10.1 Introduction

When you have brought your new Apple iRobot\(^1\) home for the first time, you are faced with the challenging task of introducing the robot to its new home/workspace. Of course the robot knows about homes and typical tasks. That’s why you bought the sleek, stylish robot, in addition to the fact that it promised a simple interface. It prompts you to take it on a tour of the house, naming the rooms, pointing out the appliances, and identifying the occupants of the house.

The promise of the now discontinued Aibo whose communication and basic behaviours show that even simple visual sensors using strong features (SIFT\([17]\)) can enable visual tracking and recognition. Built in to the home robot will be the necessary concepts – tasks, objects, contexts, locations. Your home vacuum robot “knows” only about stairs, objects, infrared walls, and random search. Your iRobot knows about kitchens, doors, stairs, bedrooms, beer (you have the party version of the iRobot that can bring you beer in the entertainment room). How does the robot tie the sensory flow it receives to its plans, names, and goals in its repertoire?

This fanciful thought experiment is not so far in the future. For some applications such as assistive technologies\([21, 1]\) which operate in contexts where rich visual sensing is deployed, the range of objects may be limited to a care facility where patients’ rooms are typically constrained in their contents. Here it may be effective to learn the connection between features of the visual stream and the object in the scene, and how they influence the actions of the robot.

---

James J. Little  
Department of Computer Science, University of British Columbia, Vancouver, BC, Canada, e-mail: little@cs.ubc.ca

Tristram Southey  
Department of Computer Science, University of British Columbia, Vancouver, BC, Canada, e-mail: southeyst@cs.ubc.ca

\(^1\) Apologies to two major corporations that may be clashing over naming rights in the future.
What is needed is knowledge of the structure of the environment: the objects in the world, their functional relations, their spatial layout and their role in tasks. Visual sensing offers some solutions. There are many ways to compute features or keypoints useful for localization and detection/recognition[17]. See Mikolajczyk and Schmid[22] for a discussion of the performance of feature detectors. Many techniques used in recognition, mapping and localization begin with these interest points.

![Image](image.png)

**Fig. 10.1** SIFT features found by our stereo-equipped robot. SIFT features in the three stereo images are matched; horizontal and vertical lines indicate the horizontal and vertical disparities respectively (from [16]).

Laser and other active sensors are not only becoming more compact, less power hungry, and less expensive, but also our techniques for solving mapping and localization (SLAM) have become increasingly powerful[35]. Solutions based on passive visual sensors, both monocular[2] and stereo[27, 29, 30], are capable also of delivering both excellent localization in and high-quality geometrical descriptions of unstructured environments. Figure 10.1 shows a set of features during a robot’s tour of our lab, with their disparities. As the robot moves in the world, it aggregates local stereo information into an occupancy grid[5] where the robot plans paths and represents the structure of solids and voids. A stereo sensor delivers real-time data about the presence of obstacles and the full arsenal of probabilistic methods for localization and estimation of actions are available[29, 30]. Figure 10.2 shows the occupancy grid constructed for the maximum-likelihood sample at the end of exploration, and the landmark map constructed for the maximum-likelihood sample at the end of a traversal of the lab; error ellipses show the spatial uncertainty of the SIFT feature points (projected from 3D). Note that the stereo cameras provide a high density of image features which match between the separate camera views and as the robot moves.