

Semantic and Geometric Scene Understanding for Single-view Task-oriented Grasping of Novel Objects

Renaud Detry Jeremie Papon Larry Matthies
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
Email: detry@jpl.nasa.gov

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 plan grasps 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 [13, 14, 21] determine grasping points that are suitable for picking up an object, while task models [3, 18] 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. The work discussed here is presented in greater detail in a follow-up publication by the same authors [7].

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 his own grip. 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 [7]. 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. [6], whereby an artificial agent learns



Fig. 1. Grasping for a *transport* task. Left: Depth gradient, input to the CNN. Middle: the green overlay indicates task-compatible regions encoded by the CNN (here: grasp handles only). Right: Planning and executing a 7DOF grasp (pose and preshape) within a compatible region.

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. [17]. We built the CNN above the MultiNet architecture proposed by Teichmann et al. [24]. 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 depth image gradients. Depth captures shape, and shape is better correlated to our task constraints than color. As a result, using depth facilitates generalization across objects. 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 (mechanically) realizable.

Previous studies of task-oriented grasping [1, 4, 10, 11, 22, 23, 26] relied on physics-based simulation [4, 20], visual features and learned statistical models [10, 11, 16, 22], or explicit semantic rules or ontologies [2, 12, 25]. 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 [9, 14, 15, 19]. One limitation of CNNs for grasping is the difficulty of learning a structured output. The work of Dehban et al. [5] is technologically close to ours, with a denoising-autoencoder-based model of object/robot affordances.

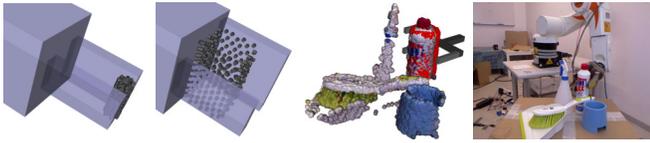


Fig. 2. Geometric model: The two leftmost images show two of the six grasp prototypes used in this work [6]. The two rightmost images illustrate the application of this model for grasping a new shape: fitting all prototypes, and executing the grasp that corresponds to the best-fitting prototype. The best-fitting prototype is shown in red in the third image.

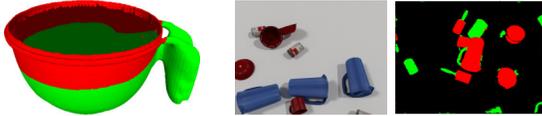


Fig. 3. Left: Task-constraint labels for a bowl and the *pour* task, where red means grasp away from the bowl opening. Middle and right: synthetically generated image and labels.

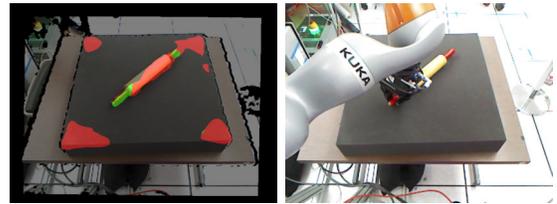
III. TASK-ORIENTED GRASPING EXPERIMENT

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 CNNs 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 [8], which provides a realistic depth-camera sensor model. This process allowed us to produce 5000 training images while keeping the annotation effort within reason. We trained one CNN for each of the four tasks listed above.

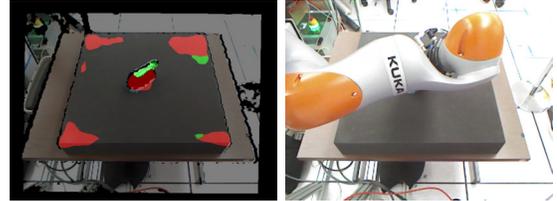
We tested our method on a robot composed of a 7-DOF articulated arm, and a three-finger gripper from Robotiq. Depth data are provided by a Kinect 1 camera that is rigidly connected to the robot base. In this experiment, we executed thirty-two grasps on novel objects that differed in size and shape from 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

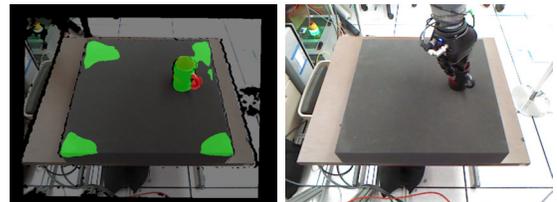
We discussed an original solution to task-oriented grasping, that addresses geometric and semantic planning jointly. Our



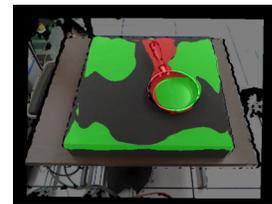
(a) *transport* task, success



(b) *transport* task, success



(c) *handover* task, success



(d) *pour* task, failure: the semantic model excluded the handle, that is the only part of the object compatible with the pouring task

Fig. 4. Task-oriented grasp examples. The first three examples are successful. The fourth example failed, for lack of similar examples in the training set (objects for the pouring task in the training set were a mug and a pitcher).

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.

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. © 2017 California Institute of Technology. Government sponsorship acknowledged.

REFERENCES

- [1] Laura Antanas, Plinio Moreno, Marion Neumann, Rui Pimentel de Figueiredo, Kristian Kersting, José Santos-Victor, and Luc De Raedt. High-level reasoning and low-level learning for grasping: A probabilistic logic pipeline. *arXiv preprint arXiv:1411.1108*, 2014.
- [2] Dmitry Berenson, Siddhartha Srinivasa, and James Kuffner. Task space regions: A framework for pose-constrained manipulation planning. *The International Journal of Robotics Research*, 30(12):1435–1460, 2011.
- [3] Sylvain Calinon, Irene Sardellitti, and Darwin G Caldwell. Learning-based control strategy for safe human-robot interaction exploiting task and robot redundancies. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.
- [4] Hao Dang and Peter K Allen. Semantic grasping: Planning robotic grasps functionally suitable for an object manipulation task. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.
- [5] Atabak Dehban, Lorenzo Jamone, Adam R Kampff, and José Santos-Victor. Denoising auto-encoders for learning of objects and tools affordances in continuous space. In *IEEE International Conference on Robotics and Automation*, 2016.
- [6] Renaud Detry, Carl Henrik Ek, Marianna Madry, and Danica Kragic. Learning a dictionary of prototypical grasp-predicting parts from grasping experience. In *IEEE International Conference on Robotics and Automation*, 2013. doi: 10.1109/ICRA.2013.6630635.
- [7] Renaud Detry, Jeremie Papon, and Larry Matthies. Task-oriented grasping with semantic and geometric scene understanding. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017. under review.
- [8] Michael Gschwandtner, Roland Kwitt, Andreas Uhl, and Wolfgang Pree. Blensor: blender sensor simulation toolbox. In *International Symposium on Visual Computing*. Springer, 2011.
- [9] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *IEEE International Conference on Robotics and Automation*, 2016.
- [10] Martin Hjelm, Renaud Detry, Carl Henrik Ek, and Danica Kragic. Representations for cross-task, cross-object grasp transfer. In *IEEE International Conference on Robotics and Automation*, 2014. doi: 10.1109/ICRA.2014.6907697.
- [11] Martin Hjelm, Carl Henrik Ek, Renaud Detry, and Danica Kragic. Learning human priors for task-constrained grasping. In *International Conference on Computer Vision Systems*. Springer, 2015. doi: 10.1007/978-3-319-20904-3_20.
- [12] Rigas Kouskouridas, Theodora Retzepi, Eleni Charalampoglou, and Antonios Gasteratos. Ontology-based 3d pose estimation for autonomous object manipulation. In *IEEE International Conference on Imaging Systems and Techniques*, 2012.
- [13] O. Kroemer, E. Ugur, E. Oztop, and J. Peters. A kernel-based approach to direct action perception. In *IEEE International Conference on Robotics and Automation*, 2012.
- [14] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.
- [15] Jeffrey Mahler, Florian T Pokorny, Brian Hou, Melrose Roderick, Michael Laskey, Mathieu Aubry, Kai Kohlhoff, Torsten Kröger, James Kuffner, and Ken Goldberg. Dexnet 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, 2016.
- [16] Nikos Mavrakis, Marek Kopicki, Rustam Stolkin, Ales Leonardis, et al. Task-relevant grasp selection: A joint solution to planning grasps and manipulative motion trajectories. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, 2016.
- [17] Jeremie Papon and Markus Schoeler. Semantic pose using deep networks trained on synthetic rgb-d. In *IEEE International Conference on Computer Vision*, 2015.
- [18] Jan Peters, Jens Kober, Katharina Mülling, Oliver Krömer, and Gerhard Neumann. Towards robot skill learning: From simple skills to table tennis. In *Machine Learning and Knowledge Discovery in Databases*, pages 627–631. Springer, 2013.
- [19] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, 2016.
- [20] Sebastian Rockel, Stefan Konecný, Sebastian Stock, Joachim Hertzberg, Federico Pecora, and Jianwei Zhang. Integrating physics-based prediction with semantic plan execution monitoring. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.
- [21] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic Grasping of Novel Objects using Vision. *International Journal of Robotics Research*, 27(2):157, 2008. URL http://ai.stanford.edu/~asaxena/learninggrasp/IJRR_saxena_et_al_roboticgraspingofnovelobjects.pdf.
- [22] Hyun Oh Song, Mario Fritz, Chunhui Gu, and Trevor Darrell. Visual grasp affordances from appearance-based cues. In *IEEE Workshop on Challenges and Opportunities in Robot Perception*, 2011.
- [23] Alexander Stoytchev. Behavior-grounded representation of tool affordances. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, 2005.
- [24] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. *arXiv preprint arXiv:1612.07695*, 2016.
- [25] Karthik Mahesh Varadarajan and Markus Vincze. Afrob:

The affordance network ontology for robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[26] Li Ying, Jiabin L. Fu, and Nancy S. Pollard. Data-

driven grasp synthesis using shape matching and task-based pruning. *IEEE Transactions on Visualization and Computer Graphics*, 13(4):732–747, 2007.