Leveraging Non-Prehensile Tactile Data for Object Retraction in Constrained Clutter Using Imitation Learning

Dane Brouwer, Joshua Citron, and Mark Cutkosky

***Abstract*— Retracting objects from dense collections of movable objects can be difficult for humans, let alone robots. A key sensing modality that assists humans during these tasks is the tactile sensing present on the non-prehensile surfaces of our hands and arms. We propose an investigation of the role of non-prehensile tactile sensing for training robots to retract objects in constrained clutter. We have built hardware and simulation environments that closely mimic each other and utilize custom triaxial tactile sensors. We use imitation learning to train policies both with and without tactile information on 259 demonstrations gathered in simulation and compare their performance. Preliminary results indicate that non-prehensile tactile information assists navigating to the target object despite object jamming. Implementing these learned policies on the hardware setup is ongoing work.**

I. Introduction

Robots typically struggle when reaching through dense, constrained clutter since visual occlusions and nonlinear contact phenomena make it difficult to predict how the scene will evolve.

Prior work has demonstrated that slender end-effectors equipped with suction cups at the tip are effective at retracting objects at the back of cluttered lateral-access scenes [1]. This approach, however, relies on single object interactions and requires iterative pushes and retraction, since vision and proprioception are the sole sensing modalities. In comparison, other recent work uses tactile sensing on a robot arm to reach toward target locations through dense scenes of movable obstacles, without vision [2].

Other work has highlighted the value of tactile, force-torque, and audio data to augment vision for contact-rich manipulation using imitation learning [3, 4]. These approaches, however, do not use non-prehensile tactile information—despite experiencing many non-prehensile contacts.

More generally, we anticipate that imitation learning is a promising candidate for contact-rich, non-prehensile tasks to (1) improve a strategy's use of features of tactile information and (2) provide an agnostic evaluation of how much the tactile information improves performance in comparison to other sensor modalities. In particular, we investigate whether non-prehensile tactile information is crucial for learning strategies to reach and retract objects in constrained clutter.

II. Methods

To investigate this hypothesis, we built a hardware setup consisting of a dense collection of movable obstacles with various physical and visual properties (Fig. 1a). We reach into these scenes using a 7DOF industrial robot arm, an external isometric camera, an eye-in-hand fisheye camera, and a slender paddle-like end-effector equipped with a suction cup. We also incorporate suction pressure sensing and non-prehensile triaxial tactile sensing, as presented in [5]. We gather demonstrations using a 3Dconnexion SpaceMouse® Compact to command the desired pose of the arm and the suction state.

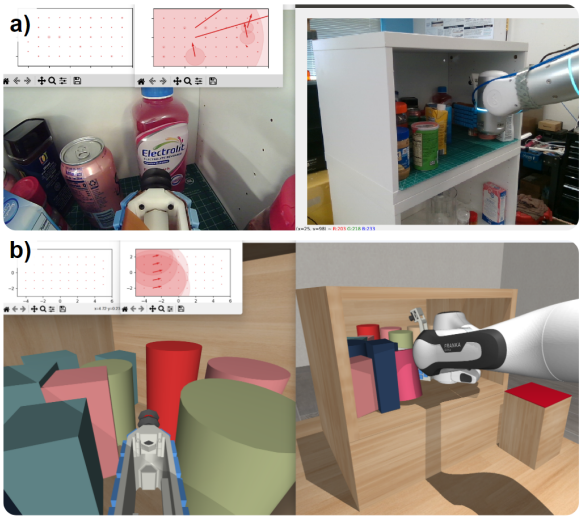

Fig. 1: Example constrained clutter scenes a) on hardware and b) in MuJoCo. The normal force of the tactile sensors is represented by the diameters in the top left inset and the shear forces correspond to the arrow direction and magnitude.

We have replicated the scenes in MuJoCo, as seen in Fig. 1b. As others have asserted, there is a potential to train in sim and have results carry over, low-shot, to hardware with a sufficiently accurate simulation [6].

We randomize obstacle locations throughout the simulated scene and define a task of retracting a red target object located at a random location at the back of the scene. We similarly gather demonstrations with the simulated robot and sensors. We use an adhesion actuator to mimic the suction cup behavior and only allow the adhesion force to turn on when the suction cup is aligned with the target object surface. We approximate the suction cup pressure sensing with a binary signal corresponding to this contact alignment.

For initial validation of the proposed investigation, we gathered 259 demonstrations in simulation on unique, randomly generated scenes. We then trained two policies for equivalent durations on the same demonstrations but with access to different subsets of data. We used a diffusion policy [7] imitation learning framework. The first policy we call

The authors are with Stanford University, USA {daneb, jcitron, cutkosky} @stanford.edu

vision only, which has access to only proprioceptive and eye-in-hand visual data. The second policy we call *full info*, which has access to proprioception, eye-in-hand visual data, non-prehensile tactile data, and the binary suction pressure approximation. To maintain the same data structure between the policies, the *vision only* policy contains zero matrices corresponding to the additional information in the *full info* policy. We evaluated the performance of the two trained policies for 5 attempts each on 10 unseen scenes.

III. Results

The process of training policies to effectively complete this task is ongoing work, but we have several interesting preliminary findings.

The task is difficult: Even an expert human demonstrator often requires several attempts to successfully generate an adequate demonstration. Especially when the target is a rectangular solid, aligning the end-effector to enable effective use of the suction cup can be challenging. This is complicated by the fact that impeding obstacles are densely distributed and may jam, constraining robot motion. The maximum duration of 90 s (chosen to make trials relatively short) also makes it difficult to retract and re-enter if the initial approach angle was not suitable to complete the task. We hypothesize that a lack of haptic feedback to the demonstrator contributes to the task difficulty. When training, it means that any reactions to tactile information will be learned implicitly by the policy rather than from explicit reaction strategies by the demonstrator.

Approaching the target: A preliminary evaluation indicates that the inclusion of non-prehensile tactile information can improve a policy's ability to approach target obstacles. This is shown in Fig. 2. The vision-only policy generally fails to get within one object diameter and makes little final progress toward the goal, which is in part due to object jamming. Policies with access to tactile information consistently do better.

Retracting the target: We also have a preliminary indication that adding information from the suction contact state improves the success of acquiring an object and being able to remove it from the scene. This is demonstrated through a 0% overall success rate for the *vision only* policy whereas the *full info* policy successfully completes the entire task 6% of the time.

IV. Discussion

As robots begin to work in homes and offices, robust interaction with unstructured environments becomes crucial. We hypothesize that non-prehensile tactile information is important to accomplish this goal and have built environments and experiments to investigate this hypothesis. Despite the substantial difficulty of the task—as witnessed by the difficulty in gathering successful demonstrations and the very low early success rates for trained policies—the impact of the tactile and suction pressure information shows an encouraging trend. These results indicate that an understanding of contacts, not only on the nominal grasping surfaces but anywhere that contacts may occur, is likely important for operating in cluttered and contact-rich environments.

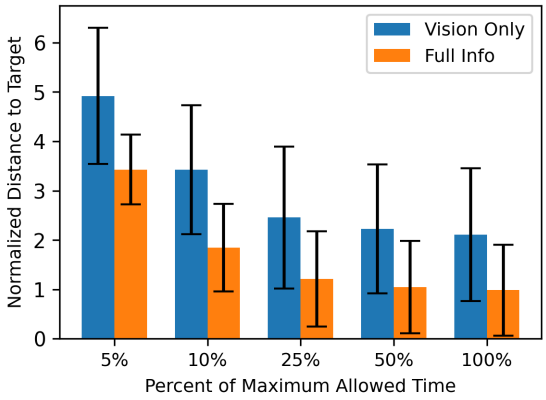

Fig. 2: Preliminary simulated results of distance to the target object (normalized by average object diameter) as a function of reaching time (normalized by maximum allowed duration, 90 s). Each distribution consists of 50 data points corresponding to the 5 attempts each on 10 unseen scenes.

It remains to be seen whether demonstrating the task in hardware will be harder or easier than in simulation. In particular, we are interested in the specific benefits from tactile data and suction data alone. To this end, we intend to conduct additional ablation studies to elucidate the relative merits. We are also interested in exploring how best to relay the robot's sensed tactile data, including contacts on the arm, to the human demonstrator via haptic feedback.

The preliminary results of this work raise several research questions: Will the earlier reported results concerning sim-to-real transfer hold for our task? Does force-torque information provide a comparable benefit to distributed non-prehensile tactile information? Is normal tactile sensing alone sufficient, or does shear information substantially improve performance? Does the quality of the demonstration data improve when (simplified) haptic feedback is displayed to the human demonstrator? Will task decomposition into phases (such as "reach", "acquire", and "retract") enable satisfactory success rates? We are addressing these questions in our ongoing work.

References

[1] H. Huang *et al.*, "Mechanical search on shelves using a novel "bluction" tool," in *IEEE ICRA*, 2022, pp. 6158–6164.

[2] D. Brouwer *et al.*, "Tactile-informed action primitives mitigate jamming in dense clutter," *IEEE ICRA (Accepted)*, 2024.

[3] H. Li *et al.*, "See, hear, and feel: Smart sensory fusion for robotic manipulation," *arXiv preprint:2212.03858*, 2022.

[4] M. A. Lee *et al.*, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," *IEEE T-RO*, 2020.

[5] H. Choi *et al.*, "Deep learning classification of touch gestures using distributed normal and shear force," in *IEEE IROS*, 2022.

[6] S. Höfer *et al.*, "Perspectives on sim2real transfer for robotics: A summary of the rss 2020 workshop," *arXiv preprint:2012.03806*, 2020.

[7] C. Chi *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," in *RSS*, 2023.