The UBC Semantic Robot Vision System

Scott Helmer, David Meger, Per-Erik Forssén, Tristram Southey, Sancho McCann, Pooyan Fazli, James J. Little, David G. Lowe

Introduction

This abstract outlines the algorithms and robot hardware used in the UBC robot competing in the Semantic Robot Vision Challenge (SRVC), held at the AAAI'07 conference in Vancouver, Canada. Successfully completing the SRVC involves smooth integration of tasks such as *data acquisition*, *training*, *obstacle avoidance*, *visual search*, and *object recognition*. Given that these tasks span several research disciplines, successful integration is a formidable task. The value of working on these problems jointly is that assumptions built into an isolated method will be exposed when it is integrated, thus highlighting where further research is required. In addition, this will focus research on robots that can navigate safely and identify objects in their environment.

Our approach is decomposed into five primary modules, each of which relies on the success of other modules, avoiding some of the unrealistic assumptions that are sometimes made when the tasks are tackled independently. The five primary modules are:

1. **Web-crawling**. Searches the web for images of objects to be located and photographed by the robot.
2. **Appearance learning**. Learns appearance based models for the objects to be located and photographed.
3. **Roaming**. Controls the robot so that the robot can explore the contest area, whilst taking care not to bump into obstacles.
4. **Attention and approach**. Directs the robot to approach and capture high-quality images of visual elements in the scene that stand out in depth and colour.
5. **Recognition**. Uses the learned object models to locate examples of each object class in the set of images captured by the robot.

Hardware

The robot is an ActiveMedia PowerBot, equipped with a SICK LMS 200 planar range finder. The robot's cameras are mounted at the top of a tower on a PTU-D46-17.5 pan-tilt unit from Directed Perception. See Fig. 1.

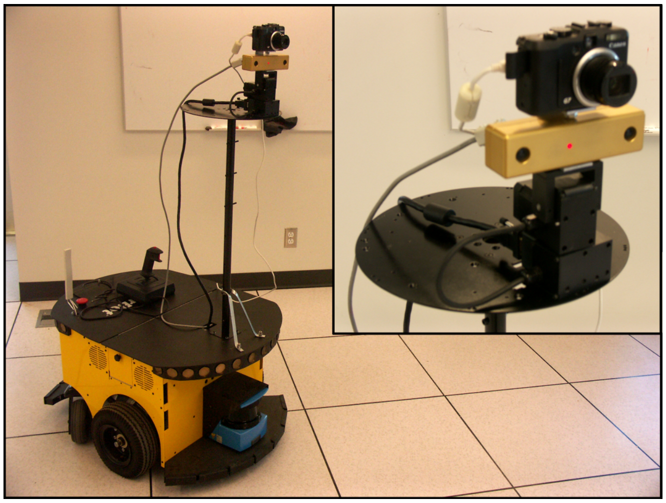

Figure 1: The UBC robot platform.

For peripheral vision, the robot has a Bumblebee colour stereo camera from PointGrey Research, with 1024×768 resolution, and a $60°$ field-of-view. For foveal vision, the robot has a Canon PowerShot G7 still image camera, with 10.0 megapixel resolution, and $6\times$ optical zoom.

Web-Crawler

Training images are acquired for the classifiers from Google's Image search website, using various techniques to reduce noise and remove mislabelled images due to incorrect labels, missing objects, or stylised appearance. In particular, linguistic techniques based upon related words in the text of the webpage containing the image are employed to identify and remove mislabelled images. Another technique uses colour histogram analysis to remove line drawings and artists renderings. In addition, a face detector (Viola & Jones 2004) is also used to identify images in which the main focus is a person, and remove them when appropriate.

It is important to note that since the robot may not be able to observe objects from all possible views during the robot exploration phase, it is important that the dataset acquired contains a variety of views if possible. Thus any technique utilized should focus on removing images that are not examples of the object rather than removing images that contain non-standard views.

Appearance Learning

Learning an object appearance model from relatively unstructured data poses significant problems, particularly when coupled with the time constraints of the competition. These challenges include mislabelled images, lack of pose information, inconsistent pose, and clutter, among others. The web-crawling phase may reduce the number of mislabelled images, but it will be of little help for many of the other problems. To deal with these issues, region based features are extracted, such as patches described by the SIFT descriptor, contours, and colour histograms. These features are used to learn an object classifier using an approach similar to (Zhang *et al.* 2007). Since the dataset collected may not be amenable for learning a general classifier of all views, image features are retained for direct matching approach, as in (Lowe 2003). This approach compensates for poor classifier performance, particularly in the case of images that are collected by the robot of non-standard object views.

Roaming

The robot is equipped with numerous sensing devices which enable safe and efficient navigation and obstacle avoidance. The SICK LMS 200 planar laser range finder allows for highly accurate detection of the position of obstacles within its 180° field of view in front of the robot. Laser range finding will be the primary approach used by the robot for navigation and obstacle avoidance. The Powerbot is also equipped with an array of sonar range finders and a stereo camera, each of which can be used to gather supplemental navigation information. Since neither of these sensing modalities are likely to be sufficiently robust in the potentially noisy and cluttered contest environment, they will not be made the focus of navigation effort.

Attention and Approach

During the robot exploration phase, the robot will attempt to collect high quality images of the potential objects in the environment. A visual attention system will be used to determine the potentially interesting locations. The system is similar to (Walther & Koch 2006), but based on appearance as well as structural information obtained using depth-from-stereo. At attended locations, *object discovery* techniques (Southey & Little 2006) will determine likely object extents. A combination of robot motion and gaze planning (Forssén 2007) is then used to centre each potential object in the foveal camera, so that a small set of high quality images can be captured of the same potential object.

Recognition

In the recognition phase, the system examines the images acquired during the robot exploration phase, and extracts the same features used during training. For each image, these features will be used to detect what objects exist in the image and establish the confidence in the detections. The local features will also be used to determine a likely extent of each object in the image. This information, along with stereo images, will then be used to place a bounding box around potential objects. If multiple images are labelled as containing a particular object, the one with the highest confidence will be selected.

Concluding remarks

The successful integration of the above mentioned modules should lead to a robot that can successfully locate and identify its target objects. Those interested in further details should make sure they get to see the robot in action at the AAAI'07 conference.

References

Forssén, P.-E. 2007. Learning saccadic gaze control via motion prediction. In *4th Canadian Conference on Computer and Robot Vision*. IEEE Computer Society.

Lowe, D. 2003. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, 91–110.

Southey, T., and Little, J. J. 2006. Object discovery through motion, appearance and shape. In *AAAI Workshop on Cognitive Robotics, Technical Report WS-06-03*. AAAI Press.

Viola, P., and Jones, M. 2004. Robust real-time face detection. *International Journal of Computer Vision* 57(2):137 – 154.

Walther, D., and Koch, C. 2006. Modeling attention to salient proto-objects. *Neural Networks* 19(9):1395–1407.

Zhang, J.; Marszalek, M.; Lazebnik, S.; and Schmid, C. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision* 73(2):213–238.