Special Issue: Robotics Collaborative Technology Alliance (RCTA) Program

Field Report

Grasping and Transport of Unstructured Collections of Massive Objects

Joseph Bowkett, Sisir Karumanchi and Renaud Detry
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109

Abstract: We present a collection of sensorimotor models which, when paired with a custom, mobile manipulation platform, collectively enable the autonomous deconstruction of piles of debris to facilitate mobility in cluttered spaces. The models address the problems of visual debris segmentation, object selection, and visual grasp planning. They also exercise proprioceptive grasp control, and force-controlled object extraction to enact the grasping plan. The object selector segments a debris pile into a set of parts that appear disjoint from one another and executes a rule-based decision program to select the object that appears easiest to extract. Then a grasp planner identifies a grasping point and hand preshape at a location where the gripper configuration fits the shape of the object, while also satisfying kinematic and collision-avoidance constraints. Our geometric grasp-prototype concept allows the planner to establish grasp suitability by fitting a set of shape-grasp primitives to a 2.5D depth image of the pile. The robot then applies the grasp reactively, pushing against the object until a preset resistance is met. Finally, an admittance controller guides object extraction, allowing a prescribed end-effector compliance along task-frame axes to minimize adverse forces and torques on the hand. We show experiments demonstrating the applicability of those models in isolation and in concert, including grasp tests conducted on objects representative of of a human-scale urban environment.

Keywords: mobile manipulation, grasping

1. Introduction

The field of robotic manipulation has grown and progressed impressively in the last decade (Kleeberger et al., 2020; Smith et al., 2012; Mason, 2018). However, research in robot manipulation is largely siloed. Challenging problems such as grasp planning, tactile control, or motion planning (to name a few) received vigorous attention and experienced multiple breakthroughs. By contrast, complex tasks that combine one or more of those domains are less often studied, and naïvely combining off-the-self packages to create a complex behavior often yields a brittle system. Understanding and solving the hard problems that stand between our robots and complex, multi-step manipulation tasks is one of today's key challenges in the field.

This paper offers a glimpse into the realization of a robotic system that addresses a multi-step manipulation task without supervision. The overall task consists of selecting an extractable object

Received: 23 October 2020; revised: 7 July 2021; accepted: 31 August 2021; published: 28 March 2022.

Correspondence: Joseph Bowkett, Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, Email: bowkett@jpl.nasa.gov

from a pile, removing it, and releasing it at a pre-set location. In the rest of the text, we will refer to this task as *T1*. We explain the design process that led us to decompose T1 into semi-isolated modules, and we describe the function and performance of each. We expose the numerous difficulties that one must face when integrating a set of components to form an end-to-end behavior, and we quantify the relationship between the performance of each component in isolation and in concert.

While this paper focuses on the extraction of objects from a pile (T1), it is embedded in a larger project that aims to solve the problem of indoor-outdoor mobility in cluttered spaces, deconstruction of a pile of debris, and recovery of an object of interest, possibly stored in a container. We will refer to this broader task as *T0*. T0 encompasses T1, and it also includes the problems of navigating through tight spaces, identifying debris piles, locating and opening a container and recovering its contents, and manipulating certain known objects according to a predefined policy. While the goal of the present paper is to discuss the design and performance of a robotic solution to T1, certain design choices represent a trade-off between constraints pertaining to T0. Those will be noted in the text, and the reader will be redirected to a second publication (Kessens et al., 2021) that provides a summarized discussion of all aspects of T0 jointly.

T1 (and T0) target an open-ended environment. We assume that a pile of debris blocks the robot's access to a target location, e.g., a hallway or road. In the latter case, the ground may consist of flat concrete or uneven dirt. The pile also spans the entire passage and may comprise items made of plastic (pipes, benches, traffic cones), metal (chairs, scrap, trusses), wood (boards, pallets), or concrete (cinder blocks). Note that, contrary to other efforts in this research doimain, we do not limit our scope to light objects or foam copies, and aim to handle objects that weigh up to 25kg total, as afforded by the RoboSimian derived arms and end-effector (Burkhardt et al., 2018). Our objective involves a tradeoff discussed more fully in Section 2. Objects are piled on top of one another in an arbitrary fashion. This scenario contrasts with more traditional, open-ended, industrial settings, where a larger set of constraints typically apply, for instance on the classes that objects may belong to, their maximum weight, texture and appearance, lighting conditions, tabletop/conveyor-belt location priors, etc. Our objective is to design a complete hardware and software robot stack that enables the removal of debris elements, weighing up to 8 kg each, within a multi-item environment.

The contributions of this paper are as follows:

- We present a novel robotic hardware/software system that is capable of autonomously deconstructing piles of debris, to facilitate mobility in cluttered spaces.
- We detail the design of the system's planning and control components.
- We explain our approach to calibrating the robot. Calibration is a known problem that offers a range of solutions, and selecting a solution that affords the required accuracy at a reasonable cost remains non-obvious, which explains why our approach is detailed here.
- We demonstrate the applicability of our work on grasp-region selection and grasp synthesis of large objects within a pile.

This paper is largely of the *systems* flavor. Our work provides an account of the challenges that field manipulation continues to face. Fielding a manipulation system remains a difficult endeavor, as shown by the mitigated performance of the system discussed below. Today, a substantial amount of engineering is required to bring a system to higher success rates, an effort that many applications cannot afford. This paper reveals areas where principled improvements may help raise success rates and reduce the need for problem-specific engineering.

2. System Design

Our goal is to design a system that can achieve mobility on uneven terrain, and deconstruct a pile of the nature of those shown in Figure 1. The objects that form the pile may be convex or concave,

Figure 1. Example manipulation scenes cluttered with objects of varied masses and geometric properties. Left: A pile of several aluminum truss segments, odd pieces of wood, a traffic cone, and a bucket. Right: An aluminum truss segment, a length of 4x4, and an overturned safety barrier upon a wooden pallet.

stiff or flexible, and exhibit complex topologies that may include void spaces between substructures, for example a ladder or truss. There is no restriction to objects passing through one another (a pipe can pass through a ladder), such that interlocking between objects may result in a lifting action on one being subject to the mass of multiple others. However, as discussed below, we limit our work to cases where objects can be pulled out along a vector aligned with the direction of the road. Situations where objects interlock and must be untangled to be extracted are beyond the scope of this work. The relevant objects within the manipulation domain are termed "massive" as they are of sufficient mass that lifting one or multiple (when interlocked) in certain orientations may exceed either the actuation force limits of the system, or cause mechanical damage to the gripper. This necessitates consideration of pile mass while planning grasps, and reactive control to detect when strength limits risk being exceeded.

A noteworthy aspect of the problem we study is its holistic consideration of the platform to conduct trades. Our main objective being to deconstruct a pile of debris, a driving aspect of our design is that neither grasping nor mobility need to be perfect, but the mobility system has to be capable of overcoming smaller objects that the manipulator is not dexterous enough to extract. We opted for a manipulation system composed of a 3-finger Robotiq gripper, a 4-finger high-torque "claw" gripper, and an Intel Realsense camera. The camera works best with objects larger than 1cm that are within its capability to resolve and manipulate objects larger than 10cm. In turn, we implemented mobility with a Talon tracked platform, that is able to overcome objects smaller than 10cm, as well as stairs and uneven ground. The range of the potential object space each of these modalities accommodates is visualized in Figure 3.

A second driving aspect of our design resides in our intention to develop a platform that is capable of operating on realistic objects, that are heavier than most objects typically considered in research tasks. Concurrently, T0 dictates that the platform must be able to pass through doorways and navigate indoors, which imposes a hard limit on the platform's operating volume. Those two requirements disqualify many industrial arms that are either designed for payloads lighter than 10kg, or are too large and heavy to fit on a mobility platform that can pass through doors. Instead, we opted for a custom arm design: a 7-DoF arm composed of actuators that are able to effect a torque of up to 500Nm. The actuators are custom-made and based on those that were designed by JPL for their Robosimian DARPA entry (Karumanchi et al., 2018). Figure 2 shows the complete system, with a JPL 7-DoF arm affixed to either side of the torso of the "RoMan" platform, with a 4-finger "claw" and 3-finger Robotiq gripper as the left and right end effector respectively.

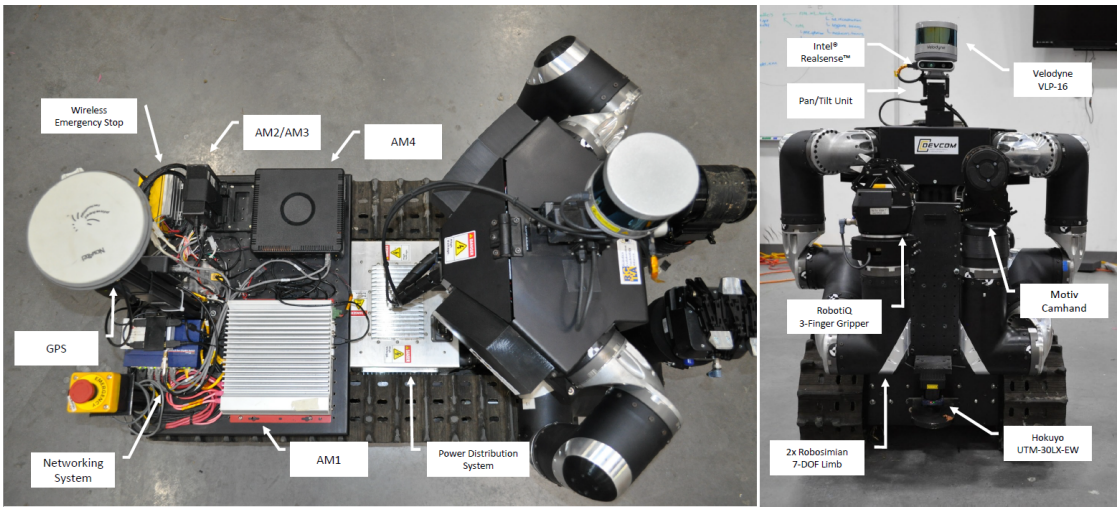

Figure 2. The "RoMan" platform upon which the algorithms herein were developed and demonstrated (Kessens et al., 2020). ©SPIE 2020.

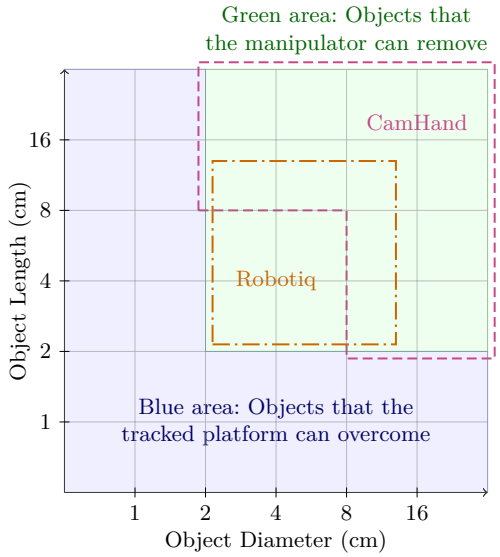

Figure 3. Representation of debris strategy as a function of debris size, assuming objects that exhibit a rotational degree of symmetry, such as tubes or cubes, or objects that exhibit a grasping fixture that has an approximate rotational degree of symmetry. The Robotiq gripper is capable of removing objects that present a grasping fixture whose diameter is comprised between 2cm and 15cm. The CamHand applies to objects with a diameter between 8cm and 18cm. Object whose diameter or length is smaller than 2cm would be more difficult to grasp, but those can be simply driven over.

3. Prior Art

Grasp selection in unstructured environments has proven a challenging task, and is complicated further when lacking *a priori* knowledge of manipuland shape and its mass properties.

Much of the earliest research on manipulation focused on grasp synthesis; the problem of finding a single or set of hand configurations that suitably constrain a target object relative to the agent, subject to any manipulation task specific constraints. Any advanced interaction with an environment first requires suitable contact to be made, for which reason, robust grasping is a pivotal concern for

autonomous robotics, and has therefore seen many different approaches developed. While this work addresses the selection of object candidates for grasping, the generation of grasps on previously unseen objects experiences similar challenges in the need to predict resultant wrenches from what is frequently incomplete geometric information. Means of grasp synthesis have historically been divided into two camps; that of explicit analytical representations, and more recent data-driven or empirical approaches (Bohg et al., 2014). One key example of geometry variation tolerant grasping comes from Eppner and Brock (Eppner and Brock, 2013) where shape adaptation is employed to allow a wider variety of geometric primitives to be applied to a previously unseen object set, though is focused on manipulands of low mass compared to those considered in this work.

Analytic approaches typically focus upon representation of the ability of a given end effector to both restrain, and manipulate, objects relative to a wrist or palm (Bicchi and Kumar, 2000). Perhaps the most important concept within deterministic analysis is that of grasp force or wrench closure, where contact forces can counteract all other external forces such as gravity (Nguyen, 1988; Kumar and Waldron, 1988). A stronger condition is that of form closure, described by Trickle as existing if and only if it is force closed with frictionless contacts (Trinkle, 1992).

Up until the turn of the millennium, the majority of grasping research focused on robotic grasping centered around model-based and mechanics-based approaches (Bohg et al., 2014), at which point a shift towards empirical or data-driven methods occurred. This may have been in part due to the rapid progression of computational power available to research labs, as well as the emergence of the simulation platform Graspit! (Miller and Allen, 2004). Empirical approaches initially used classical metrics derived from analytical formulations such as the ϵ-metric proposed by Ferrari and Canny (Ferrari and Canny, 1992).

While many of the above works have addressed a variety of means of grasp synthesis, largely agnostic to considerations beyond force or form closure (be it analytical or empirical), attention has more recently turned toward selection of grasp *regions*, with suitability to particular tasks or affordances. An affordance may be informally considered as an 'opportunity for action', as were first proposed by psychologist J.J. Gibson in 1966 (Gibson, 1966), and are used to describe the actions an agent may take with a given object or environment. They have been employed in the study of robotic traversal and object avoidance, grasping, and object manipulation (Horton et al., 2012). One of the first to combine higher level reasoning with lower-level geometric planning were Antanas et al., who proposed using symbolic object parts to semantically reason about pre-grasp configurations for particular tasks (Antanas et al., 2014). Detry, Papon, and Matthies then presented a model that computes a distribution of geometrically compatible 6D grasp poses from a depth map, and then applies a CNN-based semantic model to select those configurationally compatible with a given task (Detry et al., 2017). The selection of grasp type and location during a *handover* task were also recently investigated by Cini et al., who had human subjects pass and receive a range of objects (Cini et al., 2019).

Much of the study of manipulation affordances has emphasized grasp selection respective of the end-goal of composite or collaborative tasks, such as handing over an object to another agent, or pouring from a container (Detry et al., 2017). In unconstrained environments, where the mass of potential manipulands may vary greatly, the affordance of lifting may arguably be considered more important, as inability to lift an object typically precludes any other action. Some early works applied pure proprioception to identify the center of mass (COM) of an object in-hand (Atkeson et al., 1985), but there has been little use of exteroceptive sensing means to predict and inform this process. One example is found from Kanoulas et al., who address the question of wrench minimizing grasp selection on a single object by exteroceptively predicting the COM and then iteratively lifting and updating the estimate with wrist torque measurements (Kanoulas et al., 2018).

Unconstrained manipulation environments are commonly cluttered, with many potentially graspable or confounding objects present within a scene. Much focus has been placed on this problem within the context of the 'Amazon picking challenge', which sought to advance logistics technology capable of sorting through heterogeneous boxes of consumer items. One of the key takeaways from early iterations of the challenge was the importance of combining reactive control with deliberative

planning (Correll et al., 2018). Development by various teams led to the demonstration of the ability to recognize and grasp both known and novel objects in cluttered environments (Zeng et al., 2019), though these operate with manipulands well within the grasping system's capability.

Of perhaps closest relevance to this work is that of Boularias et al. (Boularias et al., 2015), who approach the problem of grasping objects in dense clutter with no prior information through the application of reinforcement learning. The robot learns online how to manipulate objects through trial and error, in particular through the application of *pre-grasping* actions that seek to expose objects for easier geometric access to suitable grasps, without giving consideration to the mass of candidate manipulands. Additional recent work on deep learning (Mahler et al., 2017a) and deep reinforcement learning (Zeng et al., 2018; Kalashnikov et al., 2018) provides capabilities that are similar in spirit and demonstrate a substantial capacity to encode fine visuomotor relationships. By contrast, the formulation presented in this chapter does not require any learning through repeated application, but seeks to leverage geometric context that can be provided by vision algorithms, even in the absence of any other prior information, and apply a model of contact physics to predict viable grasps, without extensive interaction with the environment. We recognize that this approach may be outperformed by deep-RL solutions on delicate tasks. However, given the coarse nature of debris removal, we opted to trade some representational capacity in the grasp planner for the ability to reconfigure it manually in the field, which, in our case, can easily be done by manually creating an additional prototype (Section 6.1).

Recent work by Holladay et al. looked at representing the kinematic and wrench constraints of the use of different tools to enable *forceful manipulation* (Holladay et al., 2019). While this leverages the concept of a wrench space surface to describe the forces that must be applied to a given tool to operate it, the analysis assumes exact prior knowledge of the manipulands and their interaction with the environment, which cannot be assumed working in unstructured environments.

Zhang and Trinkle propose to use a particle filter to simultaneously estimate the physical parameters of an object and track it while it is being pushed. The dynamic model of the object is formulated as a mixed nonlinear complementarity problem (Zhang and Trinkle, 2012). While they address uncertainty in the physical properties of a manipuland, they depend primarily on tactile information over the course of a motion to infer these properties, and are only concerned with the in-hand manipulation of a single object, rather than selection of viable grasps among a set of candidate manipulands as is the focus of the work in this chapter.

Berenson et al. investigated planning articulated arm motions in the presence of wrench constraint manifolds in the arm configuration space, such as might be imposed when lifting massive objects between two end effector poses (Berenson et al., 2009). This is distinct from the work in this chapter in that it focuses on wrenches imposed on the end effector by the mass of a manipuland during large scale motion, rather than predicting the restraining wrench imposed on a manipuland by its surroundings, but is complementary to the work presented here in that it could inform arm trajectories once an object is extracted.

The problem of identifying object properties in-hand through properceptive inference was investigated by Burkhardt et al. (Burkhardt et al., 2018), who also employed a Gaussian Process Implicit Surface (GPIS) in their representation. They focused on localising the center of mass through changing the orientation of a grasped object with respect to gravity, through wrench measurements from the wrist mounted force-torque sensor, while also inferring the geometry of the object through probing. In contrast to the work presented here, the GPIS was applied to the geometric representation of the object's surface, rather than in describing the wrenches restraining an object. They also rely solely on proprioceptive information, rather than attempting to leverage any information gleaned from visual sensing modalities as in this chapter.

4. Hand-Eye Calibration

Of paramount importance in the planning of autonomous interaction with any environment is concordance between the perceived location of objects of interest, and the positioning of the self-state

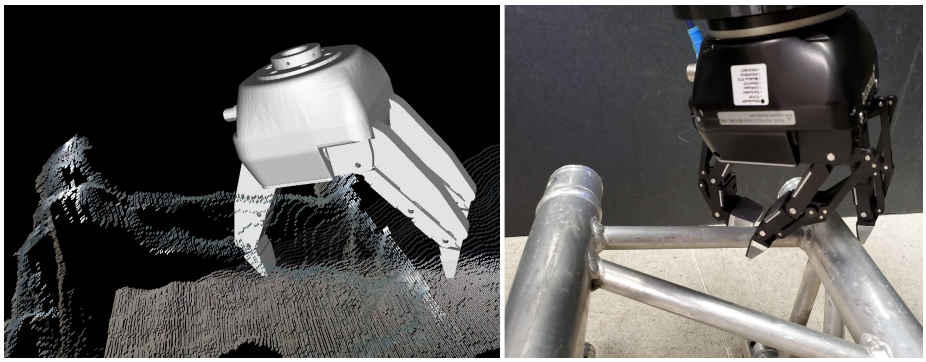

Figure 4. Left: Visualization of a grasp planned on a pointcloud of an aluminum truss segment. Right: The realization of that planned grasp. Note that the closure location is both significantly above, and offset towards the camera, relative to the planned location; a result of hand-eye-calibration having drifted.

relative to that location. This is particularly true of the field of manipulation, as manipulands often possess complex geometries of which only a subset may comply with a specific affordance; while available manipulators may only afford small margin for error in finger positioning (as explored in Section 2). Within this system, drift of calibration was manually observed by the operator and the below described pipeline triggered.

An example of this is depicted in Figure 4, where in the left image we see the gripper superimposed on the pointcloud derived from an RGBD image of the object in question (a truss segment). The gripper is placed at a pose proposed by the grasp planner (Section 6.1) where the fingers should well encompass the extremeties of the object; however, in the image to the right we can see that the resulting pose of the gripper when moved to the desired pose in the robot frame is both translationally (5cm measured) and rotationally offset from the intended location relative to the object. This results in closing of the fingers either missing, or closely grazing, the object, imparting insufficient prehensile contact to affect a lifting action that overcomes the weight of the object.

The prime cause of the error in hand-eye calibration within the RoMan system in particular is the mechanical calibration of the pan tilt unit (PTU), which employs counter-tightened bolts instead of a pin to retain position on the pan and tilt axes. When coupled with the, at times, juddering motion of the torso joint and arms, this imperfect retention mechanism results in slow drift of any sensor mounted atop the PTU with respect to the hard-stop positions that are used to calibrate position at start-up.

While addressing the mechanical cause of such drift would be a fruitful pursuit in design of future systems, the many other sources of discrepancy between intended, and actual, gripper pose relative to the pose of sensed objects, including drift of articulated arm actuators, inaccuracy of kinematic model, slop/compliance of sensor mounts. Eliminating or mitigating each of these in isolation is desirable, but their potential to compound into a pose error that results in failed grasp attempts motivates a means of addressing these inaccuracies in concert via a principled, closed loop, correction. A means to achieve this can be afforded by positioning the articulated actuation system such that a particular geometric or visual element termed a *fiducial* is identified/placed on the system, then the arm moved through a series of joint configurations such that the fiducial is kept within view of the sensor.

The expected pose of the fiducial marker at each of these joint configurations is computed through the kinematic model of the platform, ideally in the frame at which the kinematic tree bifurcates between the path to sensor and end-effector, for instance at the shoulder, so as to minimize the transformations between them and eliminate additional error from subsystems outside the path from sensor to hand. As an example of this, the fiducial chosen for the RoMan platform was the blue power LED on the side of the robotiq gripper, the location of which in the gripper base frame can be seen in Figure 5 top left. The fiducial location at each joint configuration must then be

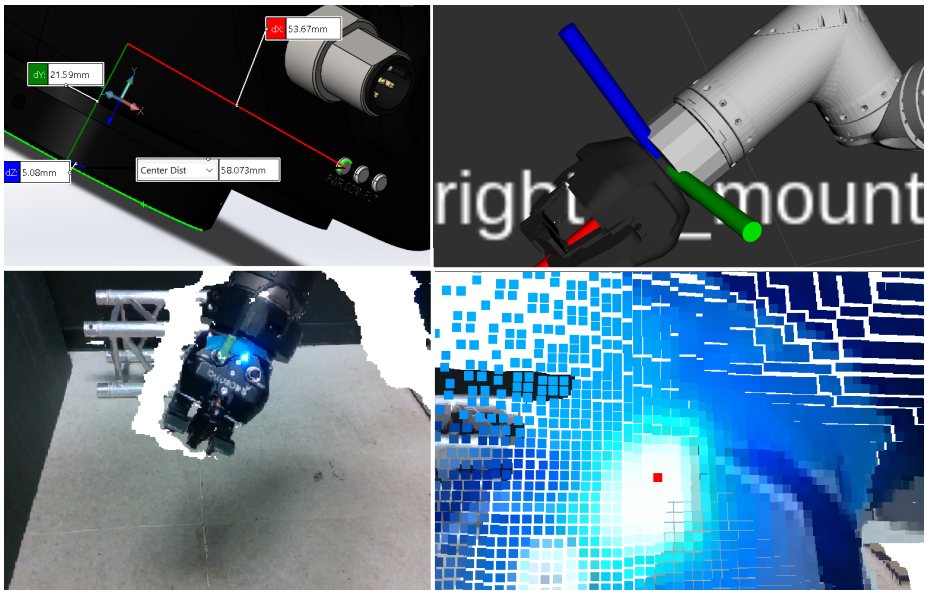

Figure 5. Top Left: Location of fiducial marker (green circled LED) in wrist mount frame. Top Right: Wrist mount frame pictured against the 7-DoF arm. Bottom Left: Pointcloud from RoMan's PTU mounted RGBD camera including fiducial marker. Bottom Right: Point corresponding to fiducial marker selected within operator's interface.

transformed back from the wrist frame, seen Figure 5 top right, into the shoulder frame. A series of pointclouds is captured viewing the fiducial in the range of joint configurations chosen to provide sparse coverage of the visible workspace (Figure 5 bottom left), then point corresponding to the fiducial tagged to provide a location in the camera frame, as in Figure 5 bottom right.

We then run an L2 optimization to find the the camera-to-shoulder transform that minimizes the root mean square error (RMSE) of the Euclidean distance between the fiducial location predicted via forward kinematics, and the location measured by the sensor. The transform between the shoulder and camera is then replaced with this result, which minimizes the error across all fiducial locations tested. In the case of the RoMan platform, 5 joint configuration/fiducial pose pairs were tested in each calibration set, typically producing an RMSE of <=4mm, and markedly improving hand accuracy after drift was detected.

5. Grasp Selection In Clutter

Grasp selection in unstructured environments has proven a challenging task for the robotics community, and is complicated further when a system lacks *a priori* knowledge of the shape and mass properties of objects it may be called upon to manipulate.

To dislodge a candidate extraction object from a pile of objects, the system must be capable of breaking the *stiction* restraining it. This describes the static friction that must be overcome to enable motion of a given stationary object relative to its contacting surroundings, and is a portmanteau of static and friction. Failure to suitably predict or detect excessive stiction or mass of a candidate manipuland also has the potential to cause catastrophic damage to a system, as occurred in Figure 12 right, which could render an autonomous agent inoperative.

Prior art has sought to address the problems of object agnostic grasp synthesis (Bone et al., 2008; Bohg and Kragic, 2010; Mahler et al., 2017b), grasping of known and unknown objects amongst clutter (Zeng et al., 2019; Boularias et al., 2015), as well as lifting of massive objects with wrench constrained end effectors or actuators (Kanoulas et al., 2018). This work seeks to address the intersection of these, in particular the disassembly of unstructured piles of massive

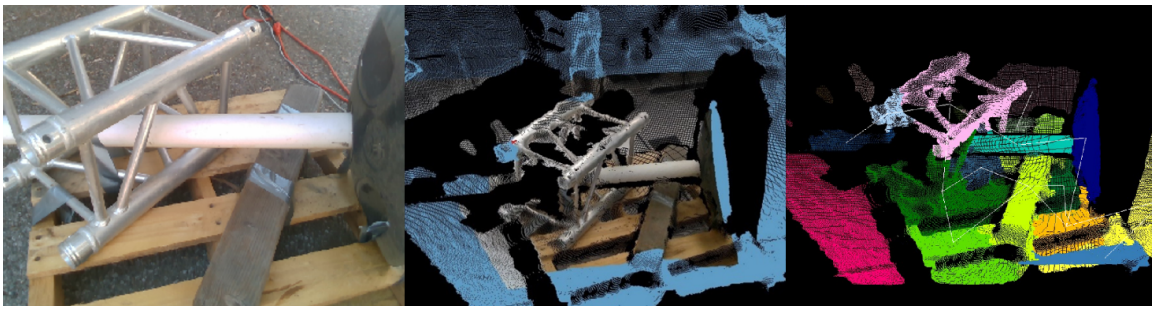

Figure 6. Left: RGB image of example debris pile containing aluminum truss segment, safety barrier, 4x4 wood section, and pallet. Center: Point cloud of workspace captured with RealSense D435. Right: Singulated object candidates with geometric adjacency from Locally Convex Connected Patches algorithm. (Stein et al., 2014)

objects (e.g. Figure 1), where lifting one object may induce lifting or pulling of other objects, which in turn increases the required grasp wrench, and may exceed the capabilities of the manipulation system.

Of first concern when selecting a grasp is the problem of singulating objects within the pile and determining adjacency, which is furnished in this implementation by the Locally Convex Connected Patches algorithm (Stein et al., 2014) in Figure 6 right. Once the objects within the pile have been suitably singulated, they may then be selected between in order to determine a suitable candidate for object extraction from the pile, via designation of a volume of the manipulation workspace termed the 'region of interest'.

5.1. Region of Interest Selection

Here we define the volume that is supplied to the grasp planner for the purpose of synthesizing form compliant grasps within its bounds. Points within the specified volume are matched against the designated grasp prototypes, while points outside that region are checked against the collision volumes used to reject grasp poses that would conflict with surrounding geometry. In order for elements of the pile within the manipulation workspace to be chosen, they must first be discriminated into candidate objects, or at least geometric regions that would be viable for grasping, as achieved through the Locally Convex Connected Patches algorithm.

A baseline approach to selecting grasp planning regions was then developed which allowed deconstruction of 'simply stacked' piles; where objects rest upon those immediately above them along the line of gravity without interlocking, a minimal example of which is seen in Figure 7 left. This employed a designated 'priority point', typically defined above the manipulation workspace of the platform in the base frame, as pictured in Figure 7 left, and the visually segmented object with the center of mass (average of point positions) of shortest Euclidean distance to the priority point would be selected first. If the grasp planner were incapable of planning grasp poses of a minimum score upon the pointcloud of the topmost selected object, it would be deemed presently infeasible for hand closure, and the next most distant region would be selected.

Let the operator specified 'priority point' be $P \in \mathbb{R}^3$, and the center of mass of each N segmented objects (mean pixel value of each colored region in Figure 6) be termed $O_i \in \{1, \ldots, N\}$. The index of the initial candidate object region selected for extraction from the pile, e_1, is then taken as

$$e_1 = \underset{i}{\operatorname{argmin}} \left\| P - O_i \right\|_2 . \tag{1}$$

Experiments showed this perfectly suited to dealing with the aforementioned 'simply stacked' piles, but failed when any more complex inter-object geometric relationships were introduced, including other objects incident upon the topmost object adding their weight to it, or yet more complex coming from the interlocking of objects, though the latter of these two is not addressed

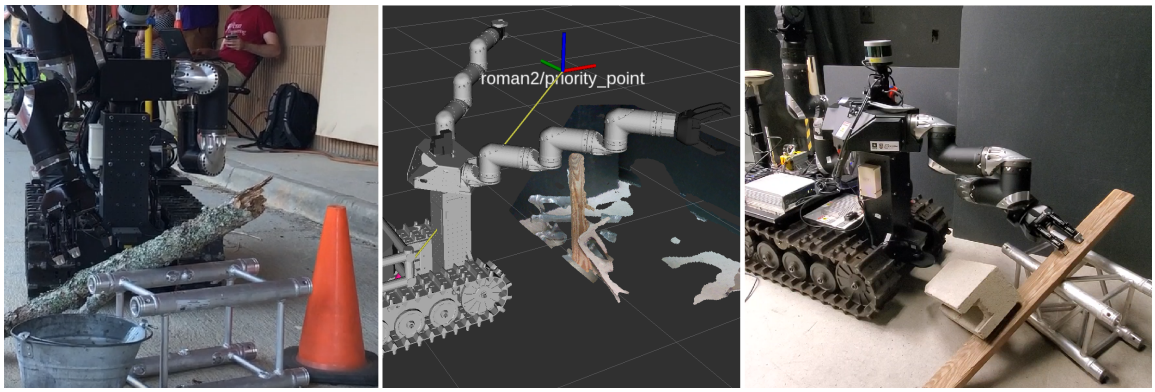

Figure 7. Left: Example of a 'simply stacked' pile, where the piece of wood lies directly atop the truss along the line of gravity, without any interlocking or other items resting upon it below its highest point. Center: Location in manipulation workspace from which the object centroid of smallest Euclidean distance is assumed to rest atop the pile, and therefore be easiest to extract. Frame of priority point is pictured above the RGBD pointcloud of a pile of objects below. Right: Debris pile configuration demonstrating failure case of priority point region selection method. The incident weight of the cinderblock on the wood piece causes the pinch grasp on the wood piece to fail during a lift motion, where similar unencumbered grasps succeed.

here. A minimal example of an incident object adding weight to the topmost is demonstrated in Figure 7 right, where a heavy cinderblock rests against a piece of wood that appears to be topmost on the pile, when selected using the priority point metric.

While the 'priority point' approach could not account for the interlaying of objects within a pile as seen in Figure 7 right, the mechanical capabilities of the system proved to be sufficient to dislodge encumbered items such as the wood length pictured. An example of this is captured in supplementary media C. In this regard, the strength of actuation of this particular system obviated the need for greater interpretation of the object pile structure with the classes of objects within the test set; however, application to a less capable system, plus manipulands of greater mass, or greater propensity for interlocking would motivate an enhanced evaluation of the singulated regions within the workspace.

6. Grasp Planning

The objective of the grasp planner is to compute grasps that will allow the robot to extract objects from the pile. As such, it is not necessary to plan accurate finger-object contact points. Instead, we implement grasping in an open-loop fashion: The planner computes a gripper pose that is such that closing the gripper is likely to yield a form- or force-closure grasp. The robot executes the grasp by moving to the planned pose, closing the fingers, and applying a constant squeezing force. The movement to the grasp pose is guarded by monitoring forces and torques at the wrists in real-time

As discussed above, a manipulation setup composed of a Realsense D435 camera and the 7-DoF RoMan arm is capable of an end-to-end accuracy of approximately 1cm. The Robotiq gripper is capable of encompassing objects whose diameter is comprised between 2cm and 15cm; the CamHand's fingers open are capable of turning by 180° and therefore provide a larger opening that can handle objects of up to 18cm. Given the grasping strategy outlined in the previous paragraph, the 1cm accuracy exhibited by the manipulation system allows the robot to grasp objects that are up to 13cm-wide (i.e., $15 - 2$cm) with its Robotiq gripper, without risk of colliding with the object. It can grasp objects that are up to $18 - 2$cm wide with the CamHand. The 1cm accuracy exhibited by the manipulation system allows the robot to avoid obstacles that are a few centimeters or further from the target object.

6.1. Grasp Synthesis and Quality Computation

The central component of the grasp planner is a model that tests the quality of a given grasping pose, which amounts to the likelihood that this grasping pose will yield a force- or form-closure grasp on the object. Let us denote by t an $\mathbb{R}^3$ transformation, and r an $SO(3)$ rotation. Given a scene s, obtained from a depth image and represented by a point cloud, we denote the quality of a grasp at (t, r) in s by $Q_{(t,r)}(s)$. In the rest of the text, we will assume that s is constant, and write $Q_{(t,r)} = Q_{(t,r)}(s)$.

$Q_{(t,r)}$ is driven by three factors: (1) suitability of the shape of the object directly facing a gripper positioned at (t, r), (2) reachability of (t, r) (from an arm kinematics standpoint), and (3) whether placing a gripper at (t, r) results in a collision or not. Let us denote those three factors by $S_{(t,r)}$, $R_{(t,r)}$, $C_{(t,r)}$, where $S_{(t,r)}$ returns a suitability value between 0 and 1, and $R_{(t,r)}$ and $C_{(t,r)}$ are binary functions whose value is 0 or 1 depending on the existence of an inverse-kinematics solution at (t, r) and whether placing the hand at (t, r) creates a collision with the scene s. $C_{(t,r)}$ is implemented by verifying whether the volume occupied by the hand intersects with 3D points measured by the robot's vision system. We note that this solution allows us to discard grasps whose *final* hand pose collides with *visible* obstacles. Failures resulting from collisions with hidden obstacles are managed at the task level.

Ideally, we would define $Q_{(t,r)}$ as:

$$Q_{(t,r)} = S_{(t,r)} R_{(t,r)} C_{(t,r)}, \tag{2}$$

and compute a grasp with

$$\underset{t \in V, r \in SO(3)}{\operatorname{argmax}} \; Q_{(t,r)}, \tag{3}$$

where V is the volume defined in Section 5.1, and the *argmax* operator implements a suitable optimizer. Unfortunately, the modular nature and simplicity of the RoMan arm design required the sacrifice of a spherical wrist, and it deprived us from fast, analytical inverse kinematics (IK). Computing a numerical IK solution for the RoMan arm costs 20ms on average, compared to the 20μs typically needed for analytical IK. This cost has a prohibitive impact on Equation 3. As a result, we factorize the search for a grasp into two independent steps.

In step 1, we search for grasps that minimize an auxiliary quality $\hat{Q}_{(t,r)}$ defined as

$$\hat{Q}_{(t,r)} = S_{(t,r)} C_{(t,r)}. \tag{4}$$

We conduct this search via simulated annealing (Kirkpatrick et al., 1983) on a Markov chain (Andrieu et al., 2003) whose invariant distribution is an increasing power of $\hat{Q}_{(t,r)}$. The chain is defined with a mixture of two local- and global-proposal Metropolis-Hastings transition kernels. We a set of grasps G located near the modes of $\hat{Q}_{(t,r)}$ by extracting states of the chain whose quality is above a fixed threshold. Intuitively, the method alternates between a hill-climbing policy that reveals local pose maxima, and random jumps in the pose space that allow multiple local maxima to be discovered. This approach is discussed in greater detail in our prior work (Detry and Piater, 2010), and a free implementation is available at https://github.com/renauddetry/nuklei.

In step 2, we select the grasp from G that maximizes our original quality metric $Q_{(t,r)}$ (which includes reachability).

The shape-suitability measure $S_{(t,r)}$ is the key factor of Q (and $\hat{Q}$). Our definition of S is data-driven: we define S to be proportional to the similarity between the shape of a grasp example and the shape of the gripper. Data-driven grasp metrics are commonplace nowadays (Detry et al., 2013; Herzog et al., 2014; Mahler et al., 2017a; Viereck et al., 2017). In prior work, such metrics have been defined by (a) exhaustively comparing the query grasp to all examples of a training set (Herzog et al., 2014), (b) by compressing the training set into an implicit score function such as a deep network (Mahler et al., 2017a; Viereck et al., 2017), or (c) a combination of both (Detry et al., 2013). In this work, we provide the robot with a set of grasp *prototypes* (Figure 8) that represent a range

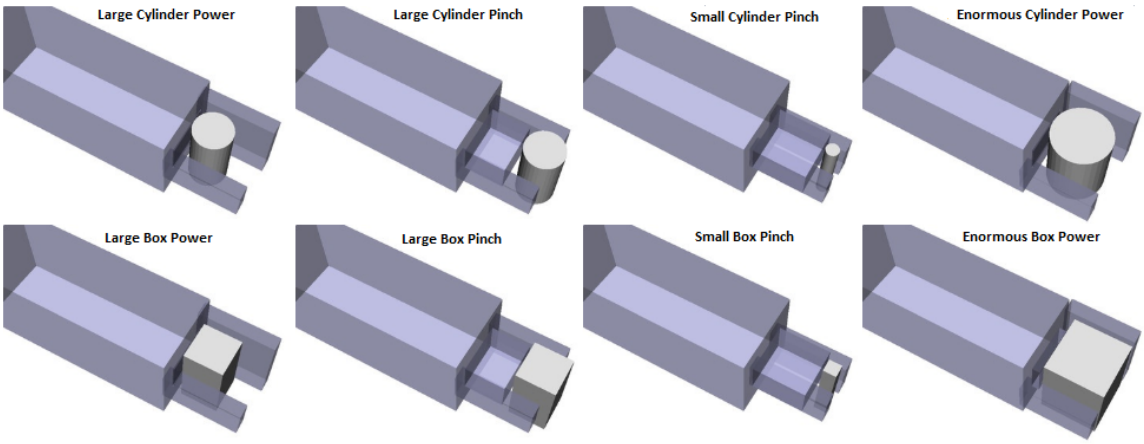

Figure 8. Example prototype geometry and grasp pose pairs applied to the manipulation planning for generic objects. Objects are referenced against the palm of the gripper, which is formed by the object-ward face of the largest rectangular prism.

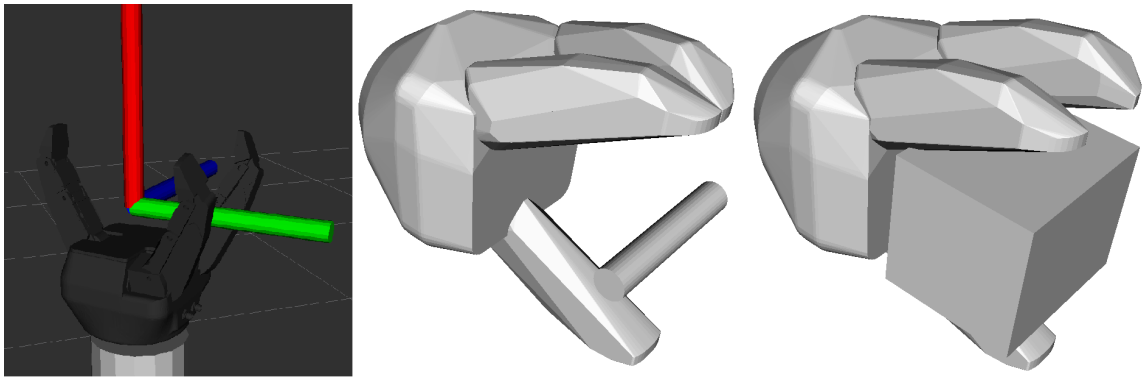

Figure 9. Left: Grasp object frame relative to Robotiq gripper geometry. Center: Hand preshape for pinch grasp on 2cm diameter cylinder prototype. Right: Hand preshape for power grasp on 9cm-wide box prototype.

of geometric primitives that the objects the robot encounters may be expected to be comprised of. While we have, in this paper, constructed this dictionary by hand, our previous work showed how such a dictionary can be learned from training examples (Detry et al., 2013). A grasp prototype is composed of a shape model, a gripper pose, and a gripper preshape. We encode the shape as a point cloud, and, by convention, we express it in the reference frame of the gripper (see Figure 9 left).

A grasp 'preshape' defines the positioning of the fingers of the gripper specific to the geometric primitive prototype for which it is designed. For example, in Figure 9 center, a pinch grasp on a 2cm cylinder positions the two finger pair together, so as to directly oppose the "thumb", both reducing susceptibility to deviation about the X grasp axis (red in Figure 9 left) and minimizing risk that applying a bending moment to the object through widely spaced fingers may damage it. Conversely, for a larger and assumedly more rigid object box-like object in Figure 9 right, the two fingers opposing the thumb are spaced wider apart to provide greater stabilization against adverse torques about the X grasp axis that might work to wrest the object from the grasp (gravity or otherwise).

Let us denote by I a point cloud issued from a depth image and containing an object of interest, by $P = \{p_i\}_{i\in[0,N-1]}$ a dictionary of N prototypes, and by $T_{t,r}(x)$ a function that rotates the point cloud x by r then translates it by t. The shape-suitability measure $S_{(t,r)}$ of a grasp at pose (t, r) is

Table 1. Tabulation of outcome modes for grasp planner validation experiment campaign when using generic prototype library on small pile of varied objects, representative of human scale manipulation environment targeted by the platform.

Outcome	Count	% over all faults	% over grasp pipeline faults
Success	19	44.2%	73%
Bad grasp	7	16.3%	27%
Operator error - Grasp parameters	5	11.6%	Total valid grasp tests: 26
Operator error - Gripper initialization	4	9.3%	
Mechanical fault	3	7%	
Joint planning failure	2	4.7%	
Calibration	2	4.7%	
Hardware driver	1	2.3%	
Total over all faults	43		

then defined as:

$$S_{(t,r)} = \max_{i \in [0, N-1]} c\left(T_{(t,r)}\left(p_i\right), I\right) \tag{5}$$

where $c(\cdot, \cdot)$ is a measure of similarity between two point clouds, which is equal to the average distance between points of $T_{(t,r)}\left(p_i\right)$ that are visible from the camera's viewpoint, and their nearest neighbor in I. For more detail, we refer the reader to our prior work (Detry et al., 2013).

6.1.1. Grasp Planner Validation Experiments

While the grasp planner was employed extensively across a variety of manipulation behaviors documented in (Kessens et al., 2021), a restricted set of experiments was conducted to directly validate its capacity to grasp representative objects for the human-scale urban environment within which the system is intended to operate. Trials were conducted on the aluminium truss segment, 2"x4" wood section, and safety barrier pictured in Figure 1, both in isolation and when combined into a pile. Due to failures of the grasp pipeline originating from a wide variety of subsystems within the software and hardware architecture of the platform, assessment of the success of the pipeline is broken down by cause of failure. This affords the ability to introspect the relative reliability of different subsystems, in order to perform a combinatorial system reliability analysis.

As an example, the first grasp validation experiment campaign experienced consistent hand-eye calibration issues over course of 20 trials, motivating the work described in Section 4. This manifested as frequent finger-object collisions during the Cartesian motion from pre-grasp to grasp pose, indicating that the alignment of the collision boxes used within planning was in poor agreement with the realized tool-object relative pose. Enacting a regimented hand-eye calibration procedure mitigated these frequent misalignments, allowing the remainder of the grasping pipeline to be evaluated in subsequent experimental campaigns. The outcome modes across the subsequent campaigns are tabulated in Table 1, where column 3 represents percentage incidence across all initiations of the grasp planning pipelines, and column 4 represents percentage incidence across all valid tests of the grasp pipeline itself, where no other subsystems faulted during planning and execution of a grasp.

Note that the 5 'Operator error - Grasp parameters' outcomes were caused by an optional ground plane subtraction processing step having been inadvertently enabled, while incoming pointclouds had already been cropped to remove the floor. This resulted in a ground plane being fitted to the object pile instead, and the elimination of most points in the manipulation workspace, as alluded to in Section 8 3. When solely evaluating the success rate of the grasp attempts that were actually enacted by the system, against the bad grasps that were enacted by the system, we attain an outcome of 73%.

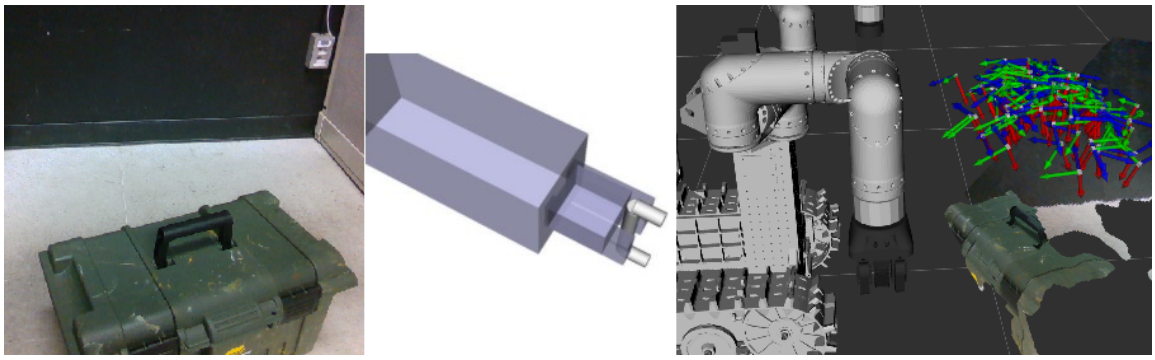

Figure 10. Left: Example of specific use-case object, in the form of a hinged lid crate, where a specific region of the geometry must be grasped to enact the desired affordance (lifting or opening). Center: Custom grasp prototype designed for the purpose of realizing the necessary affordance for the given object class. Right: Grasps planned on the specific use-case object visualized within the manipulation workspace.

6.2. Object Specific Prototypes

While the library of generic prototypes pictured in Figure 8 allows the platform to synthesize grasps on geometric primitives with multiple axial symmetries, there may be times when the platform is required to utilize affordances (Gibson, 1966) of a particular class of objects that necessitate precision grasps on specific geometries common to the class that are known *a priori*. An example of this comes from the hinged-lid crate pictured in Figure 10 left, which may be both lifted, and opened, from the protruding handle; as may be necessary to inspect and extract any contents in the scenarios the platform was tasked with.

Figure 10 center shows a custom grasp prototype that was constructed for the purpose of exploiting the affordances of lifting and opening objects equipped with a handle geometry common to one-handed crates and bags. A key distinction from the cylindrical prototypes of the generic library is that the addition of the end supports of the handle collapses the continuous axial symmetry of the geometry to a single approach vector colinear with the handle supports, thereby typically normal to the body of the object so as to allow maximum surrounding space for ease of grasp closure. Figure 10 right visualizes a set of grasps synthesized on the crate using the custom prototype.

While this affordance specific prototype allows the platform to grasp preordained regions of particular objects, the quality metric produced by applying the prototype, $Q_{(t,r)}$, is not sufficient to accurately identify the presence of the affordance specific geometry in the scene. Additional context must be supplied from higher level reasoning, for example in the form of deep learning networks that identify the presence of the object class in the scene, in order for the affordance specific prototype to be engaged within the manipulation region of interest (Do et al., 2018; Kokic et al., 2017).

6.2.1. Object Specific Prototype Validation Experiments

In a similar regard to general clutter grasping, (Kessens et al., 2021) documents experiments relating to the grasping of a handle of a crate in order to utilize the affordance of opening; therefore, this paper only seeks to validate the grasp planner with regard to grasping the object class associated with the handle affordance prototype, shown in Figure 10 center. The ability of the pipeline to synthesize suitable grasps on the object specific prototype was analyzed using a dataset that was derived from experimental exercise in removing the test article with specific handle type, shown in Figure 11a, from within the container pictured in Figure 10 left. The experimental results are presented in Table 11b, and include 147 datapoints across the test set. It should be noted, however, that some of the joint planning and hardware driver failures occurred during a preparatory behavior of opening the container lid, the design of which is not considered in this article, but the failure of which still results in the system failing to perform the end-goal of grasping the test article. The

(a) Test article for object specific grasp prototype testing.

Outcome	Count	% over all faults	% over grasp pipeline faults
Success	95	64.6%	96%
Bad grasp	4	2.7%	4%
Joint planning failure	34	23.1%	Total outcomes of grasp pipeline: 99
Hardware driver	10	6.8%	
Operator error	4	2.7%	
Total over all faults	147		

(b) Tabulation of outcome modes for grasp planner validation experiment campaign with object specific prototype, a small handle, when object is grasped from inside a container.

Figure 11. Object specific grasp pipeline validation experiment results

Figure 12. Left: RoMan platform grasping a "Czech Hedgehog" anti vehicle barrier prior to a drag behavior. Failure to account for the motion of the grasp point relative to the ground support contacts during lifting motion results in a large torque being exerted on the gripper about the Y axis of the FLU robot frame. Right: Broken proximal joint on digit 3 of a Robotiq 3-finger gripper after heavy object lift exceeded safe torque levels.

kinematic complexity of the container opening prior to grasping results in a higher incidence of joint planning failures than during multiple generic object experiments in Section 6.1.1, whereas hardware driver errors are of a similar order of magnitude. Operator error is lower than multi-object generic prototype grasping due to a more regimented test procedure. Accounting for all sources of fault results in a success rate of 64.6%.

Of prime concern, however, is the performance of the grasp pipeline within this operation. The pipeline produced 95 successful grasps out of the 99 total grasps attempted on the test article and, when eliminating the faults not related to the grasping pipeline, results in a grasp pipeline success rate of 96%. This is somewhat higher than that of the multi-object grasping with generic prototypes, at 73%, and may be attributed to the greater conformance of the object specific prototype to the geometry of the test article given the bespoke design. A video of an object specific grasp being conducted on the test article, while outside the container used within experiments, is included in supplementary media B.

7. Intrinsic Controller

In order to lift constituent objects within debris piles (Figure 1), as well as drag larger items that may be encountered within urban environments (Figure 12 left), the manipulation system must be capable of imparting sufficient force to exceed the mass of such objects, without incurring damage to the end-effectors used to apply that force. Exceeding safe force or torque limits of gripper mechanisms, as is easily possible during forceful manipulation of large objects, can lead to catastrophic failure of actuators as in Figure 12 right.

While gross, free space motions of the end-effector with a massive object in-hand are possible, any anisotropic wrench limits of the gripper may impose constraints on end-effector orientation during

a motion. This is a tractable motion planning problem, but outside the scope of this work, such that forceful manipulation of these massive objects was conducted entirely through use of inverse Jacobian Cartesian motions of the arm, coupled with planned motions of the tracked base. Joint trajectories for gross end-effector motions were produced using the Search-based Motion Planing Library (SMPL) and MoveIt! as described in greater detail in the companion paper (Kessens et al., 2021) section 3.6.2, while inverse Jacobian Cartesian motions were computed through a software module termed the "Intrinsic Controller".

Baseline operation of the Intrinsic Controller produces linear end-effector paths between the starting pose and the prescribed goal pose via fixed length step interpolation mapped through Jacobian pseudo inverse based numerical IK.

7.1. Wrench Reactive Control

Compliance of the end effector trajectory during a controlled motion is specified per axis for translation and rotation. For ease of operator use, the wrench upon which the compliance acts is defined in an 'End effector Centric, Robot Aligned' (ecra) frame.

Deflection is calculated from the change in wrench experienced by the wrist mounted force torque sensor over the course of the motion; this is achieved by taring the instantaneous wrench, W_{ft}^{inst}, against that measured upon the controller being engaged, W_{ft}^{init}. Exponential smoothing is used to filter the force torque wrench measurements and produce a damped deflection response, with smoothing factor $\alpha = 0.01$ in the 250Hz control loop, where W_{ft}^{filt} is then used to calculate instantaneous deflection.

$$W_{ft}^{tare} = W_{ft}^{inst} - W_{ft}^{init}, \tag{6a}$$

$$W_{ft}^{filt} = (1-\alpha)W_{ft}^{filt} + \alpha W_{ft}^{tare} \tag{6b}$$

The adjoint of the homogeneous transform between two frames allows conversion of the dual space of wrenches between those frames (Murray et al., 1994), which brings the wrench measured at the force torque sensor into the end effector frame, within which we must operate within payload limits of the fitted tool. The wrench experienced in the end-effector frame is then given by W_{ee} where a *homogeneous transformation* is termed $g = \begin{bmatrix} R & \bar{p} \\ 0 & 1 \end{bmatrix}$. A homogeneous transformation may then be used to describe the transformation of a wrench between coordinate frames via the *adjoint transformation* $Ad_g = \begin{bmatrix} R & \hat{p}R \\ 0 & R \end{bmatrix}$, which permits the wrench transformation $W_a = Ad_{a2b}^T W_b$. Here $a2b$ is the name of a homogeneous transform from coordinate frame b to coordinate frame a.

For example, the wrench measured at the force torque sensor, W_{ft}, may be transformed into the end-effector frame via

$$W_{ee} = Ad_{ee2ft}^T W_{ft} \, . \tag{7}$$

We then define an intermediate *hyb* as the hybrid frame of EE (end-effector) position with robot orientation via the purely rotational adjoint

$$Ad_{hyb2ee} = \begin{bmatrix} R_{robot2ee} & 0 \\ 0 & R_{robot2ee} \end{bmatrix} , \tag{8}$$

$$W_{hyb} = Ad_{hyb2ee}^T W_{ee} \tag{9}$$

The benefit of this hybrid frame is that it allows the operator to specify rotational and/or translational compliance components relative to the world frame, meaning they do not have to consider the present orientation of the end effector. The hybrid frame is then applied across the directed Cartesian motion by defining the tensor

$$C_{hyb} := diag(c_{f_x}, c_{f_y}, c_{f_z}, c_{\tau_x}, c_{\tau_y}, c_{\tau_z}) \, . \tag{10}$$

Within the control loop for the Cartesian motion, a deflected goal pose $G_{robot2ee_{defl}}$ is then computed in response to the specified compliance tensor and instantaneous wrench measurement by first attaining the twist imposed in the end effector frame by the compliance

$$T_{hyb}^{defl} = C_{hyb} W_{hyb} \ , \tag{11}$$

$$T_{ee}^{defl} = Ad_{ee2hyb} T_{hyb}^{delf} \ . \tag{12}$$

This may then be taken back into homogeneous coordinates ($G \in SE(3)$) via the matrix exponential and applied to the nominal goal pose of the Cartesian motion without any compliance $G_{robot2ee}$ to produce the deflected goal pose

$$G_{ee2ee_{defl}} = e^{T_{ee}^{defl}} \ , \tag{13}$$

$$G_{robot2ee_{defl}} = G_{robot2ee} G_{ee2ee_{defl}} \ . \tag{14}$$

This control approach enabled safe lifting of a variety of objects by specifying compliance in the task-relevant frame. The chief example within the course of the program was the clearing of "Czech hedgehog" from Figure 12, where no additional mechanical failures were encountered during lifting of the heavy and cumbersome anti-vehicle barriers after wrench-reactive Cartesian control was applied.

8. Lessons Learned

8.1. The supreme importance of hand-eye calibration.

Much as articulated in Section 4, any and all computational efforts to synthesize mechanically and task compliant grasps is for naught if the corporeal end-effector's positioning does not coincide with the planned grasp in the object frame. For this reason, ensuring accurate hand-eye calibration is of paramount importance to the success of any manipulation system, either through rigorous mechanical design, utilizing components that do not drift in alignment, or realizing efficient recalibration pipelines, such as that described in Section 4, which we eventually adopted.

8.2. Compatibility of kinematic planner and controller configurations.

The grasping pipeline described herein adopted a common arm-motion paradigm, wherein a planner directs gross motions with collision avoidance to reach the *pregrasp* gripper pose then switches to a Jacobian-based controller and moves the gripper into contact with the object of interest. A recurrent problem within the RoMan system was that the transition between these joint command schemes would sometimes fault, due to either their managing module failing to engage the controller, or joint limit ranges having been reached in the planner that exceeded the limits with the controller, or vice versa. For consistent operation, the management of these schemes and joint-range limits should be designed to minimize risk of conflict between the schemes.

8.3. Be wary of varying context when making assumptions.

Assumptions made within one operational environment cannot – and should not – be applied to all conceivable environments. A standard operation within the preparation of each point cloud for grasp planning was to identify the dominant plane and subtract all points within and beneath it, premised on the ground or a table surface always being largely visible within the cloud. This assumption proved valid for extensive single-arm, tabletop testing, and mobile operations on flat ground, but failed when conducting pile-decomposition experiments in a congested laboratory environment leading to all points being cropped from the cloud. Where possible, such assumptions and processing steps should ecosystem, requiring them able to communicate be evaluated programmatically, based on

additional contextual information available, such as the surface upon which the platform base rests being used to infer the plane plane.

8.4. Avoid adopting software components that are overly prescriptive on the rest of the system.

Underlying drivers that command joints within the platform's RoboSimian-style arms utilize a library that only supports Linux kernels up to Ubuntu version 14.04, which imposed a relatively arbitrary restriction on all remaining software packages within the ecosystem, requiring them able to communicate with an Ubuntu 14.04 (ROS Indigo/Jade) OS. While some less common or open source drivers for hardware protocols (such as EtherCAT) inevitably place constraints on the operating system they can operate within, the merits of maintaining robustness to operating system version cannot be overstated.

8.5. Ensure reactive force protections and controls are measured relative to safe levels.

By virtue of the slow, consistent thermal drift of the force-torque sensors adopted for the wrists of the RoMan platform, force and torque safety stops and reactive controllers were always calculated with respect to those values on commencing motion. While this shortcut performed admirably in autonomous operation, where only a single forceful operation would occur before a release and regrasp if thresholds were exceeded, excitable operators occasionally executed multiple forceful motions in sequence (to achieve some placement goal or otherwise that was "ever so close"). This human eagerness had the effect of compounding force and torque thresholds with each execution, at times leading to mechanical failure of end effectors.

8.6. Make multi-dimensional operator parameters intuitive, or task relevant.

In defining the anisotropic compliance afforded by wrench-reactive control in Section 7.1, early iterations simply mapped compliance terms to the primary axes within the end-effector frame, which proved highly challenging for operators to specify. For that reason, transformation into the hybrid frame of EE location and robot-base orientation afforded operators a far more intuitive means of specification of axis specific rotational or translational compliance. During autonomous operation, operator intuition is not a factor, but a similar requirement may be applied when availing the platform of particular, orientationally constrained, object affordances; an example being the lift of an anti-vehicle barrier, where compliance must be specified about the two supporting legs.

9. Conclusions

- The object removal pipeline presented here proved capable of removing human-scale objects, both within a pile of general objects using generic grasp prototypes and when removing a specific object with a specialized grasp prototype.
- Applying our generic-prototype strategy to grasp and lift operations on a set of human-scale objects yielded good performance. Overall system success (including hardware and calibration faults) was 42%. When examining only the grasp pipeline in isolation, success reached 73%
- The object-specific prototype grasp experiments produced a net system grasp success rate of 64.6% when including all faults, and 96% when examining only the grasping pipeline.
- The "region of interest" selection algorithm – segmenting then selecting items along an approach vector – proved able to select suitable objects from atop a pile, save in edge cases where a heavier object rested directly on an upward-tilted candidate. System actuation capability proved sufficient to overcome even the latter cases, but the same would not be true in the instance of more massive manipulands, thus motivating further study.

- The "wrench-reactive" controller enabled intuitive operator specification of compliance parameters when lifting then dragging a large object, such as an anti-vehicle barrier. Here rotation about the initial grasp orientation was necessary to mitigate adverse wrench and prevent damage to end effectors.

Acknowledgments

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0016. 2021 California Institute of Technology. Government sponsorship acknowledged.

Contributors

The authors would like to thank:

- Chad C. Kessens, Matthew Kaplan, Trevor Rocks, and Philip R. Osteen of the Army Research Laboratory for their facilitation of the research program, coordination of development campaigns, and technical support in collecting some of the object specific grasp data presented herein.
- Long Quang, Long Quang, Mark Gonzalez, Jaymit Patel, Michael DiBlasi, Shiyani Patel, Matthew Weiker, and Dilip Patel of General Dynamics for their tireless development and maintenace of the research platforms, facilitation of experiments, and administrative support.
- Karl Schmeckpeper and Kostas Daniilidis of the University of Pennsylvania for their accommodating several collaborative experiment campaigns.
- Bryanna Yeh and David Handelman of the Johns Hopkins University Applied Physics Laboratory for their support in generating media to depict the synthesis of grasps using the grasping pipeline described herein.

ORCID

Joseph Bowkett https://orcid.org/0000-0002-3101-489X
Sisir Karumanchi https://orcid.org/0000-0002-0685-4125
Renaud Detry https://orcid.org/0000-0003-0597-1167

References

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43.

Antanas, L., Moreno, P., Neumann, M., de Figueiredo, R. P., Kersting, K., Santos-Victor, J., and De Raedt, L. (2014). High-level Reasoning and Low-level Learning for Grasping: A Probabilistic Logic Pipeline.

Atkeson, C. G., An, C. H., and Hollerbach, J. M. (1985). Rigid Body Load Identification for Manipulators. *Proceedings of the IEEE Conference on Decision and Control*, (December):996–1002.

Berenson, D., Srinivasa, S. S., Ferguson, D., and Kuffner, J. J. (2009). Manipulation planning on constraint manifolds. i:625–632.

Bicchi, A. and Kumar, V. (2000). Robotic grasping and contact: A review. *Proceedings-IEEE International Conference on Robotics and Automation*, 1:348–353.

Bohg, J. and Kragic, D. (2010). Learning grasping points with shape context. *Robotics and Autonomous Systems*, 58(4):362–377.

Bohg, J., Morales, A., Asfour, T., and Kragic, D. (2014). Data-driven grasp synthesis-A survey. *IEEE Transactions on Robotics*, 30(2):289–309.

Bone, G. M., Lambert, A., and Edwards, M. (2008). Automated modeling and robotic grasping of unknown three-dimensional objects. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 292–298.

Boularias, A., Bagnell, J. A., and Stentz, A. (2015). Learning to Manipulate Unknown Objects in Clutter by Reinforcement. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence Learning*, pages 1336–1342.

Burkhardt, M., Karumanchi, S., Edelberg, K., Burdick, J. W., and Backes, P. (2018). Proprioceptive Inference for Dual-Arm Grasping of Bulky Objects Using RoboSimian. *IEEE International Conference on Robotics and Automation*, pages 4049–4056.

Cini, F., Ortenzi, V., Corke, P., and Controzzi, M. (2019). On the choice of grasp type and location when handing over an object. *Science Robotics*, 4(27):1–18.

Correll, N., Bekris, K. E., Berenson, D., Brock, O., Causo, A., Hauser, K., Okada, K., Rodriguez, A., Romano, J. M., and Wurman, P. R. (2018). Analysis and observations from the first Amazon picking challenge. *IEEE Transactions on Automation Science and Engineering*, 15(1):172–188.

Detry, R., Ek, C. H., Madry, M., and Kragic, D. (2013). Learning a dictionary of prototypical grasp-predicting parts from grasping experience. In *IEEE International Conference on Robotics and Automation*.

Detry, R., Papon, J., and Matthies, L. (2017). Task-oriented grasping with semantic and geometric scene understanding. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Detry, R. and Piater, J. (2010). Continuous surface-point distributions for 3D object pose estimation and recognition. In *Asian Conference on Computer Vision*, pages 572–585.

Do, T.-T., Nguyen, A., and Reid, I. (2018). Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE.

Eppner, C. and Brock, O. (2013). Grasping unknown objects by exploiting shape adaptability and environmental constraints. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4000–4006.

Ferrari, C. and Canny, J. (1992). Planning optimal grasps. *Proceedings - IEEE International Conference on Robotics and Automation*, 3(May):2290–2295.

Gibson, J. J. (1966). The senses considered as perceptual systems.

Herzog, A., Pastor, P., Kalakrishnan, M., Righetti, L., Bohg, J., Asfour, T., and Schaal, S. (2014). Learning of grasp selection based on shape-templates. *Autonomous Robots*, 36(1-2):51–65.

Holladay, R., Lozano-Pérez, T., and Rodriguez, A. (2019). Force-and-Motion Constrained Planning for Tool Use. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Horton, T. E., Chakraborty, A., and St. Amant, R. (2012). Affordances for robots: A brief survey. *Avant*, 3(2):70–84.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., et al. (2018). Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*.

Kanoulas, D., Lee, J., Caldwell, D. G., and Tsagarakis, N. G. (2018). Center-of-mass-based grasp pose adaptation using 3d range and force/torque sensing. *International Journal of Humanoid Robotics*, 15(4): 1–26.

Karumanchi, S., Edelberg, K., Baldwin, I., Nash, J., Satzinger, B., Reid, J., Bergh, C., Lau, C., Leichty, J., Carpenter, K., Shekels, M., Gildner, M., Newill-Smith, D., Carlton, J., Koehler, J., Dobreva, T., Frost, M., Hebert, P., Borders, J., Ma, J., Douillard, B., Shankar, K., Byl, K., Burdick, J., Backes, P., and Kennedy, B. (2018). Team robosimian: Semi-autonomous mobile manipulation at the 2015 DARPA robotics challenge finals. *Springer Tracts in Advanced Robotics*, 121:191–235.

Kessens, C., Kaplan, M., Rocks, T., Osteen, P. R., Rogers, J., Stump, E., Hurwitz, A., Fink, J., Quang, L., Gonzalez, M., Patel, J., DiBlasi, M., Patel, S., Weiker, M., Patel, D., Bowkett, J., Detry, R., Karumanchi, S., Matthies, L., Burdick, J., Oza, Y., Agarwal, A., Dornbush, A., Saxena, D. M., Likhachev, M., Schmeckpeper, K., Daniilidis, K., Kamat, A., Mandalika, A., Choudhury, S., and Srinivasa, S. S. (2021). Human-scale mobile manipulation using roman. *Field Robotics*, To appear: Special Issue - RCTA.

Kessens, C. C., Fink, J., Hurwitz, A., Kaplan, M., Osteen, P. R., Rocks, T., Rogers, J., Stump, E., Quang, L., DiBlasi, M., Gonzalez, M., Patel, D., Patel, J., Patel, S., Weiker, M., Bowkett, J., Detry, R., Karumanchi, S., Burdick, J., Matthies, L., Oza, Y., Agarwal, A., Dornbush, A., Likhachev, M., Schmeckpeper, K., Daniilidis, K., Kamat, A., Choudhury, S., Mandalika, A., and Srinivasa, S. (2020). Toward fieldable human-scale mobile manipulation using RoMan. In Pham, T., Solomon, L., and Rainey, K., editors,

Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II, volume 11413, pages 418 – 437. International Society for Optics and Photonics, SPIE.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kleeberger, K., Bormann, R., Kraus, W., and Huber, M. F. (2020). A survey on learning-based robotic grasping. *Current Robotics Reports*, 1(4):239–249.

Kokic, M., Stork, J. A., Haustein, J. A., and Kragic, D. (2017). Affordance detection for task-specific grasping using deep learning. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 91–98. IEEE.

Kumar, V. R. and Waldron, K. J. (1988). Force Distribution in Closed Kinematic Chains. *IEEE Journal on Robotics and Automation*, 4(6):657–664.

Mahler, J., Liang, J., Niyaz, S., Laskey, M., Doan, R., Liu, X., Ojea, J. A., and Goldberg, K. (2017a). Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics. In *Robotics: Science and Systems (RSS)*.

Mahler, J., Matl, M., Liu, X., Li, A., Gealy, D., and Goldberg, K. (2017b). Dex-Net 3.0: Computing Robust Robot Suction Grasp Targets in Point Clouds using a New Analytic Model and Deep Learning. *arXiv preprint arXiv:1709.06670*.

Mason, M. T. (2018). Toward robotic manipulation. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(1):1–28.

Miller, A. T. and Allen, P. K. (2004). Graspit: A versatile simulator for robotic grasping. *IEEE Robotics and Automation Magazine*, 11(4):110–122.

Murray, R. M., Li, Z., and Sastry, S. S. (1994). *A Mathematical Introduction to Robotic Manipulation*. CRC Press (1700).

Nguyen, V.-D. (1988). Constructing force-closure grasps. *The International Journal of Robotics Research*, 7(3):3–16.

Smith, C., Karayiannidis, Y., Nalpantidis, L., Gratal, X., Qi, P., Dimarogonas, D. V., and Kragic, D. (2012). Dual arm manipulation—a survey. *Robotics and Autonomous Systems*, 60(10):1340–1353.

Stein, S. C., Schoeler, M., Papon, J., and Worgotter, F. (2014). Object partitioning using local convexity. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 304–311.

Trinkle, J. C. (1992). On the Stability and Instantaneous Velocity of Grasped Frictionless Objects. *IEEE Transactions on Robotics and Automation*, 8(5):560–572.

Viereck, U., Pas, A. t., Saenko, K., and Platt, R. (2017). Learning a visuomotor controller for real world robotic grasping using simulated depth images. In *Conference on Robot Learning*.

Zeng, A., Song, S., Welker, S., Lee, J., Rodriguez, A., and Funkhouser, T. (2018). Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Zeng, A., Song, S., Yu, K. T., Donlon, E., Hogan, F. R., Bauza, M., Ma, D., Taylor, O., Liu, M., Romo, E., Fazeli, N., Alet, F., Chavan Dafle, N., Holladay, R., Morona, I., Nair, P. Q., Green, D., Taylor, I., Liu, W., Funkhouser, T., and Rodriguez, A. (2019). Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *International Journal of Robotics Research*, pages 1–16.

Zhang, L. and Trinkle, J. C. (2012). The application of particle filtering to grasping acquisition with visual occlusion and tactile sensing. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 3805–3812.

How to cite this article: Bowkett, J., Karumanchi, S., & Detry, R. (2022). Grasping and transport of unstructured collections of massive objects. *Field Robotics, 2*, 385–405.