Acting and Interacting in the Real World

Jeannette Bohg, Niklas Bergström, Mårten Björkman and Danica Kragic

A characteristic of robots that distinguishes them from other intelligent agents like the chess playing *Deep Blue* [1] or any 2D image classifier [2] is its physical embodiment in the real world. A robot can act in or interact with the environment and thereby change it. To do this in any meaningful way, it needs to be able to perceive and represent the 3D structure of its surrounding. This facilitates processes like motion planning of actuators or grasping and manipulation.

Recently, through the release of devices like the *kinect* [3], sensing of dense and high quality 3D data became cheap, fast and easy. Certain constraints introduced by other 3D sensing devices are removed. However, the basic challenges of *processing* and *understanding* 3D data remain. These are (i) the high dimensionality of the data, (ii) noise, (iii) occlusions and most importantly (iv) how to get from the low-level set of 3D points to high-level semantic information.

In the following, we will outline how we deal with these challenges given the specific task of grasping and manipulation in a table top scenario.

I. Segmentation

For picking up objects from a table, segmenting them from the table plane and each other facilitates grasp planning but also helps processes like object recognition, categorization and pose estimation that attach high-level semantic information to the low-level data. In our previous work [4], we posed segmentation as an optimization problem that maintains three hypotheses: figure, ground and a supporting plane. The proposed approach is an iterative two-stage method that first performs pixel-wise labeling using a set of model parameters and then optimise's these parameters in the second stage until convergence. The model parameters are describing distributions in hue, saturation and disparity space. The approach is robust to noise and occlusions. Recently, this approach got extended to maintain not just one but several foreground hypotheses. Furthermore, we showed that this approach works well for 3D data of very different quality coming from different sensing devices (see Fig 1).

Compared to the popular plane removal approach [5], there are several advantages. First, the segmentation is not purely geometrical but uses multimodal information for optimizing the pixel labeling. Therefore even object parts close to the table will be detected as foreground. Second, the 2D ordering of the 3D data in the pixel plane is used to efficiently represent the data and solve for the neighborhood relationships. However, it needs some top-down information for initializing the foreground hypotheses. We show that this can either come from attention in which salient scene point serve as initial foreground points or from user input.

This work was supported by the EU through the project GRASP, IST-FP7-IP-215821, the Swedish Foundation for Strategic Research and the Swedish Research Council. The authors are with the Centre for Autonomous Systems and Computational Vision and Active Perception Lab, School of Computer Science, KTH in Stockholm, Sweden. bohg,nbergst,celle,danik@csc.kth.se

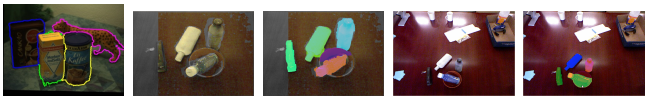

Fig. 1. Segmentation results given input data from different sensing devices. From Left to Right: (i) KTH vision system [4]. (ii) PR2 narrow stereo-image overlayed with co lour information transferred from wide-field. [6] (iii) PR2 Segmentation Result. (iv) Kinect input image mounted on the PR2. [3] (v) Kinect Segmentation Result.

II. Prediction of Occluded Object Shape

Given segmented object hypotheses from the segmentation described above, the robot now might want to know whether it has seen this or similar objects before and what their pose is. Due to occlusions, even dense 3D data will only provide a partial object reconstruction. Estimating the occluded and unknown part of an object can support 3D object or classification approaches but has also advantages for collision detection and grasp planning. Psychological studies suggest that humans are able to predict the portions of a scene that are not visible to them through *controlled scene continuation* [7]. The expected structure of unobserved object parts are governed i) visual evidence and ii) completion rules gained through prior visual experience. A very strong prior that exists in especially man-made objects is symmetry.

In our previous work [8], we show how we can use this prior to successfully predict whole object shape and evaluate their plausibility based on visibility constraints. Furthermore, we reconstruct a mesh based on which grasps can be planned in simulation using traditional grasp quality measures. However in our previous work, we have not taken the plausibility rank of each point explicitly into account. Here, we will show how this can be useful in a probabilistic grasp planning framework like the one presented in [9].

References

[1] IBM, "Deepblue," http://www.research.ibm.com/deepblue/.
[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, Jun. 2010.
[3] Microsoft, "Kinect," http://www.xbox.com/en-US/kinect.
[4] M. Björkman and D. Kragic, "Active 3D scene segmentation and detection of unknown objects," in *ICRA*, 2010.
[5] R. Rusu, A. Holzbach, R. Diankov, G. Bradski, and M. Beetz, "Perception for mobile manipulation and grasping using active stereo," in *Humanoids*, Paris, 2009.
[6] W. Garage, "Pr2 robot," http://www.willowgarage.com.
[7] T. P. Breckon and R. B. Fisher, "Amodal volume completion: 3d visual completion," *CVIU*, 2005.
[8] J. Bohg, M. Johnson-Roberson, B. León, J. Felip, X. Gratal, N. Bergström, D. Kragic, and A. Morales, "Mind the gap - robotic grasping under incomplete observation," in *ICRA*, May 2011, to appear.
[9] P. Brook, M. Ciocarlie, and K. Hsiao, "Collaborative grasp planning with multiple object representations," in *ICRA*, 2011, to appear.