Learning to Locate Objects in 3D from Image Sequences

Dimitris Margaritis, Sebastian Thrun

Problem: Robots of the future will need to interact with the world and in particular humans. They will need to navigate in a complex three-dimensional environment. They will need to locate and manipulate objects as instructed by humans, learning the tasks with minimal training. As such they need to build representations of the world around them in an accurate and efficient manner. So far a lot of research has been devoted to 2D maps, which is a reasonable approximation of the environment. However, there are situations where a 3D representation is necessary. For example, an overhanging table might cause a mobile robot to fail. Equally important is the issue of interacting with objects, where a 3D representation of the object is necessary when deciding on how to grasp and manipulate it.

Impact: A solution to the abovementioned problem will enable robots to interact and assist humans in many situations ranging from training industrial robots to instructing home robots with minimal effort. My interests lie more in the latter end of the spectrum. I believe that people tend to embrace new technologies as soon as they are made accessible and easy to use (as exemplified by the relatively recent widespread use of home computers and the Internet). My hope is that my research will contribute in making robot instruction by humans simple and natural, and their presence more widespread.

State of the Art: Existing approaches to 3D object localization in the robotics domain already exist. However, they center around modeling obstacles, not particular objects. Typically they use stereo sensors to determine obstacle depth and model occupancy using a fine-grained occupancy grid. A major hurdle in this type of research (which we should also overcome) is the enormity of the memory size needed to represent the 3D occupancy grid.

Reconstructing 3D environments has also been researched in the context of computer vision. However, the emphasis on recognition and less on learning.

Approach: I address the problem of *training* robots to recognize and locate user-specified objects. More specifically, we propose an approach that enables people to train robots by simply showing a few poses of the object. Once trained, the robot can recognize these objects and determine their location in 3D space. In contrast to existing approaches to mobile manipulation, which usually assumes that objects are located in floor or table-height, this approach does not make restrictive assumptions as to where the object is located. This poses new challenges on the ability to localize objects, as a single camera image is insufficient to determine the location of an object in 3D space.

The approach we propose uses probabilistic representations to estimate the identity and location of the target object from multiple views. It maps camera images into 2D probabilistic maps, which describe, for each pixel in the camera image, the likelihood that this pixel is part of the target object. This mapping is established by a decision tree applied to local image features, which is constructed during the training phase from labeled images. The 2D probabilistic map is then projected into the 3D work space, based on straightforward geometric considerations. Since a single camera image is insufficient to determine the location of an object in 3D, our approach integrates information from multiple images, taken from multiple viewpoints (see Fig. 1). The Bayes rule is employed to generate a consistent probabilistic 3D model of the workspace. Our approach also takes into account the uncertainty introduced by robot motion, by using a probabilistic model of robot motion. As the robot moves in the environment taking images, it gradually improves the estimation of the identity and location of an object, until it finally knows what and where the

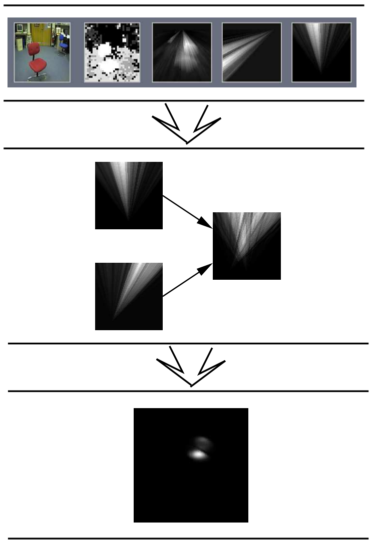

Figure 1: **(a)** Probability map that is the output of the decision tree trained to recognize a red chair. The brightest tiles in the probability map (second picture from left) correspond to probability greater than 0.9. Projection of the map in 3D are shown in the last three columns, as averages along the x, y and z (rightmost picture) axis respectively. **(b)** An illustration of the way information from images taken from two different viewpoints is integrated in the occupancy grid. Shown to the left are two single projections applied to an "empty" grid. The picture on the right shows how they are combined together. The images depict the average values of grid cell probabilities when viewed from above (i.e. averaging probability values along the z-axis). **(c)** The result after several iterations of the procedure, when viewed from the side. The two parts of the chair (back and seat) are discernible.

object is. Experimental results using a RWI B21 robot equipped with a color camera show that multi-part objects can be located robustly and with high accuracy.

Future Work: A promising extension of the current approach would be to devise methods that *actively* control the robot so as to maximize information gain. Currently, a human manually positions the robot. It would be advantageous if the robot decided its next position based on the abovementioned criterion. In the context of object localization, such a method could lead to a behavior where a robot investigates the object from multiple viewpoints, in order to estimate its location accurately.

A major hurdle of 3D object location research of this kind is the enormity of size of the grid at fine resolutions and, subsequently, the cost for accurately updating it. An interesting extension of the current approach would be to use variable-resolution representations, such as oct-trees, for representing object location. Such representations could balance the computational and memory resources, by modeling regions coarsely that are unlikely to contain a target object. If the density of target objects is low (which is usually the case), such an extension could improve the computational efficiency of the approach substantially.

Selected References:

[1] D. Margaritis and S. Thrun. Learning to locate an object in 3d space from a sequence of camera images. In *International Conference on Machine Learning*, 1998. To appear.