Semantic and Geometric Scene Understanding for
Task-oriented Grasping of Novel Objects from a Single View

Renaud Detry  Jeremie Papon  Larry Matthies
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

Abstract—We present a task-oriented grasp model, that learns grasps that are configurationally compatible with a given task. The model consists of a geometric grasp model, and a semantic grasp model. The geometric model relies on a dictionary of grasp prototypes that are learned from experience, while the semantic model is CNN-based and identifies scene regions that are compatible with a specific task. A key element of this work is to use a deep network to integrate contextual task cues, and defer the structured-output problem of gripper pose computation to an explicit (learned) geometric model. Jointly, these two models generate grasps that are mechanically fit, and that grip on the object in a way that enables the intended task.

I. INTRODUCTION

This paper addresses task-oriented grasping onto new objects using a single depth image (single viewpoint). Our solution endows robotic agents with the ability to grasp objects that enable the execution of the intended task. This skill enables the use of new tools and objects, which is vital to our robots transition to uncontrolled environments. In this domain, our community has focused on two important issues: developing grasp models and developing task models. Grasp models [12], [11], [18] determine grasping points that are suitable for picking up an object, while task models [2], [16] often assume the pre-existence of a satisfactory grip on the object and focus on modeling the motion that realizes the task. Despite its importance, task-oriented grasping has received little attention compared to adjoining domains. The objective of this work is to bridge the gap between grasp planning and task (motion) planning, i.e., grasping objects to the end of completing a task that imposes constraints on the grip configuration.

II. TASK-ORIENTED GRASP MODEL

Our aim is to define a task-oriented grasp model, that encodes grasps whose placement on an object enable a given task. For instance, if the task is to hand over an object to an operator, the model encodes grasps that leave part of the object’s surface available for the operator to secure the object. As alluded above, the model consists of two components, a geometric model and a semantic model. The geometric model computes, from a depth image, a distribution of 6D grasp poses for which the shape of the gripper matches the shape of the underlying surface. The model relies on a dictionary of geometric object parts annotated with workable gripper poses and preshape parameters. This model builds on the work of Detry et al. [5], whereby an artificial agent learns such a dictionary from experience via kinesthetic teaching (Fig. 2). The second component is a semantic model that encodes task-compatible grasping regions. It relies on a CNN that parses a depth image into a set of task-compatible regions, building on the work of Papon et al. [15]. We built the CNN above the MultiNet architecture proposed by Teichmann et al. [21]. While we use MultiNet as our architecture, we diverge from it on the input side: rather than using RGB images as our input, we use preprocessed depth images, to facilitate generalization across object. The semantic model allows us to encode relationships such as “grasp from the handle”. The product of the geometric and semantic agents allows us to initiate manipulative tasks on previously-unseen objects by identifying grasping regions where the shape of the gripper fits the shape of the tool or object, and where the positioning of the gripper allows the robot to perform the intended task. This work advances the state-of-the art by leveraging data-driven semantic scene understanding and combining a qualitative semantic map to explicit geometric constraints, thereby providing solutions that are both contextually relevant and physically (mechani-
ally) realizable.

Previous studies of task-oriented grasping [1], [3], [8], [9], [19], [20], [23] relied on physics-based simulation [3], [17], visual features and learned statistical models [9], [8], [14], [19], or explicit semantic rules or ontologies [10], [22]. In this work, we capture task constraints with a deep convolutional neural network. Previous studies evaluated the applicability of CNNs to grasp and manipulation planning or control [12], [7], [13]. One limitation of CNNs for grasping is the difficulty of learning a structured output. The work of Dehban et al. [4] is technologically close to ours, with a denoising-autoencoder–based model of object/robot affordances.

III. TASK-ORIENTED GRASPING EXPERIMENT

To evaluate the applicability of our model, we trained models for four different tasks: transport (grasp by the handle), handover (grasp away from the handle), pour (grasp away from opening), and open (grasp away from lid). We trained the CNN on a synthetic, hand-annotated dataset. We constructed this dataset by annotating 3D object meshes with task constraints (Fig. 3, left), and generating random views of random configurations of those object synthetically (Fig. 3, right). We rendered simulated depth images using the BlenSor sensor simulation framework [6], which provides a realistic depth-camera sensor model. This process allowed us to produce a large training dataset while keeping the annotation effort within reason.

We tested our method on a robot composed of a 7-DOF articulated arm, and a three-finger gripper from Robotiq. Depth data are provided by a Kinect 1 camera that is rigidly connected to the robot base. In this experiment, we executed thirty-two grasps on novel objects that differed in size and shape from those used for training. We computed task constraints by submitting a single depth image to the CNN. We computed the grasp (hand pose and preshape) using the geometric model, restricting it to points marked as task-compatible by the CNN. We executed thirty-two tests with a single object on the table. Success was established if the constraints computed by the CNN correctly matched the task’s constraints (evaluated by inspection) and if the robot was able to transport the object to a basket situated 80cm away from the center of the workspace. Twenty-two of these grasps were successful, yielding a 69% success rate. Fig. 4 illustrates sample results.

IV. CONCLUSIONS

We discussed an original solution to task-oriented grasping, that addresses geometric and semantic planning jointly.

Our model allows the agent to grasp new objects for which the agent has no mesh model, and works on partial object images such as those captured by stock RGBD sensors. Our results show that the model is capable of transferring between objects that are globally different in shape: the geometric planner only exploits local 3D structure, and the CNN learns class traits that are not necessarily anchored in global object structure.

REFERENCES


