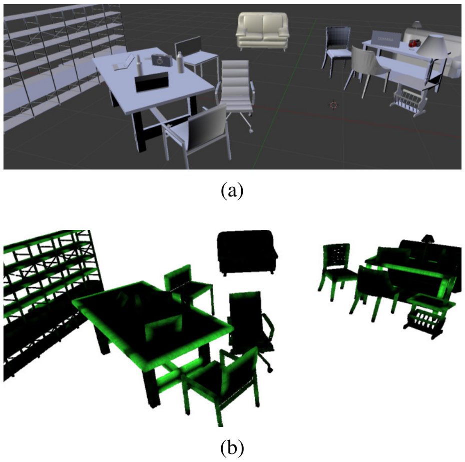
Learning to Grasp with a Deep Network for 2D Context and Geometric Prototypes for 3D Structure

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

***Abstract*— This paper discusses a hybrid approach to planning a grasp from a single view in an open-ended environment. We encode qualitative contextual grasping cues in a deep neural network, and structured gripper pose/preshape constraints in a part-based geometric model. We establish a grasp plan by combining the two models together. A key element of our approach is that instead of using the CNN to solve the complete grasping problem, we only use the CNN to compute a scalar pixel-wise grasp probability prior, which we then combine to a geometric-reasoning likelihood to establish an $SE(3)$ wrist position/orientation grasp plan. Our preliminary results indicate the feasibility of capturing scene-wide scalar graspability within a CNN.**

I. INTRODUCTION

Planning a grasp in a fully-controlled environment is a hard problem, that requires us to search through the high-dimensional space of wrist positions, orientations and finger placements, to select a hand configuration and hand/object contacts that impart a net force that counteracts gravity and other external disturbances. Grasping in an uncontrolled environment, where the shape of the object is unknown, and where we perceive the object from a single viewpoint, is even harder, primarily because of self-occlusions and cross-object occlusions: only one side of every object is visible to the camera, and the robot must devise grasps in which at least one finger is contacting a hidden surface. All grasp planners that work with partial views must, explicitly or implicitly, include means of predicting the shape of hidden surfaces. Some do this explicitly, for instance via principles of symmetry [2], [7], [17]. Others work implicitly, by fitting shape prototype (cube, cylinders, ...) to visible surfaces [3], [1], [6], [12], or with deep networks [10], [16], [5], [14], [11], [9], [15], [8]. The approach that we are currently studying is implicit. We train a CNN to map from depth images to graspability. We generate graspability labels onto full 3D synthetic scenes, and simulate a large number of partial views of those scenes in order to train the CNN on depth images that encode 2.5D data but are labeled with full-3D graspability. By contrast to approaches that rely on symmetry or on shape prototype, our approach exploits contextual data: The graspability map that we compute for a given object depends not only on points observed along the surface of the object in question, but also on neighboring objects and surfaces. By contrast to many CNN-based grasp planners, our approach only uses the CNN to compute a scene-wide scalar prior. We compute gripper parameters with a classical geometric planner. This approach allows us to compute a 7DOF grasp plan, that parametrizes the 3D position, orientation and preshape of the gripper, by contrast to CNN planners that typically compute 2D (image-space) grasping points, and use heuristics to find a 6DOF solution.


(a)

(b)

Fig. 1. Top: Synthetic training scene. Bottom: Graspability labels computed with the algorithm of Detry et al. and the two prototypes of Fig. 2; graspability is proportional to green intensity.

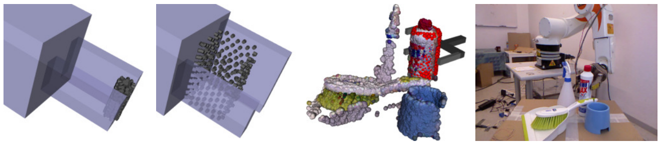


Fig. 2. Geometric model: The two leftmost images show two of the six grasp prototypes used in this work [3]. The two rightmost images illustrate the application of this model for grasping a new shape: fitting all prototypes, and executing the grasp that corresponds to the best-fitting prototype. The best-fitting prototype is shown in red in the third image.

To plan a grasp in a new scene, we first compute a graspability image with the model discussed above, then execute the part-based model of Detry et al. [3] while restricting the pose search to regions that are marked graspable by the CNN.

This paper describes (1) the generation of grasp-annotated synthetic scenes, and (2) the CNN that maps from depth images to graspability, and (3) a preliminary experiment that demonstrates scene-wide graspability.

Our current efforts are directed towards training a deep

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.

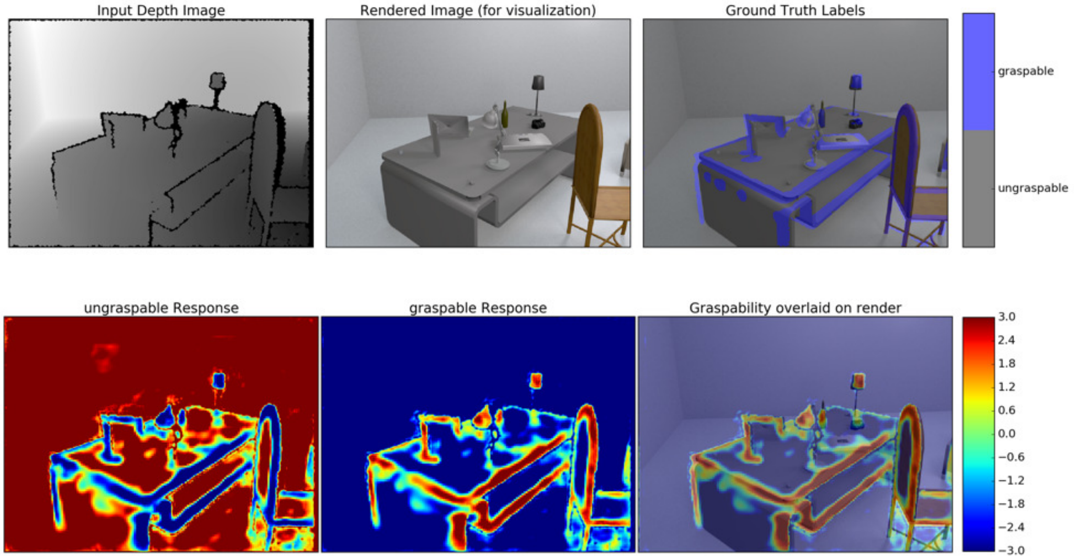


Fig. 3. Learning graspability. The first image shows an example depth image, simulated using the model of Fig. 1a. The third image shows ground truth image, simulated using the model of Fig. 1b. The bottom row shows the labels computed by the CNN for this view.

network with synthetic indoor scenes such as the one shown in Fig. 1a. For two such scenes, we have annotated each point of the scene with a graspability coefficient computed with the geometric grasp planner of Detry et al. [3]. The model of Detry et al. [3] relies on a dictionary of grasp prototypes, composed of geometric object parts annotated with a workable gripper pose and preshape parameters (Fig. 2). This model allows us to evaluate the feasibility of a given gripper pose G, by measuring the surface overlap between the scene and the prototype in pose G, and by verifying that none of the space that is spanned by the gripper collides with the scene. We define a scene point p as graspable by simulating random orientations of all prototype centered at p. Point p is graspable if there is one prototype, in one orientation, that is such that 50% of the surface of the prototype matches scene surfaces, while avoiding collisions between the scene and the gripper. Fig. 1b shows an example of a graspability map for the scene of Fig. 1a, obtained by summing graspability maps computed with the two prototypes of Fig. 2.

To train the CNN, we generated random views of the graspability point clouds using a Kinect-like camera model [4]. For each view, we generated a depth image and a graspability image. We trained the fully convolutional residual-dilated-skip CNN architecture of Papon et al. [13] on this data set, using the depth image as input and thresholded graspability image as ground truth. In this sense, the network learns to predict grasps that are compatible with the back (self-occluded) side of an object, by using cues from local context instead of multiple views. Initial results are shown in Fig. 3. The partial view shown there was held-out from the training set – that is, while the network saw this scene from other views, it never saw this viewpoint. The overall cross-validation success rate of the network is 97%. This shows that the network is learning to map from metric depth to graspability. We note that those results were obtained with a relatively limited number of training scenes. While those results are encouraging, further investigation will quantify to what extent the network learns the actual physical capabilities of the gripper – for example, by evaluating predicted graspability when surrounding objects block off potential grasps, and by excluding the background and flat surfaces from the evaluation.

II. ACKNOWLEDGEMENTS

The authors thank Tristan Thrush for his key contributions to assembling the manipulation platform, and configuring software drivers for the arm, hand and camera.

REFERENCES

[1] C. Bard and J. Troccaz. Automatic preshaping for a dextrous hand from a simple description of objects. In *IEEE International Workshop on Intelligent Robots and Systems*, pages 865–872. IEEE, 1990.

[2] J. Bohg, M. Johnson-Roberson, B. León, J. Felip, X. Gratal, N. Bergstrom, D. Kragic, and A. Morales. Mind the gap – robotic grasping under incomplete observation. In *IEEE International Conference on Robotics and Automation*, pages 686–693, 2011.

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

[4] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree. Blensor: blender sensor simulation toolbox. In *International Symposium on Visual Computing*. Springer, 2011.

[5] M. Gualtieri, A. ten Pas, K. Saenko, and R. Platt. High precision grasp pose detection in dense clutter. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, 2016.

[6] A. Herzog, P. Pastor, M. Kalakrishnan, L. Righetti, T. Asfour, and S. Schaal. Template-based learning of grasp selection. In *The PR2 Workshop (Workshop at IROS'11)*, 2011.

[7] K. Hsiao, S. Chitta, M. Ciocarlie, and E. Jones. Contact-reactive grasping of objects with partial shape information. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1228–1235, 2010.

[8] E. Johns, S. Leutenegger, and A. J. Davison. Deep learning a grasp function for grasping under gripper pose uncertainty. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, 2016.

[9] D. Kappler, J. Bohg, and S. Schaal. Leveraging big data for grasp planning. In *IEEE International Conference on Robotics and Automation*, 2015.

[10] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[11] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg. Dex-net 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.

[12] A. T. Miller, S. Knoop, H. Christensen, and P. K. Allen. Automatic grasp planning using shape primitives. In *IEEE International Conference on Robotics and Automation*, volume 2, pages 1824–1829, 2003.

[13] J. Papon and M. Schoeler. Semantic pose using deep networks trained on synthetic rgb-d. In *IEEE International Conference on Computer Vision*, 2015.

[14] L. Pinto and A. 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.

[15] J. Redmon and A. Angelova. Real-time grasp detection using convolutional neural networks. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, 2015.

[16] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic Grasping of Novel Objects using Vision. *International Journal of Robotics Research*, 27(2):157, 2008.

[17] S. Thrun and B. Wegbreit. Shape from symmetry. In *IEEE International Conference on Computer Vision*, volume 2, pages 1824–1831, 2005.