Curious George: The UBC Semantic Robot Vision System

Scott Helmer, David Meger, Per-Erik Forssén, Sancho McCann, Tristram Southey, Matthew Baumann, Kevin Lai, Bruce Dow, James J. Little, David G. Lowe
Department of Computer Science, University of British Columbia
B.C. V6T 1Z4, Canada

Abstract

This report describes the robot, Curious George, that took part in, and won, the robot league of the 2007 Semantic Robot Vision Challenge (SRVC), held at the AAAI'07 conference in Vancouver, Canada. We describe the robot hardware, the algorithms used during each of the three competition phases, as well as the results obtained by the system during the competition.

Introduction

The Semantic Robot Vision Challenge (SRVC) is a competition in which competing robots search a restricted environment and photograph objects from a target list. The challenge is divided into three phases:

- During the 30 minute *training phase*, robots are required to build visual representations for classifiers, of a previously unknown list of objects, using only images collected from the World Wide Web.
- In the 15 minute *exploration phase*, the robots examine a contest environment, which is constructed in a semi-realistic fashion, and contains the objects listed, as well as other distracting objects.
- The final, 30 minute phase, is the *classification phase*, where objects must be identified with semantic labels by matching images obtained in the first two phases.

Performance is evaluated by comparing the robotic system's classification output with a human's labeling of the objects.

Successfully completing the SRVC involves smooth integration of *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, highlighting where further research is required. In addition, the challenge will focus research on robots that can navigate safely and identify objects in their environment.

The remainder of this report is divided into five main sections. Section *Hardware* describes the robot hardware, sections *Training Phase*, *Exploration Phase*, and *Classification Phase* describe the algorithms used in the three phases of the system respectively. Finally, we present our results in the section *Contest Performance* and finish with *Concluding Remarks*.

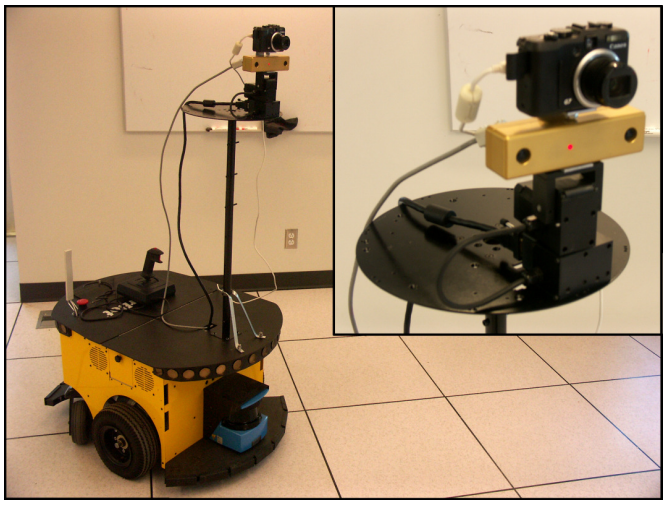

Figure 1: The UBC robot platform, Curious George.

Hardware

Hardware design is an important consideration when constructing a robot that is targeted at operating in a man-made environment. Many extant robot platforms are limited by height, navigation ability, and fixed direction sensor platforms so that interesting objects are inaccessible. For example, objects located on desks or bookshelves in an office are often too high to be seen by a robot's cameras. Our robot platform, Curious George, was designed to have roughly similar dimensions and flexibility to a human, so that relevant regions of the environment could be easily viewed and categorised. The robot is an ActiveMedia PowerBot, equipped with a SICK LMS 200 planar range finder. The robot's cameras are raised by a tower with height approximately 1.5 m. The cameras are mounted on a PTU-D46-17.5 pan-tilt unit from Directed Perception, which provides an effective 360° gaze range. See figure 1.

We employ a peripheral-foveal vision system in order

to obtain the high resolution required to recognise objects while simultaneously perceiving a large portion of the surrounding region. This choice has again been modelled after the human perceptual system, and was also inspired by design choices made in (Kragic & Björkman 2006). For peripheral vision, the robot has a Bumblebee colour stereo camera from PointGrey Research, with 1024×768 resolution, and a 60° field-of-view that provides a low resolution survey of the environment. For foveal vision, the robot has a Canon PowerShot G7 still image camera, with 10.0 megapixel resolution, and $6\times$ optical zoom that allows for high resolution imaging of tightly focused regions.

Training Phase

Web-Crawler

Classifiers are trained using images acquired from Google's Image Search. The search term we used was the text supplied in the list of desired objects. The images returned by Google were generally found to be more accurate than from other image search engines (i.e., the Google images more frequently contained the target object), but even these results were of significantly lower quality than image databases typically studied previously. In order to extract as many high quality images as possible, we decided to target product images from commercial websites. Such images are typically of relatively high resolution, have a homogenous non-distracting background, are taken with good lighting, and, if they contain the target object, show it from a view where it is highly recognisable.

Since the number of commercial websites on the Internet is so large, as was the range of objects that could be in the search list, we didn't specify which websites were likely to contain commercial images. Instead, we used a blacklist of websites that primarily contained amateur photographic images, such as Flickr. Images on these websites were frequently blurry, mislabelled, and if they contained the target object it was usually surrounded by distractor objects.

The most recognisable feature of commercial images is the presence of a homogenous, monochromatic background. Our homogenous background detector used the graph-based image segmentation technique proposed by (Felzenszwalb & Huttenlocher 2004). Their approach treats the image as a graph were each pixel is a node. The image is segmented by identifying dissimilar regions in the graph and "cutting" between these regions based on differences in the contained pixel intensity and position. Our work then analysed the size and location of these regions to identify if they were the background. Each region was tagged with two metrics: first, a ratio of the number of pixels in the region to the total image, and second, a ratio of the number of pixels in the region that lie on the image boundary to the total number of boundary pixels in the image. An homogeneous background should occupy a significant portion of the image and simultaneously occupy a significant proportion of the image boundary. These values are thus compared to a pair of thresholds to make the decision on each region. If an image segmentation contains one or more segments that exceed both of these thresholds, the image is declared to have an homogeneous background. Empirical testing showed that a pixel ratio of 20% and border ratio of 40% for a given region is sufficient to detect most homogeneous-background images. Images with a homogenous background were then labelled so they could be prioritised in the classification step.

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, we extract grey-scale regions in a similarity covariant (translation, rotation and scale covariant) frame around difference-of-Gaussian scale-space points, and describe these using the SIFT descriptor (Lowe 2003). For robustness, an object recognition system should use several kinds of features, and we initially experimented with contour features and colour histograms as well. However, in the end we abandonned these due the time constaints placed on the learning phase of the competition.

Initially, we attempted using the SIFT features to learn an object classifier using an approach similar to (Zhang *et al.* 2007). However, the large within-class variation and low number of example images collected for most of the object classes favors direct image matching similar to (Lowe 2003). Such direct image matching can be computationally intensive, so we would like to focus computation on more promising example images first. To accomplish this, we rank the training images within each object class based on their within-class similarity and between-class dissimilarity. This approach ensures that if the training data consists of multiple views or multiple modes then the best image will from each mode will be ranked near the top.

Exploration Phase

Laser-Based Mapping

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. As the robot moved through the contest environment, range scans and odometry information were used to create an occupancy-grid map (Moravec & Elfes 1985) using an estimation procedure based on a Rao-Blackwellized particle filter (Montemerlo *et al.* 2003). The occupancy-grid was subsequently used to ensure safe and efficient navigation and to enable planning through traversable regions.

Our team investigated the use of a completely visual navigation system in place of laser mapping. Since the visual appearance of a location is often much more distinctive than its geometry, visual mapping systems such as (Sim & Little 2006) offer potential for fast localisation and convergence over large areas. Under the time and hardware constraints of

the SRVC, however, our team found that the robot's cameras needed to be constantly used for object recognition, and the laser range scans were adequate for mapping.

Attention System

The attention system identifies potential objects (also known as *proto-objects* (Rensink 2000)) using the peripheral vision system, and focuses on these objects to collect detailed images using the foveal system, so that these images can be further processed for object recognition. Identifying potential objects correctly is a non-trivial problem, due to the presence of confusing backgrounds and the vast appearance and size variations amongst the items that we refer to as a objects. Our system makes use of multiple cues to solve this problem. Specifically, we obtain depth from stereo to determine structures which stand out from floor or background, and we process visual information directly with a saliency measure to detect regions with distinctive appearance. This section will describe the stereo and saliency approaches in detail, and will describe the subsequent collection of foveal images.

Stereo

The Bumblebee stereo camera is bundled with software for computing depth from stereo. We use the output disparity maps to detect obstacles and objects of interest, by detecting regions with above-floor elevations, see figure 2. This algorithm makes use of camera tilt (variable) and elevation (static) to transform the disparities to elevation values. The elevations are then thresholded at 10 cm, and the resultant binary map is cleaned up by a series of morphological operations. This helps to remove small disparity regions, which are likely to be erroneous, and also fills in small gaps in objects. The resultant *obstacle map* is used both to avoid bumping into objects and tables, and in combination with saliency to determine likely locations of objects.

Saliency

To detect potential objects we make use of the spectral residual saliency measure defined in (Hou & Zhang 2007). We extend the measure to colour in a manner similar to (Walther & Koch 2006). That is, we compute the spectral residual on three channels: intensity, red-green, and yellow-blue. The results are then combined by summing them to form a single *saliency map*. Regions of multiple sizes are then detected in the saliency map using the *Maximally Stable Extremal Region* (MSER) detector (Matas *et al.* 2002). This detector is useful since it does not enforce a partitioning of the scene. Instead, nested regions can be detected, if they are deemed to be stable. Typically, MSERs are regions that are either darker or brighter than their surroundings, but, since bright in the saliency map corresponds to high saliency, we know that only bright regions are relevant here, and consequently we only need to run half the MSER detector. Bright MSERs are shown in red and green in figure 3. Regions are required to have their smallest saliency value above a threshold proportional to the average image intensity (which is justified since spectral saliency scales linearly with intensity changes). This gives us automatic adaptation to global

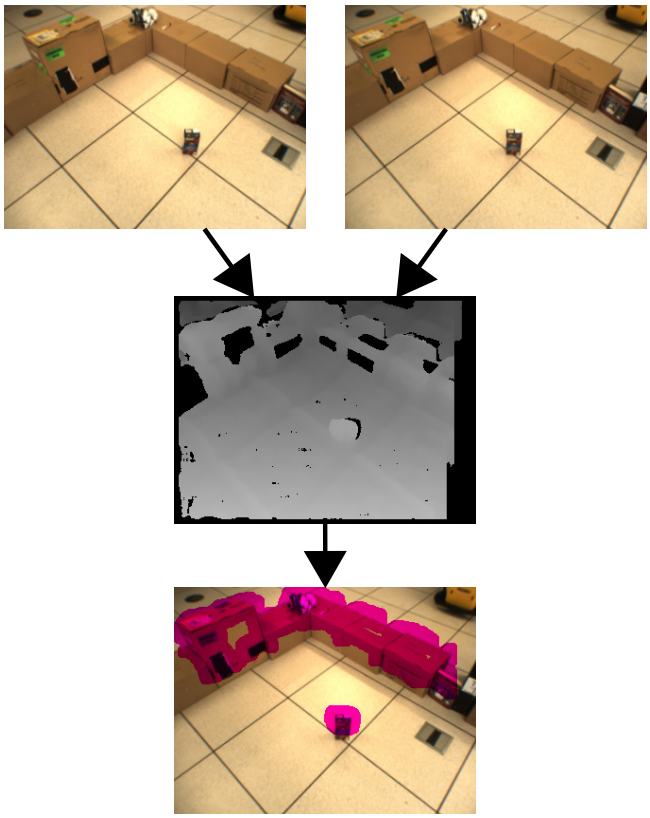

Figure 2: Stereo computation. Top to bottom: Left and right input images, disparity map, and obstacle map superimposed on right input image.

illumination and contrast changes. The regions are further required to be more than 20% smaller than the next larger nested region, to remove regions that are nearly identical. To ensure that the salient regions are not part of the floor, they are also required to intersect the obstacle map (see section *Stereo*) by 20%. Regions which pass these restrictions are shown in green in figure 3.

Compared to (Walther & Koch 2006), which can be considered state-of-the-art in saliency detection, the above described detector offers three advantages:

1. The use of spectral saliency and the MSER detector makes the algorithm an order of magnitude faster. (0.1 instead of 3.0 seconds in our system.)
2. The use of the MSER detector allows us to capture both objects and parts of objects, whenever they constitute stable configurations. This fits well with bottom-up object detection, since objects typically consist of smaller objects (object parts), and we would not want to commit to a specific scale before we have analysed the images further. The multiple sizes also map naturally to different zoom settings on the still image camera.
3. The use of an average intensity related threshold allows the number of output salient regions to adapt based on image structure. In particular, our measure can report that

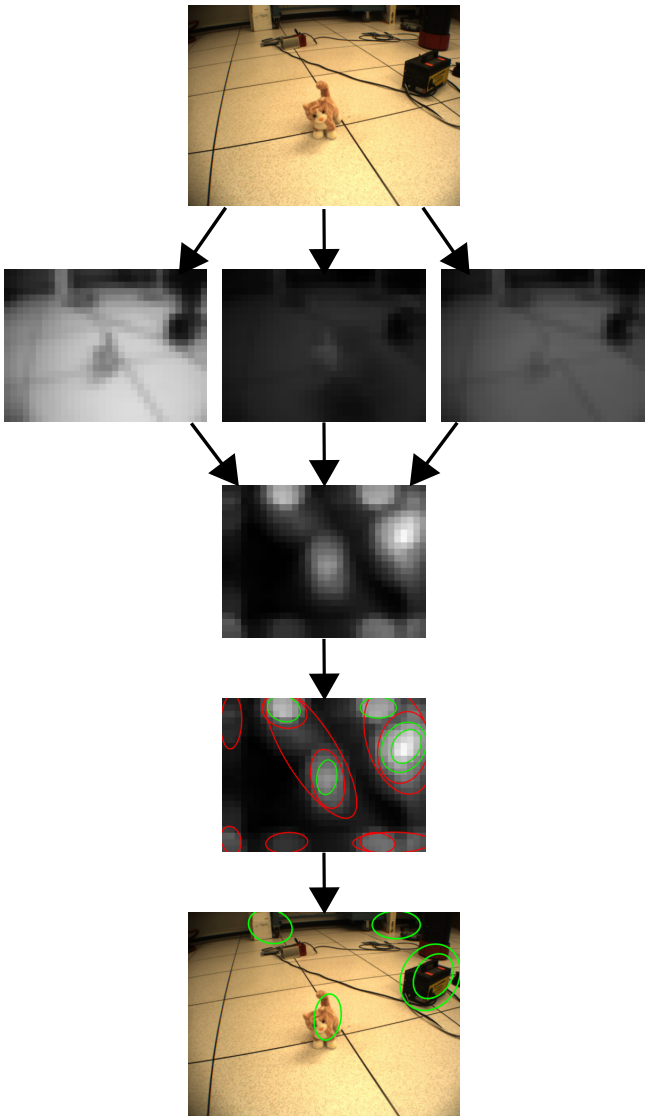

Figure 3: Saliency computation. Top to bottom: Input image, colour opponency channels (int,R-G,Y-B), spectral saliency map, detected MSERs, and MSERs superimposed on input image.

there are no salient regions within a highly uniform image, such as a picture of the floor or wall. This is in contrast to the Walther toolbox (Walther & Koch 2006), which, due to its built-in normalisation, can only order salient regions, but never decide that there is nothing interesting in the scene.

The potential objects are not necessarily what one would normally call objects — they are equally likely to be distracting background features such as intersecting lines on the floor, or box corners. The purpose of saliency is merely to restrict the total number of possible gazes to a smaller set that still contains the objects we want to find. This means that it is absolutely essential that the attended potential objects are further analysed in order to reject or verify their status as objects.

Gaze control

In order to actually centre a potential object in the still image camera, we employ the saccadic gaze control algorithm described in (Forssén 2007). This algorithm learns to centre a stereo correspondence in the stereo camera. To instead centre an object in the still image camera, we centre the stereo correspondence on the *epipoles* (the projections of camera's optical centre) of the still image camera in the stereo camera.

In order to select an appropriate zoom level, we have calibrated the scale change between the stereo camera and the still image camera for a fixed number of zoom settings. This allows us to simulate the effect of the zoom, by applying the scale change to a detected MSER. The tightest zoom at which the MSER fits entirely inside the image is chosen.

Classification Phase

In the classification phase, the system examines the images acquired during the robot exploration phase, and extracts the same types of features used during training. The system then attempts to match these test image features to those from training images from each object class, ordering matching attempts based on the image rank discussed in section *Appearance Learning*. That is, at matching attempt i, the i^{th} ranked image from all classes are attempted. In this way we are able to balance our focus amongst all classes at once. Along the way, we retain the best pair of training and test images for each object class.

The direct image matching between a test and training image consists of two parts:

- The first part compares the features from a training image to the features in the test image, and selects the top match for each training feature. To provide robustness to noise, the system normalizes the value of each match by the value of the second-best match and only retains matches which exceed a threshold. After the competition, we further improved this ratio score by replacing the second-best match in the image with the best match value in a background image set.
- The second part searches for local geometric consistency between the remaining feature matches by searching for a similarity transformation between the images to produce a score used for classification. The score is a measure of how well each training feature agrees with the transformation. More specifically, a Gaussian weighting function is applied to difference of each feature's location, scale, and a rotation from those suggested by the similarity.

The best resulting similarity transformation between the training views and test images is also used to determine a likely extent of each object in the image. This information is used to place a bounding rectangle around matched image regions.

Contest Performance

As mentioned earlier, the 2007 SRVC contest was composed of three phases: web search, exploration, and classification.

Figure 4: Recognition results recorded during the official run of the 2007 SRV Contest. (a-d) High quality views obtained by the focus of attention system, allowing for correct recognitions. (e-f) The system's best guesses at objects for which no good views were obtained – these are clearly incorrect.

The abilities of the intelligent system described in this report were demonstrated in the SRVC, where our system was the winning entry in the robot league. The list supplied to the teams in the competition contained 15 objects, and our system photographed and correctly classified 7 of these. The actual scoring was made by comparing the bounding rect-

angles output by the robots, with human drawn axis aligned bounding boxes. An overlap of the bounding boxes above 75% gave 3 points, an overlap above 50% gave 2 points, and an overlap above 25% gave 1 point.

Figure 4 demonstrates several of the objects correctly classified by our system during the final round of the contest, along with several of the misclassifications. As can be seen by the images, the contest environment was not completely realistic, but it was sufficiently complicated to present a significant challenge for current state-of-the-art recognition systems and require intelligent navigation. It was impossible to view all candidate objects from any single location, so robot motion and collection of multiple views of each object was essential. Also, many of the objects were placed in highly cluttered locations such as table tops, which would cause confusion for saliency methods that do not take into account that parts of objects may also themselves be objects. The navigation and attention systems described in sections *Attention System* and *Laser-Based Mapping* were sufficiently successful at exploring and determining the locations of interesting objects to deal with these challenges.

Concluding Remarks

In this report, we described an intelligent system capable of building a detailed semantic representation of its environment. Through careful integration of components, this system demonstrates reasonably successful and accurate object recognition in a quasi-realistic scenario. Significant work is still needed to produce a system which will operate successfully in more general environments such as homes, offices, and nursing homes, where personal companion robots are intended to operate. In such environments, challenges include the level of clutter, number of distinct objects, non-planar navigation, dynamic environments, and need to operate in real time, among many others. While the current implementation of our system is not sufficiently sophisticated to be successful in these environments, we believe there are several additional components which would bring this closer to reality.

We believe that the prospect of a useful mobile robot companion is a realistic medium term goal and that many of the components discussed in this report will be essential to the realization of such a system. It will continue to be important to evaluate approaches that extract semantic meaning from visual scenes in realistic scenarios, and also to integrate such systems with active, mobile systems, in order to achieve robustness and generality. The system described here is one step along this path.

References

Felzenszwalb, P. F., and Huttenlocher, D. P. 2004. Efficient graph-based image segmentation. *Int. J. Comput. Vision* 59(2):167–181.

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

Hou, X., and Zhang, L. 2007. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR07)*. IEEE Computer Society.

Kragic, D., and Björkman, M. 2006. Strategies for object manipulation using foveal and peripheral vision. In *IEEE International Conference on Computer Vision Systems ICVS'06*.

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

Matas, J.; Chum, O.; Urban, M.; and Pajdla, T. 2002. Robust wide baseline stereo from maximally stable extremal regions. In *13th BMVC*, 384–393.

Montemerlo, M.; Thrun, S.; Koller, D.; and Wegbreit, B. 2003. FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI)*. Acapulco, Mexico: IJCAI.

Moravec, H., and Elfes, A. 1985. High-resolution maps from wide-angle sonar. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 116–121.

Rensink, R. 2000. The dynamic representation of scenes. *Visual Cognition* 7(1/2/3):17–42.

Sim, R., and Little, J. J. 2006. Autonomous vision-based exploration and mapping using hybrid maps and Rao-Blackwellised particle filters. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*. Beijing: IEEE/RSJ.

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.