Opening Cabinets and Drawers in the Real World using a Commodity Mobile Manipulator

Arjun Gupta* Michelle Zhang* Rishik Sathua Saurabh Gupta
University of Illinois at Urbana-Champaign
https://arjung128.github.io/opening-cabinets-and-drawers

Fig. 1: This paper presents a system for opening cabinets and drawers in novel real world environments using a commodity mobile manipulator. Here we visualize an example execution of our system interacting with a novel object in an unseen environment. We include the following frames: before navigation, after navigation, pre-grasp pose, during manipulation, and at the end of manipulation. We evaluate our system across 13 unseen environments from 10 distinct buildings for a total of 31 unique articulated objects in the real world.

***Abstract*—Pulling open cabinets and drawers presents many difficult technical challenges in perception (inferring articulation parameters for objects from onboard sensors), planning (producing motion plans that conform to tight task constraints), and control (making and maintaining contact while applying forces on the environment). In this work, we build an end-to-end system that enables a commodity mobile manipulator (Stretch RE2) to pull open cabinets and drawers in diverse previously unseen real world environments. We conduct 4 days of real world testing of this system spanning 31 different objects from across 13 different real world environments. Our system achieves a success rate of 61% on opening novel cabinets and drawers in unseen environments *zero-shot*. An analysis of the failure modes suggests that errors in perception are the most significant challenge for our system. We will open source code and models for others to replicate and build upon our system.**

I. Introduction

This paper develops and evaluates a system for pulling open cabinets and drawers in diverse previously unseen real world environments (Figure 1). Opening articulated objects like cabinets and drawers presents hard technical challenges spanning perception, planning, and last centimeter control. These include accurate perception of object handles that are typically small and shiny, whole-body planning to drive the end-effector along the task constraint (*i.e.* trajectory dictated by the articulating handle), and dealing with execution errors in a task with low tolerances. All of these pieces have been studied at length in isolation [1, 5, 6]. Yet, how these modules interact with one another and what matters for successfully completing the task are not well understood. End-to-end learning via imitation or reinforcement circumvents these issues but is itself difficult because of the sample efficiency of learning and the unavailability of large-scale datasets for learning [2]. We take a modular approach and bring to bear state-of-the-art modules for perception and planning with a specific focus on studying how the different modules play with one another.

Specifically, for perception we extend a Mask RCNN model [3] to also output articulation parameters. For planning, we extend SeqIK, the recently proposed trajectory optimization framework [2] to produce whole body motion plans. Contrary to our expectation, just putting these two modules together did not lead to a successful system because of last centimeter errors in execution. Even slight inaccuracies in navigation and extrinsic camera calibration cause the end-effector to just be slightly off from the handle preventing handle grasping. To tackle this problem, we close the loop with proprioceptive feedback: predictions from visual sensors gets the end-effector in the vicinity of the handle and the actual grasping is done by leveraging contact sensors in the gripper and the arm.

Two other unique aspects of our study are a) the use of a *commodity mobile manipulator* and b) extensive testing in previously unseen diverse real world environments. Many previous papers have demonstrated specialized systems for similar problems [4]. Constructing specialized hardware for a given task can simplify the task at hand, at the cost of generality to other tasks. Therefore, we test our proposed system using the Stretch RE2, a *general purpose* commodity mobile manipulator, *without any hardware modifications*. Furthermore, this testing is conducted across 31 different articulated objects in 10 different buildings. Testing sites include offices, classrooms, apartments, office kitchenettes, and lounges. Our system achieves a 61% success rate in a zero-shot manner across this challenging testbed.

This broad study has allowed us to answer numerous

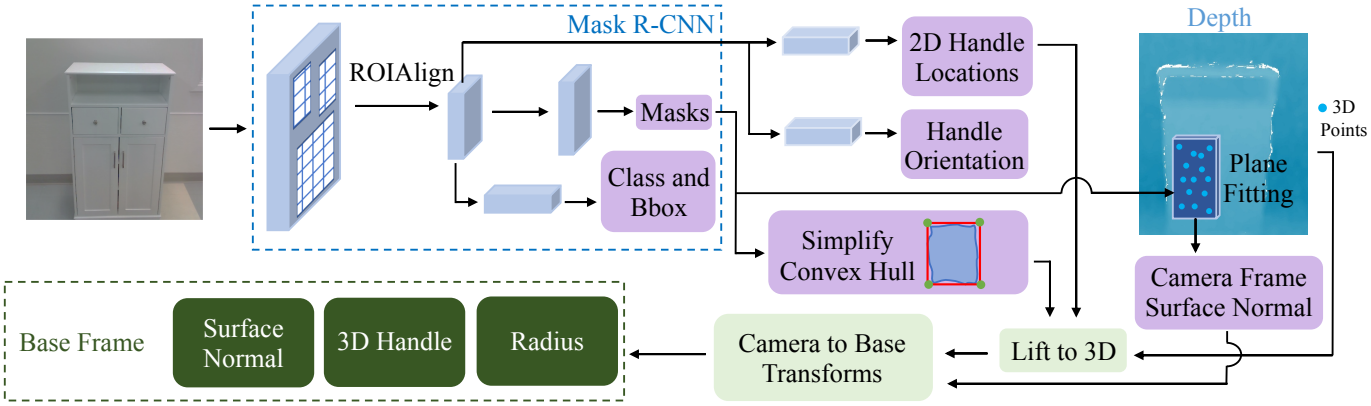

Fig. 2: Overview of the Perception Module. Given an RGB image our modified Mask RCNN detects articulated objects and predicts the articulation type, the handle orientation, the 2D segmentation mask, and the 2D handle keypoint. We fit a convex hull to the segmentation mask and simplify it to a quadrilateral. We fit a plane to the depth image points that lie inside the segmentation mask to estimate the surface normal. The 2D handle and quadrilateral corners are lifted to 3D using the depth image. All predictions are transformed to the robot base coordinate frame. The final output of the module includes the 3D handle coordinate and surface normal in the base coordinate frame for all articulated objects, and additionally the radius for cabinets.

questions about deploying such a system in the real world: a) what are the current bottlenecks in deploying a system for articulating objects in novel environments, b) how accurate should motion plans be for articulating objects, and c) what aspects of the current pipeline could benefit from machine learning? We find that perception is a major bottleneck for such a system, where inability to detect objects and handles accounts for 59% of the failures of the system. This calls for broad datasets with labels for cabinet and drawer articulation parameters in *diverse* settings. Our study also reveals that control is surprisingly robust to misestimations in the articulation parameters. Once the end-effector has acquired a firm grasp of the handle, the system is able to open the cabinet a non-trivial amount even with a radius error of up to $10cm$. Finally, errors from earlier parts of the pipeline compound to lead to failures in grasping of the handle.

II. System Design

Being able to open an arbitrary drawer or cabinet in a novel environment requires us to a) detect the object and predict its articulation parameters using on-board sensors, b) use the predicted articulation parameters to generate a whole-body motion plan, and c) adapt and execute the motion plan with the aid of proprioceptive feedback. We first present our approach for estimating articulation parameters given RGB-D images in Section II-A. We then discuss our methodology for generating a navigation target and a whole-body motion plan for opening the given object in Section II-B. Finally, in Section II-C, we describe how proprioceptive feedback from the robot is used to adapt the generated motion plan during execution.

A. Predicting Articulation Parameters using On-board RGB-D Sensors

Given an RGB-D image pair containing articulated objects, our goal is to a) detect cabinets and drawers and b) predict articulation parameters for the detected instances. These articulation parameters include the 3D handle location and surface normal for all objects, and additionally the axis of rotation and radius for cabinets. These articulation parameters help deduce the end-effector trajectory needed to open the given object. We predict 2D quantities from RGB images, and lift these predictions to 3D using the depth image.

For 2D prediction from RGB images, we adopt Mask RCNN [3]. As is, Mask RCNN predicts a 2D segmentation mask and the class of each detected object (in our case, the articulation type: drawer, left-hinged cabinet, or right-hinged cabinet). We add additional heads to Mask RCNN to.predict a) the 2D coordinate of the handle, and b) the handle orientation (horizontal or vertical).

We use the depth image to lift these 2D predictions to 3D. For the surface normal, we fit a plane to the depth image points within the predicted segmentation mask. For cabinets, we also need the radius and the axis of rotation. We compute the convex hull of the predicted 2D segmentation mask, and simplify it to a quadrilateral. We lift the vertices of this quadrilateral to 3D using the depth image and infer the rotation axis from the corners, *e.g.* for a left-hinged cabinet we use the left-most two points to define the axis of rotation. We use the distance of the handle to its projected point on the axis of rotation as the radius. Figure 2 shows an overview of the perception module. We train our modified Mask RCNN on the ArtObjSim dataset [2].

B. Motion Plan Generation

Our next goal is to generate a motion plan to open the given articulated object in a collision-free manner. We build upon past work from Gupta *et al.* [2] that assumes ground truth articulation parameters and tackles the problem of converting end-effector pose trajectories into robot joint angle trajectories. Specifically, rather than casting it as a constrained motion planning problem, Gupta *et al.* propose SeqIK, which casts it as a trajectory optimization problem.

We extend the framework presented in Gupta *et al.* [2] in three ways. First, while [2] works with the Franka Emika

Fig. 3: Topdown Navigation Target Locations. We visualize the topdown navigation target locations relative to the handle for each articulation type. We use the MPAO (No neural network) method from [2] to extract these in a data driven manner.

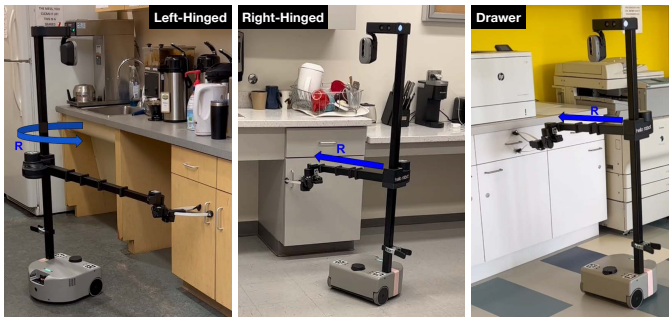

Fig. 4: Corrective Motions. We visualize the corrective motions for the different articulation types. For left-hinged cabinets, this is a counter-clockwise rotation in 1° increments. For the other objects, we extend the arm in $1cm$ increments.

Panda robot, we adopt their implementation to work with the Stretch RE2 robot, which has fewer degrees of freedom.[1] Secondly, we work with predicted articulation parameters as opposed to ground truth articulation parameters. We use our predicted articulation parameters from Section II-A and convert them to an end-effector pose trajectory in the same manner. Finally, rather than finding motion plans just for the arm assuming a fixed base position and orientation as in [2], we obtain *whole-body* motion plans for interacting with a given articulated object using our predicted trajectory. We find this essential to fully opening a wide variety of cabinets and drawers due to the limited number of degrees of freedom of the Stretch RE2.

SeqIK requires an initial base position. For the initial base position, we utilize MPAO (No neural network), a data-driven method from [2]. Figure 3 shows the base positions found by this procedure for drawers, left-hinged and right-hinged cabinets. We use these as the navigation targets for each articulation type respectively.

C. Adapting and Executing Motion Plans using Proprioceptive Feedback

Minor errors in state estimation, navigation and calibration compound to prevent handle grasping. In particular, when approaching either a horizontal or vertical handle with a wide gripper, there is sufficient tolerance in both the horizontal and vertical directions, but a much smaller tolerance in the depth direction. We develop a method for contact-based correction of the pre-grasp pose to combat this: we extend the gripper towards the object until contact is detected. For drawers and right-hinged objects, because the arm is largely parallel to the surface normal of the object, we keep extending the arm in $1cm$ increments until contact is made. For left-hinged objects, because the arm is largely perpendicular to the surface normal of the object, we rotate the base by 1° counter-clockwise until contact is made. See Figure 4 for a visualization of these contact-based correction primitives.

D. Pipeline

Here, we describe the full end-to-end pipeline. The robot begins approximately $1.5m$ away from the target articulated object. The base is arbitrarily oriented, as long as the desired object is within the field-of-view.

Our perception module, as described in Section II-A, produces a prediction for the handle location (in the camera's coordinate frame), the surface normal, and the radius (if the target object is a cabinet). We use a calibrated robot URDF to transform the 3D predictions (handle, surface normal, axis of rotation, and the navigation target) from the camera frame to the base frame. After this, we generate a whole-body motion plan using the methodology described in Section II-B. We execute the first qpose (full robot configuration), and subsequently run our contact-based correction mechanism from Section II-C. Once the handle has been grasped, we execute the rest of the motion plan.

III. Experiments

We work with the Stretch RE2 robot for our experiments. We present our full end-to-end system test results, in which our system is evaluated across 10 buildings and a total of 31 novel articulated objects. This test set of objects does not overlap with the set used for development. We assume the robot previously navigated toward the target object, so for each test it is positioned approximately $1.5m$ from the object with the camera oriented to have the target in view. We allow for some variance across tests in the exact positioning and orientation of the robot base due to environmental constraints and potential variance in the ending pose of any previous navigation. In particular, the base orientation is randomly chosen to be facing forward, oriented slightly to the left, or oriented slightly to the right. We represent each trajectory by ten end-effector waypoints, for which our whole-body motion planner attempts to find joint angles. We define a successful opening of an object if our system is able to execute at least 7 out of 10 waypoints. For cabinets, this corresponds to opening the cabinet over 60-degrees.

End-to-end System Test. Overall, our system achieves a 61% success rate across 31 unseen cabinets and drawers in unseen real world environments. For example deployments of our full pipeline in the testing environments, please refer to Figure 5.

[1]The Franka Emika has $\{3, 7, 1\}$ degrees of freedom while the Stretch RE2 has $\{2, 5, 1\}$ degrees of freedom for the base, arm, and gripper, respectively.

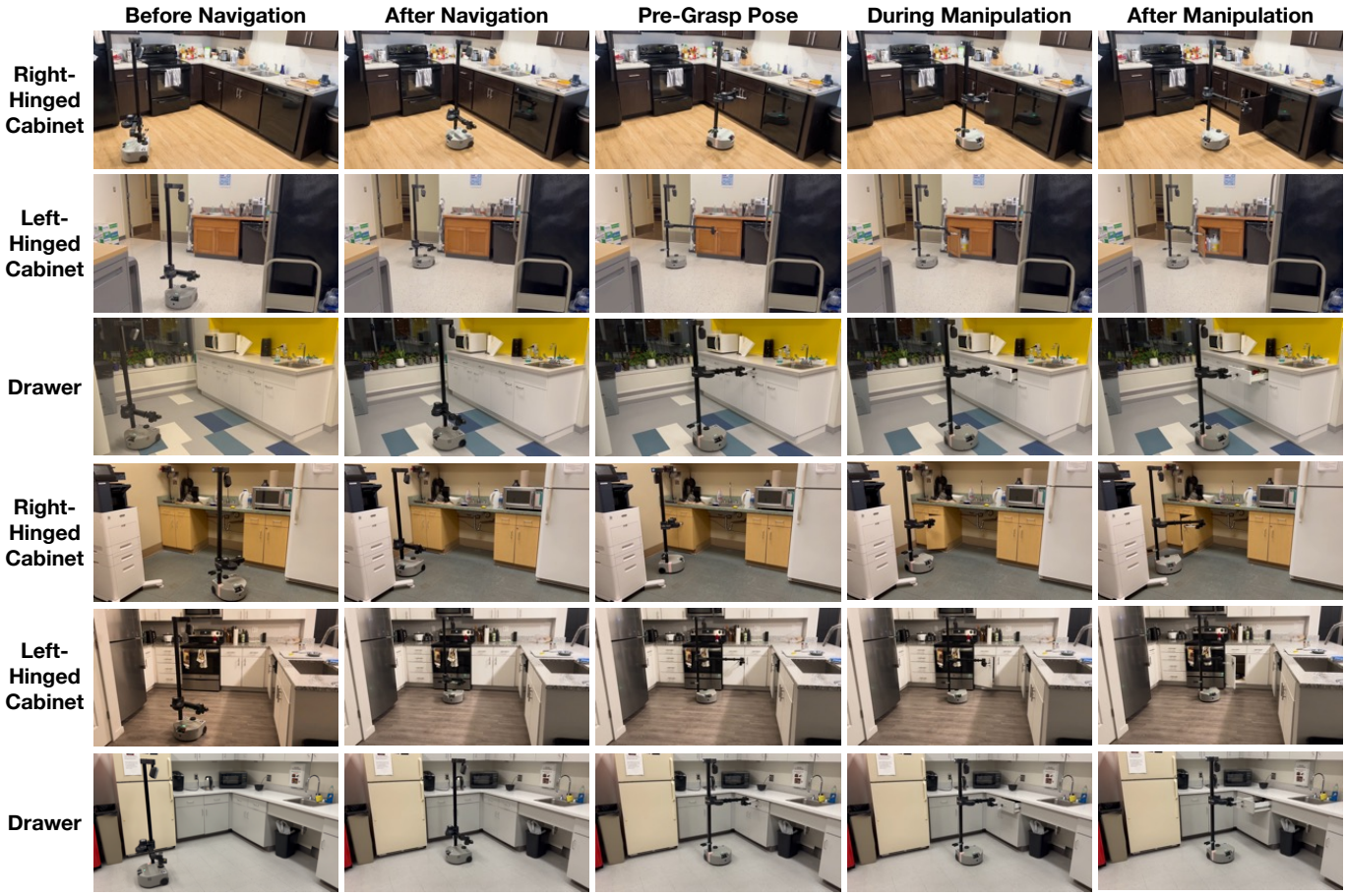

Fig. 5: Example roll outs of our full system in various unseen environments. For each environment, we show the following frames: before navigation, after navigation, pre-grasp pose, during manipulation, and at the end of manipulation.

IV. Discussion

In this work, we develop an end-to-end system for opening cabinets and drawers in novel real world environments. While our system is able to solve a majority of the novel objects we tested on, our large scale evaluation revealed unforeseen failure modes. This included failures in perception, navigation, and execution, which we describe next.

Failure in Perception. One of the main failure modes we encounter is error in perception. This failure in perception includes failure to detect the target object and erroneous handle orientation prediction by our Mask RCNN model. These perception errors are due in part to testing on out of distribution objects. The adapted Mask RCNN model is trained on luxury homes from the HM3D dataset, whereas we mainly test on the more readily available academic office buildings and apartments on campus.

Failure in Navigation. Our real world system was developed in an environment with tiled floors, so tests on carpeted floors introduced an unanticipated failure mode. When navigating on carpeted floors, the robot audibly strains during base rotations. This affects both the initial navigation and the deployment of the pre-grasp robot configuration, both of which involve base rotation, ultimately leading to a failed grasp of the handle.

Failure in Execution. In some cases, a firm, centered grasp of the handle would *not* be acquired. This would be due to imperfect calibration of the robot leading an error in lifting the 2D predictions to 3D, or minor navigation errors (even on tiled floors), which would compound. In such cases, as the cabinet was pulled open, the gripper would eventually let go of the handle. The vast majority of the cabinets we tested on had recoil, due to which the cabinet would close shut after the gripper let go of the handle, even after as many as 5/10 waypoints of the motion plan had been executed.

In summary, this paper presents the design and evaluation of a mobile manipulation system to open cabinets and drawers using a commodity mobile manipulator. Large scale testing across 13 test sites in 10 buildings and 31 different cabinets and drawers reveals guidance for practitioners aiming to build similar systems.

References

[1] Dmitry Berenson, Siddhartha Srinivasa, and James Kuffner. Task space regions: A framework for pose-constrained manipulation planning. *IJRR*, 30(12):1435–1460, 2011.

[2] Arjun Gupta, Max Shepherd, and Saurabh Gupta. Predicting motion plans for articulating everyday objects.

In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[4] Advait Jain and Charles C Kemp. Pulling open doors and drawers: Coordinating an omni-directional base and a compliant arm with equilibrium point control. In *2010 IEEE International Conference on Robotics and Automation*, pages 1807–1814. IEEE, 2010.

[5] Yiannis Karayiannidis, Christian Smith, Francisco Eli Vina Barrientos, Petter Ögren, and Danica Kragic. An adaptive control approach for opening doors and drawers under uncertainties. *IEEE Transactions on Robotics*, 32 (1):161–175, 2016.

[6] Xiaohao Sun, Hanxiao Jiang, Manolis Savva, and Angel Xuan Chang. Opdmulti: Openable part detection for multiple objects, 2023.