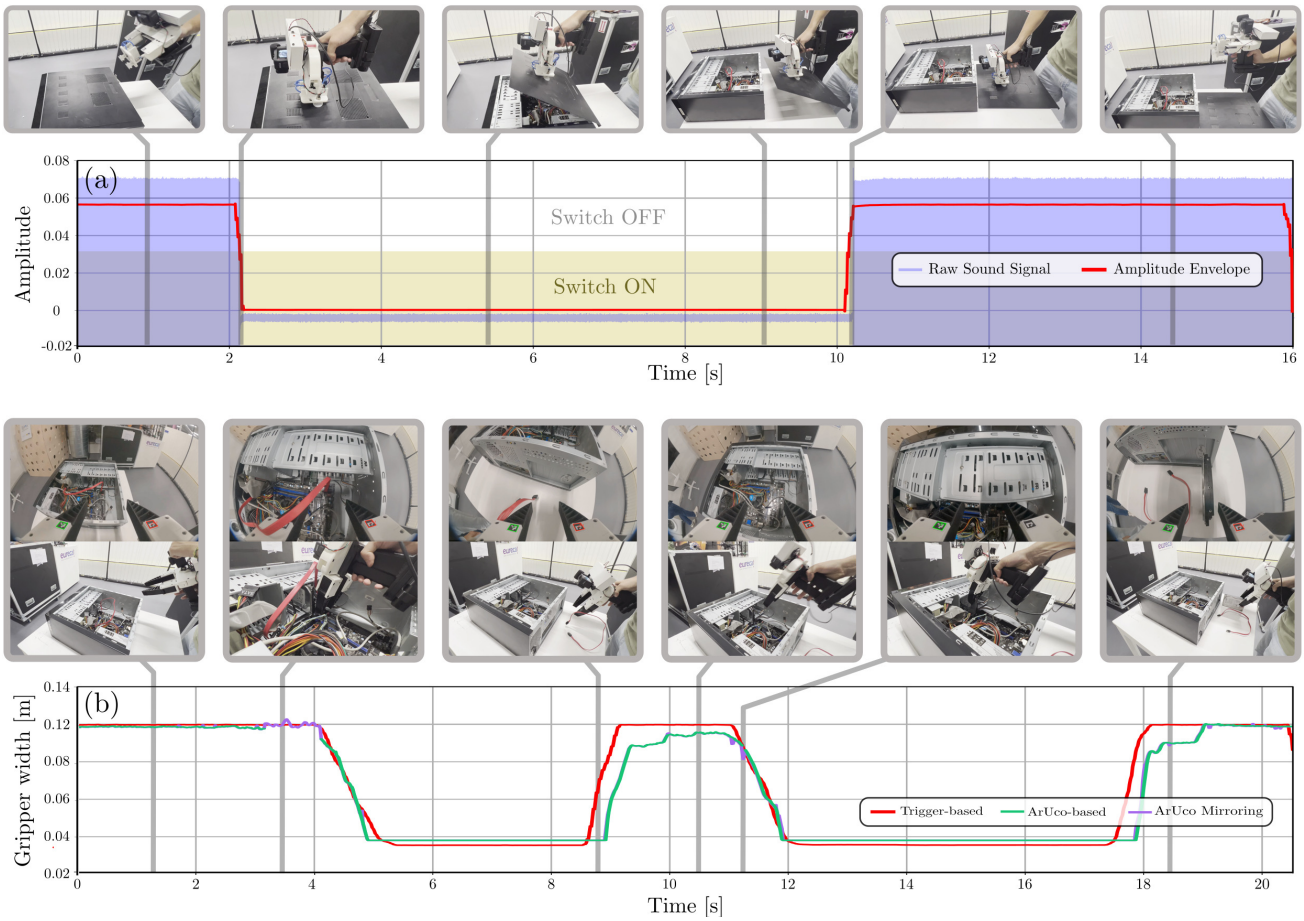

Figure 5. Desktop-computer dismantling use case using an actuated adaptation for vacuum grippers on the removal of the side lid (a) and using the parallel configuration for detaching an internal cable and hard-drive (b), showing in this last one proposed approach performance together with ArUco one and its mirroring approach.

signal and gripper widths for the set of available approaches, respectively. Besides the default ArUco-based process for gripper position, we have also included the approach presented in FastUMI [3] to tackle occlusions of one marker, based on mirroring the position of the detected one. On the first phase, it can be seen how the adaptation is able to successfully remove the lid, first by performing a side motion to unlock it and then lifting it up. On the second phase it can be seen how the ArUco-based approach does fail when one of the markers is occluded (second scene, around 3.5 [s]) or not detected (fifth scene, around 11[s]), and although the mirroring approach does overcomes these situations is also ill-posed in case both markers are not detected and suffers accuracy-loss and delay due to the interpolation process.

IV. Conclusions and Future Work

This paper proposes an upgrade of the UMI device to (i) eliminate the need of vision-based gripper position detection by embedding trigger position into an audio signal synchronized with captured video, which leads to (ii) extending its usage with any gripper configuration, that relies on a mechanical coupling and operation mode interface for a quick operation change. Presented results show that solution is reliable by design and outperforms defaults approach in terms of synchronization, and additionally cuts demonstration post-processing times to half. This is also demonstrated in the manufacturing use case presented, in which the application of the solution is exemplified by a vacuum gripper, that is required for a task that would otherwise be difficult to carry out using a parallel gripper.

Moving forward from the presented solution, our focus will be devoted to decouple the approach from the GoPro setup, leveraging on a ROS-centric integration that opens-up the integration of additional data, e.g. depth.

Acknowledgements

Acknowledgements omitted for Anonymous Review.

REFERENCES

[1] Chi, C., Xu, Z., Pan, C., Cousineau, E., Burchfiel, B., Feng, S., Tedrake, R., Song, S.: Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In: RSS (2024)

[2] Liu, F., Li, C., Qin, Y., Shaw, A., Xu, J., Abbeel, P., Chen, R.: Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. arXiv preprint arXiv:2504.06156 (2025)

[3] Liu, K., Guan, C., Jia, Z., Wu, Z., Liu, X., Wang, T., Liang, S., Chen, P., Zhang, P., Song, H., et al.: Fastumi: A scalable and hardware-independent universal manipulation interface with dataset. arXiv e-prints pp. arXiv–2409 (2024)

[4] Liu, Z., Chi, C., Cousineau, E., Kuppuswamy, N., Burchfiel, B., Song, S.: Maniwav: Learning robot manipulation from in-the-wild audio-visual data. arXiv preprint arXiv:2406.19464 (2024)

[5] Oščádal, P., et al.: Improved pose estimation of aruco tags using a novel 3d placement strategy. Sensors **20**(17) (2020). https://doi.org/10.3390/s20174825

[6] Ravichandar, H., Polydoros, A.S., Chernova, S., Billard, A.: Recent Advances in Robot Learning from Demonstration. Annual Review of Control, Robotics, and Autonomous Systems **3**(1), 297–330 (2020). https://doi.org/10.1146/annurev-control-100819-063206

[7] San-Miguel-Tello, A., Scarati, G., Hernández, A., Cavero-Vidal, M., Maroti, A., García, N.: Advances on affordable hardware platforms for human demonstration acquisition in agricultural applications. In: 2025 European Robotics Forum (ERF) (2025)