Research Statement – Ahad Jawaid

Title: Extracting Tactile Data from Hand-Object Videos for Autonomous Dexterous Manipulation

Introduction: Autonomously manipulating objects using robots has been a long-standing goal of the robotics community due to its broad applications, such as sorting packages, assembling products, and everyday robotic assistants. A key challenge in autonomous object manipulation is predicting high-precision movements applied by the robot's hand in real world settings. This challenge can be addressed by utilizing tactile (touch) sensing. Inspired by how humans use tactile feedback to adjust their hand movements when grasping an object, researchers have used tactile feedback to reduce the need of predicting very precise movements [1]. However, the ability of autonomous robotic hands to achieve good performance across a wide range of objects remains a challenge. A potential solution is to increase the data availability, as neural networks—on which these robots depend—require large datasets to generalize effectively. Unfortunately, large datasets with a wide range of objects have eluded the field due to the complexity and high cost of using tactile sensors on a large scale. Expert demonstrations, a common method for simplifying autonomous robot training, are hard to come by because of the time and expense involved in operating tactile-based teleoperation setups. Previous efforts have focused on utilizing limited data obtained from teleoperation setups more efficiently. Although these methods perform well, they often fail to generalize to other objects, leaving the issue unresolved [2]. Another approach involves learning in simulated environments and transferring the knowledge to the real world. But, a limitation of this approach is accurately modeling the real-world conditions, which limits the generalizability of such methods [3].

Objective: The objective of this proposal is to address the availability of large amounts of diverse tactile data to train tactile sensing agents that controls robotic hands to manipulate a diverse set of objects.

Relevant Literature: A recent study [1] showed the feasibility of using simple binary force sensors—which detect whether force is applied or not—for dexterous in-hand object rotation (e.g., balls, cubes, cylinders) around a predefined spatial axis, as shown in Figure 1 [1]. Another recent work [4] showed that it is possible to obtain hand joint angles from hand-object interaction videos of humans to train a vision-based agent. Together, these studies raise the question of whether it is possible to extract binary tactile expert data from videos of hand-object interactions.

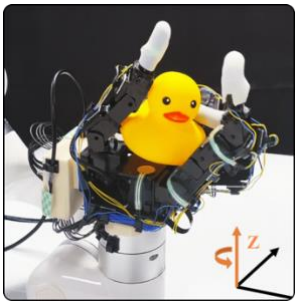

Figure 1: Robot in-hand rotation along Z axis [1]

Hypothesis: Based on the key insight, a hand's contact points with objects in video frames can be considered binary tactile feedback. I hypothesize that binary tactile data can be extracted from videos of hand-object interactions and used as expert demonstrations to train a neural network. This network would aim to perform at a level comparable to one trained with real tactile sensory data and demonstrations from a pretrained expert agent for in-hand object rotations around a predefined spatial axis using a single hand. If successful, this method could be extended to more general manipulation tasks, allow the community to be able to extract tactile expert demonstrations from hand-object interaction video, and allow for greater use of the tactile modality in robotics. If not, we can learn what makes tactile information distinct from visual modality in videos for learning to dexterously manipulate objects.

Aims / Method: This research aims to address the data availability problem for training a tactile-based agent. I will train an agent using behavioral cloning, where the agent learns to map the expert's observations to its actions. To this end, I will obtain expert data by extracting tactile observations and human hand joint angles from video frames. I will exclude the pinky in our expert cloning, as done in prior works because it introduces redundancy [5]. Due to this redundancy, robot hands typically only have four fingers, as will the one I will use. I will extract the hand joint pose and the hand's contact points from videos using an off-the-shelf MANO (Model of a human hand) estimator and contact-point predictor [6]. Then map the contact points to the tactile sensor locations to obtain the binary tactile data as show in Figure 2. The human hand joint angles will be retrieved from

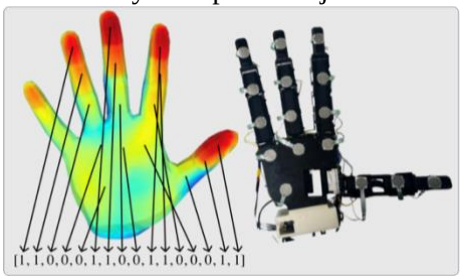

Figure 2: Contact points to tactile

the estimated MANO, which encodes pose information, including joint angles. The input to the agent will consist of the extracted tactile data and the joint angles of the human hand from the estimated pose. I will train the agent to predict the joint angles of the human hand in the next frame, as shown in Figure 3. The agent's output, the predicted joint angles, will be used to move the hand's joints to the desired positions. I will replace the human hand's joint angle and tactile inputs with the robot's during inference.

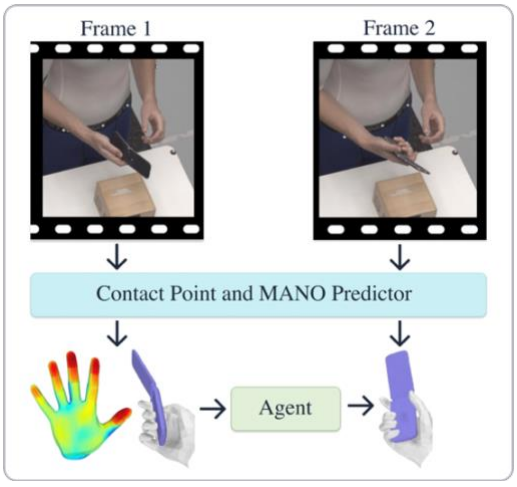

Figure 3: Behavioral Cloning Framework

Experiments: I will conduct experiments in simulation and in the physical world. I will use the Isaac Simulator by Nvidia for the simulation experiments since an existing environment from a previous work is available with a single-hand robot equipped with binary tactile sensors [1]. For the physical experiment, I will require a 16 Degree of Freedom Allegro hand, a robotic arm, 16 force-sensing resistors, and a microcontroller to collect the sensor's signals. I will use the same model architecture described by Yin et al., which employs a multilayer perceptron (MLP) and their open-source code [1]. The baselines will consist of an agent trained via behavioral cloning with expert demonstrations from a pretrained agent from Yin et al. [1]. As an ablation, I will train an agent using only joint angles as input, excluding tactile information to determine if tactile information improves the performance. Lastly, to study the quality of the extracted tactile data, I will compare the performance of the agents trained with contact points extracted from videos versus the ground truth contact points.

Data: I selected the ARCTIC dataset for the hand-object videos, which contains 2.1 million video frames of human bimanual (two-handed) object manipulation [6]. I selected this dataset for two main reasons: 1) the number of frames, since the generalization of neural networks depends on the amount of training data available, and 2) its inclusion of ground truth contact points, which will be needed for the ablation study. Since this dataset is bimanual, I will filter out the video frames that use both hands since I will only use one, which can be done by filtering out the frames where both hands' contact points are active.

Evaluation: I will use time-to-fall and cumulative rotation angles for evaluation metrics since they are standard metrics for in-hand rotation [1]. Time-to-fall tracks how long it takes for the object to fall out of the hand while it is being rotated, and it will be tracked in seconds. The cumulative rotation angle will be the total angle rotation made around the selected axis during each trial, which is a useful metric to track the rotational capability of the agent.

Intellectual Merit: This work addresses the availability of a large and diverse dataset to train tactile sensing agents. Current datasets using the tactile modality only have thousands of samples to train from, but as seen in other systems such as vision-based manipulators, increasing the diversity and amount of data to millions leads to better generalization and performance on manipulation tasks [4]. If my hypothesis is correct, this method could lead to a cheap and diverse data source for researchers in the tactile field. If the hypothesis is disproved, it will contribute to the community's understanding of the quality of tactile data necessary to train tactile feedback-based agents.

Broader Impacts: Using tactile feedback to manipulate objects in an error-robust way could reduce the engineering requirements of robotic hands, eliminating the need for highly precise, reliable movements. This approach could lead to more affordable robots making them accessible to younger students interested in STEM, providing cost-effective data, and enabling more efficient autonomous robots for daily tasks. I will mentor an undergraduate student on this project, giving them valuable research experience that could help strengthen the U.S. STEM workforce. To share the project results, I will publish and present my work at an international robotics conference. Additionally, the source code will be open sourced to facilitate replication and further development.

References: [1] Yin et al. RSS, 2023. [2] Guzey et al. CoRL, 2023. [3] Nagabandi et al. CoRL, 2020. [4] Bharadhwaj et al. ICRA Workshop, 2023. [5] Chen et al. 2024. [6] Fan et al. CVPR, 2023.