Curious George: An Attentive Semantic Robot

David Meger, Per-Erik Forssén, Kevin Lai, Scott Helmer, Sancho McCann,
Tristram Southey, Matthew Baumann, James J. Little, David G. Lowe, Bruce Dow

***Abstract*— State-of-the-art methods have recently achieved impressive performance for recognising the objects present in large databases of pre-collected images. There has been much less focus on building embodied systems that recognise objects present in the real world. This paper describes an intelligent system which attempts to perform robust object recognition in a realistic scenario, where a mobile robot moving through an environment must use the images collected from its camera directly to recognise objects. To perform successful recognition in this scenario, we have chosen a combination of techniques including a *peripheral-foveal vision system*, an *attention system* combining *bottom-up visual saliency* with *structure from stereo*, and a *localisation and mapping* technique. The result is a highly capable object recognition system which can be easily trained to locate the objects of interest in an environment, and subsequently build a spatial-semantic map of the region. This capability has been demonstrated during the Semantic Robot Vision Challenge, and is further illustrated with experimental results.**

I. Introduction

A driving motivation behind much of cognitive robotics research today is the notion of a personal robot companion that would be capable of aiding people in their daily activities. Special cases of this are systems to care for the elderly, robotic home and office assistants, and interactive robot toys for children. For each of these applications, the human and robot involved must perceive and represent the world in a similar fashion, so that they can collaborate effectively. Since humans understand the world largely based on visual information, robots targeted as personal companions should also rely on visual input. A human-like visual attention system would help a robot with both *obstacle avoidance* (e.g., noticing everyday objects it might bump into, and also spotting black-yellow warning sticker tape), and for more natural *human–robot interaction* (e.g., "Robot, fetch me my coffee mug!").

Many of the competences required for a completely visual home assistant are beyond the boundaries of current state-of-the-art research. In particular, recognising visual objects based on their semantic meaning, often referred to as object category recognition, has recently received extensive attention from computer vision researchers [1], [2], [3], [4], [5]. The focus of much of this research has been on learning appearance from large databases of static images or on indexing images from the web based on their meaning. This scenario is significantly different from the one faced by a robot in an ever-changing home environment where recognition, navigation, planning (both for robot motion and the robot's view), and interaction must all occur simultaneously. One example of a robotic system capable of object recognition in realistic settings is [6], which is similar in spirt it our system. Robotics researchers have also recently considered producing semantic maps based on the locations of objects (for example [7]), but there are still many remaining challenges related to learning visual representations of objects and integrating these semantic concepts with other robot behaviours. This paper presents an integrated solution to many of these challenges and describes a system which is capable of performing real-world object recognition in realistic scenarios.

All authors are with the Department of Computer Science, University of British Columbia, 201-2366 Main Mall, Vancouver, B.C. V6T 1Z4, Canada

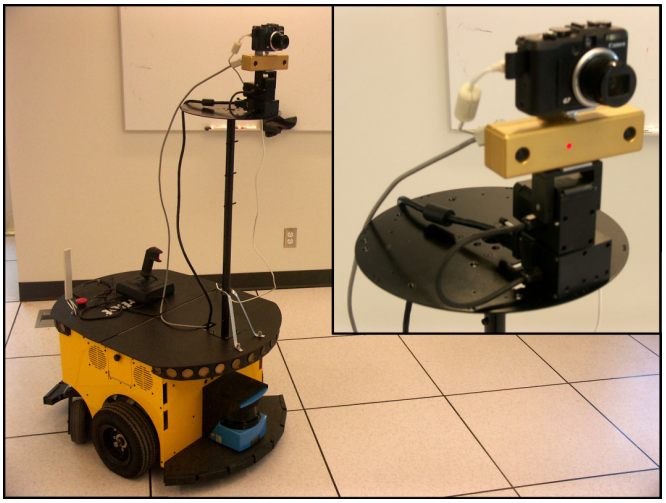


Fig. 1. The "Curious George" robot platform.

Our efforts have been motivated and directed by the Semantic Robot Vision Challenge (SRVC) [8], recently held at the Association for the Advancement of Artificial Intelligence (AAAI) conference. This challenge is divided into three phases. During the training phase, robots are required to build visual representations of a previously unknown list of objects in a short time frame, using only images collected from the World Wide Web. In the 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 phase is recognition, 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.

The physical system described in this paper finished first in the robot-league of the 2007 SRVC. Many of the design choices and physical specifications have been made

somewhat specific to that scenario, and should be changed for a more general-purpose application. Specifically, the SRVC separated the recognition problem into three phases, whereas running all components in parallel during the operational lifetime would be desirable for a robot companion. Also, the strict time requirement meant that mapping needed to occur as quickly as possible, and that highly accurate sensors were desirable. For this reason, the mapping procedure described in section IV uses a laser range finder, whereas visual mapping such as the method of Sim et al. [9] would be preferable both for reasons of cost and safety.

The focus of this paper is a description of the behaviour used during environment exploration phase of the SRVC. The goal during this phase was to collect numerous, high-quality views of each of the objects. Due to the time-constraints of the contest, these views had to be collected *without performing object recognition*, but instead by quickly identifying promising objects and regions, which we will refer to as *potential objects*. This pre-semantic identification of interesting regions was inspired by the model of human visual attention proposed by Rensink [10], where *proto-objects* are detected subconsciously in the visual periphery, and attention shifts between these to allow more detailed consideration. Our potential object detection method can be considered a simplified version of object discovery, such as the method described by Southey et al. [11], which attempts to faithfully segment meaningful objects using numerous cues. In comparison, we produce a less precise segmentation with less computation and rely on subsequent recognition to refine the result.

The remainder of this paper will provide a detailed description for each component of our method. Section II describes the hardware system. Section III describes the potential object selection method, which serves to direct the attention for our system. This is followed by a description of the navigation, mapping and coverage algorithm in Section IV and then by a brief description of the visual object recognition approach in Section V. Section VI presents results obtained during the SRVC as well as during further testing conducted in our lab, which provide validation of our approach. Finally, future work and perspectives will be discussed.

II. Hardware

Hardware design is an important consideration when constructing a robot which 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. Our 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 [12]. For peripheral vision, the robot has a Bumblebee colour stereo camera from PointGrey Research, with 1024×768 resolution, and a 60° field-of-view which 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 which allows for high resolution imaging of tightly focused regions.

III. Attention System

The attention system identifies potential objects 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.

A. 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.

B. Saliency

To detect potential objects we make use of the spectral residual saliency measure defined in [13]. We extend the measure to colour in a manner similar to [14]. 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 [15].

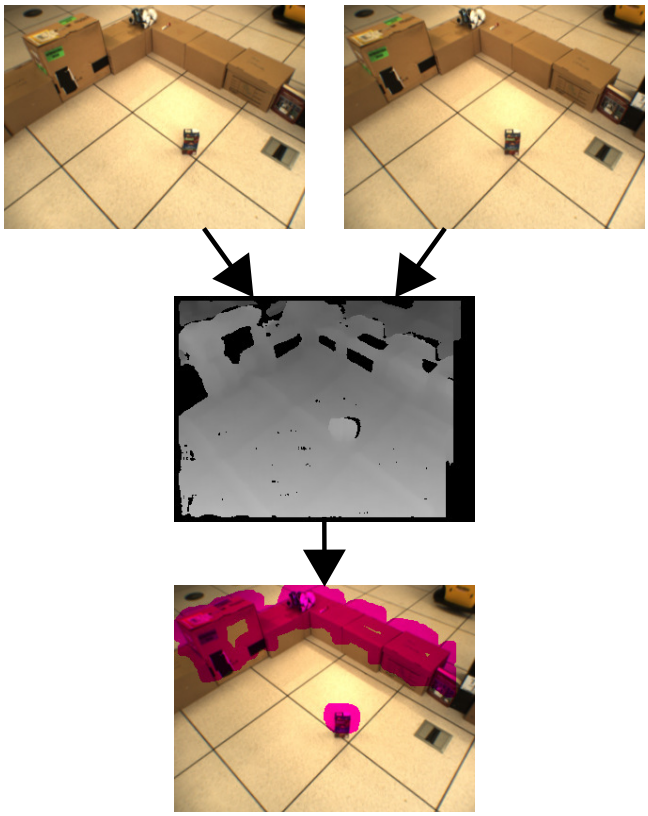


Fig. 2. Stereo computation. Top to bottom: Left and right input images, disparity map, and obstacle map superimposed on right input image.

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 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 intersect the obstacle map (see section III-A) by 20%. Regions which pass these restrictions are shown in green in figure 3.

Compared to [14], 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 us to output zero or many salient regions, depending on the image structure. This is in contrast to the Walther toolbox [14], which, due to its built-in normalisation, only can order salient regions, but never decide that there is nothing interesting in the scene.

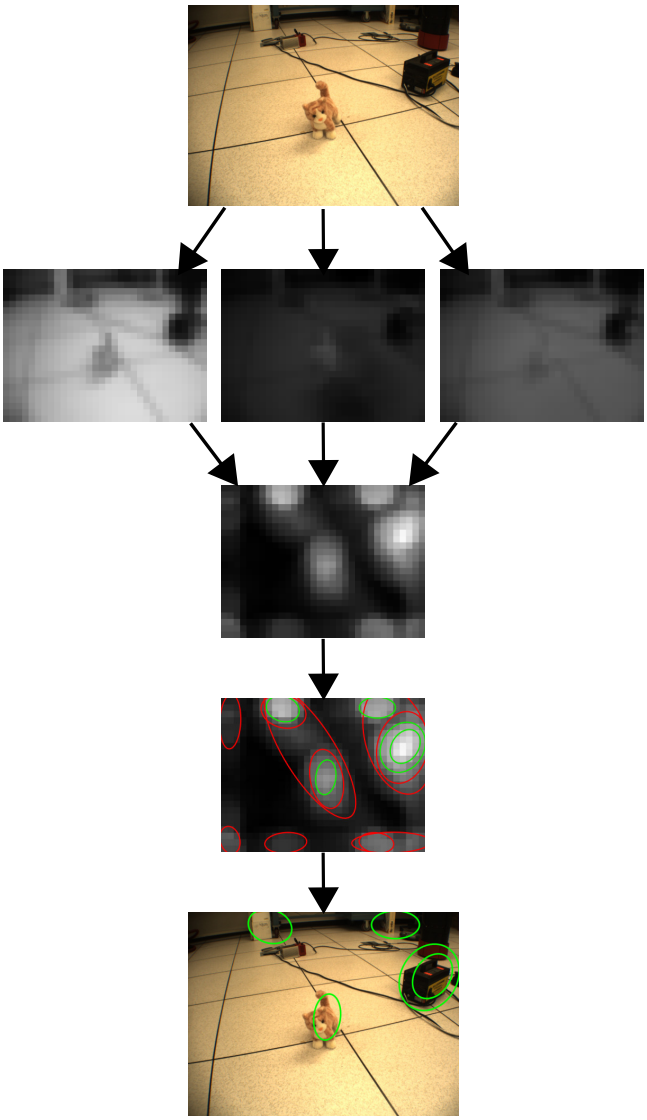


Fig. 3. Saliency computation. Top to bottom: Input image, colour opponency channels (int,R-G,Y-B), spectral saliency map, detected MSERs, and MSERs superimposed on input image.

Note that 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.

C. Gaze control

In order to actually centre a potential object in the still image camera, we employ the saccadic gaze control algorithm described in [16]. 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.

IV. Spatial Representation

An embodied recognition system must do more than simply recognising semantically meaningful objects which are directly in its field of view at a single moment in time. It must additionally move safely through its environment, record the locations of detected objects, and plan its motions to discover new objects. That is, it must be able to represent spatial-semantic information. Our system accomplishes this by: 1) building a geometric map representation of the space it has so far encountered; 2) using this map to guide further planning and exploration; 3) covering the space with the visual attention system to search for objects; 4) annotating objects in the map when they are first discovered; and 5) updating the object locations and properties over time by looking-back from different viewpoints. This section will describe each of these components in detail.

A. Geometric Mapping

Our system performs mapping with FastSLAM, a Rao-Blackwellized Particle Filter implementation [17], which builds a probabilistic occupancy grid [18] based on the laser range finder readings and the robot's odometry, and tracks the robot's position within the map. An occupancy grid is well suited to guide navigation and planning tasks for a mobile robot moving on a flat surface since it mirrors the inherently 2D nature of this environment. We have implemented a layered planning architecture where goals proposed by one of the high level behaviours described below are achieved by following a lower level path produced by A^*-search through the occupancy grid. Finally, the Vector-Field Histogram local planner described by Borenstein et al. [19] is used for local obstacle avoidance and to adapt to dynamic changes in the environment.

B. Exploration Planning

We employ the frontier based exploration technique described by Yamauchi et al. [20] to quickly cover the environment with the laser scanner and produce an initial map. As is illustrated in figure 4(a), a frontier is defined as the border between explored and unexplored space. For our system, these frontiers will be the locations just beyond the range of the laser scans, and in the laser shadows created behind objects or around corners. The frontier planning technique identifies candidate locations where laser scans would be most likely to uncover new regions to explore. First, one of these promising locations is chosen, then the robot moves to this location, and the map is updated. This process is iterated, until all regions have been explored.

C. Coverage Planning

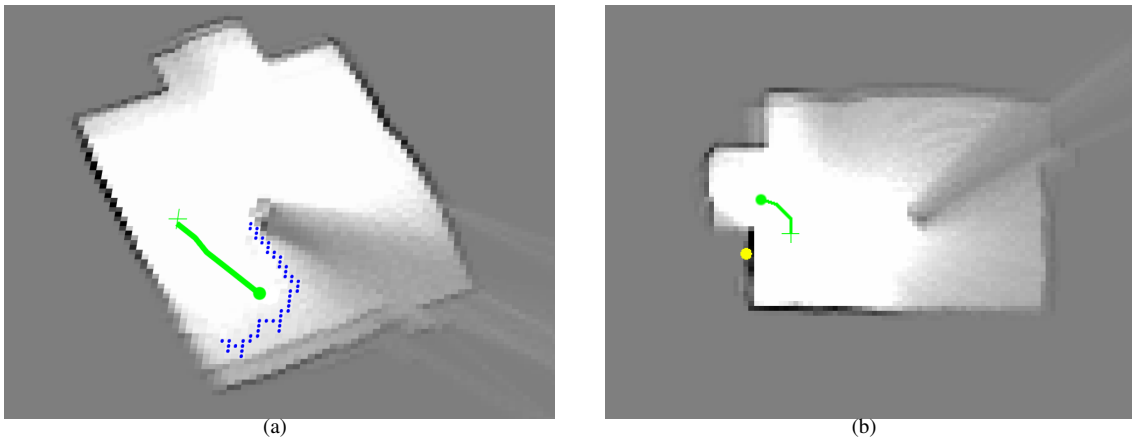
Each time a region of the environment is observed with the peripheral camera, the attention system has the opportunity to detect potential objects within that area. In order to maximise these opportunities, the camera should be pointed in directions which cover as much new territory as possible. We use an iterated greedy search based on visible area weighted by the number of previous observations, to select favourable directions. This approach causes the camera to cover the environment roughly uniformly and give an equal chance of detecting potential objects in any location.

D. Object Permanence

The set of available object poses in visual training data collected from the web is often incomplete. One tends to get the *characteristic views* [21] (e.g., a shoe is normally photographed from the side, and hardly ever from the front), rather than a uniform sampling of views. In order to recognise objects from such limited training data, we attempt to direct the foveal camera towards previously detected potential objects from various views. This behaviour ensures that the object is recognised even if the training data is biased towards one, or a small number of viewpoints. To allow collection of highly distinctive viewpoints, the previous views of an object vote for nearby angles into a histogram with values in the range $[0, 2\pi]$, and histogram bins with low scores are selected. That is, views from a completely new direction are favoured over those from similar angles. We again employ greedy search over histogram values and iterate the procedure to obtain roughly uniform coverage of viewing angles. Once a direction is selected, the hierarchical planning method moves the robot to the desired viewing position and a foveal image is collected. Figure 4(b) shows an example of a path produced during this behaviour.

V. Object Recognition

While we focus on robot exploration and image collection in this paper, it is also important to briefly discuss our method for subsequently recognising objects in the images. This section will outline our approach for training object classifiers and for evaluating these on the images collected by our robot. The current object recognition subsystem collects its training



(a) (b)

Fig. 4. Paths are planned to achieve numerous goals. (a) Path towards frontier of unexplored space (indicated by blue dots) allow for exploration. (b) A path to another clear view of an object (indicated by a yellow dot) can be used to obtain multiple views. Legend: + start of path. • end of path.

data from images returned by text-based queries to internet image search engines. These search results will likely contain numerous images containing the desired object, but will also almost certainly contain some mislabelled images, cartoon representations of the object, and extensive clutter. Such unstructured image data makes learning an object appearance model quite challenging, particularly when coupled with the time constraints of the competition. Heuristic re-ranking of the images can focus attention on more useful images. Our system uses colour histogram analysis to demote images with few colours, which are likely to be an artist's renderings and promotes uncluttered images with homogeneous backgrounds using colour image segmentation.

Unfortunately, the training data will remain challenging even after such processing, so general classifier learning approaches such as Zhang et al. [22] were found to be ineffective. It was still possible to recognise particular images using the local feature matching approach described by Lowe [23]. So, our training phase currently consists of a computation of SIFT features and their geometric relations for the training examples. Of course, this method is as likely to match incorrectly labelled images as the correct ones. So, we search for consistency in the training set by evaluating pairwise similarity. Images which mutually agree upon object appearance are ranked highly and are tried first for recognition.

Our recognition system searches for matching features between training and test images. These matching features allow us to detect objects and evaluate the confidence in the detections. That is, a large number of geometrically consistent feature matches indicates a high likelihood that the object is present in the image. If multiple images are labelled as containing a particular object, the system outputs the one with the highest confidence. The locations of matched features in the robot's images are used to determine the likely position and extent of the object and produce a bounding box. The quality of the bounding boxes (compared to ones manually drawn by the judges) were used to determine the score at the Semantic Robot Vision Challenge.

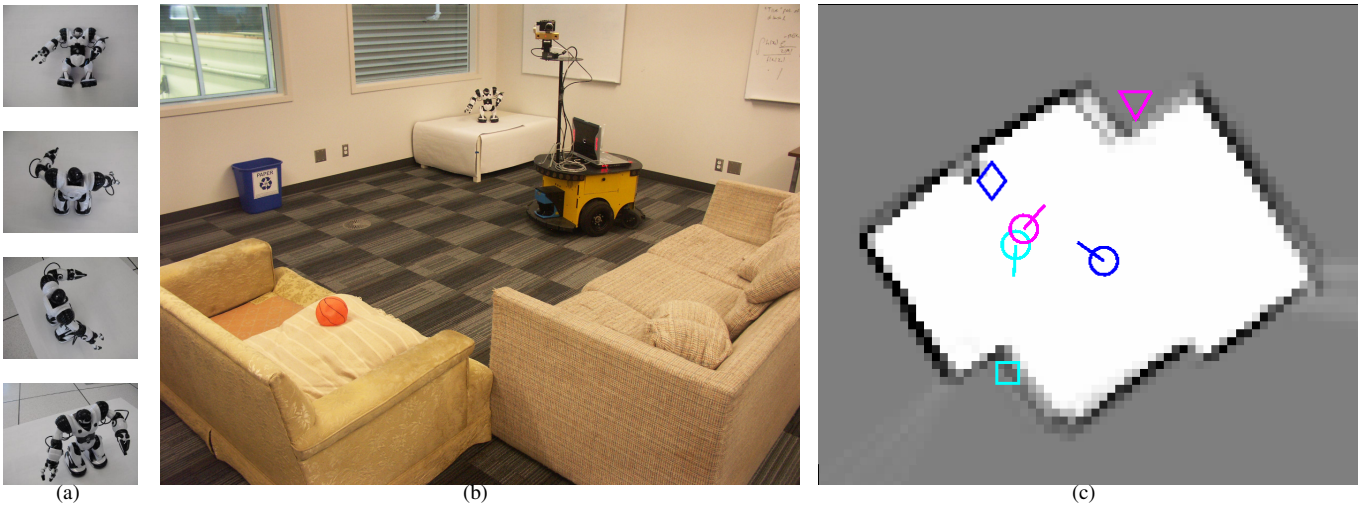
VI. Experimental Results

A. Semantic Mapping

The combination of techniques described in the previous sections endow a mobile agent with the ability to explore its environment and to recognise the objects it discovers. This behaviour can be easily extended to spatial-semantic mapping by back-projecting the recognised objects into the robot's map representation of the world. In our case, the probabilistic occupancy grid constructed from laser range scans fed through the FastSLAM algorithm can be augmented with the locations of visual objects. For example, figure 5(b) and 5(c) illustrate the locations of objects matching the labels "robosapien", "basketball", and "recycling bin". The object recognition subsystem was provided with between 2 and 4 example views of each object, see figure 5(a) for an example. Each of the objects was recognised from various locations, giving several pieces of information about the objects positions, and allowing for collection of numerous views which can be used for recognition or future matching. We envision that the types of maps illustrated here could be easily used in a human-robot interaction system where the human operator would be able to relay commands to the robot in semantically meaningful terms.

B. SRVC Contest Performance

As mentioned earlier, the 2007 SRVC contest was composed of three phases: web search, exploration, and classification. The abilities of the intelligent system described in this paper were demonstrated in the SRVC, where our system was the winning entry in the robot league. Figure 6 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. It



(a) (b) (c)

Fig. 5. Combining the spatial awareness provided by SLAM with object recognition, meaningful object labels can be assigned to locations in the map. (a) Training data for object "robosapien". (b) Overview photo of the room the robot is exploring. (c) The map with three objects, and the locations from which they were observed. Legend: ⊖ robot poses where objects were first seen, □ object "basketball", ◊ object "recycling bin", ▽ object "robosapien".

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 III and IV were sufficiently successful at exploring and determining the locations of interesting objects to deal with these challenges.

VII. Conclusions

In this paper, 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.

The object permanence ability is useful for looking back at previously seen objects while moving about, and thus to get new views of objects. These new views can be used, either to establish object identity in cases where this fails from a particular view, or to extend the object models with more training examples, and thus make them more robust.

Online and life-long learning are both promising directions towards developing a truly useful home companion which is able to interact with people and visually ground objects of common interest. A large challenge to such a system is to learn to recognise the set of objects and classes which are useful in the particular scenario of interest. This problem is confounded by the fact that the set of such objects will change, and further, that their appearances may change over time. Adaptation is clearly needed to overcome these challenges, but an even stronger comment can be made. It is, in many cases, an easier visual task to recognise objects when trained in the particular circumstances and on the particular objects which will be required during operation. Active training data acquisition as facilitated by object permanence is needed to extend the crude models obtained from the web, and to adapt to changing object appearances, (e.g., due to wear and tear).

Context is a currently untapped source of information which can be used to aid the spatial-semantic recognition task. Contextual information such as the type of room being examined would help to prioritise recognition effort towards those objects likely to be present. Spatial context allows for preferential search based on the height and position at which an object is normally found. Some interesting attempts to incorporate context using the *gist* descriptor [24] are given in [25].

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 paper 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.

Fig. 6. 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.

REFERENCES

[1] S. Helmer and D. Lowe, "Object recognition with many local features," in *In Proceedings of Generative Model Based Vision (GMBV) (workshop at CVPR)*, Washington, D.C., USA, 2004.

[2] K. Mikolajczyk, B. Leibe, and B. Schiele, "Multiple object class detection with a generative model," in *In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[3] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective," in *In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[4] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from googles image search," in *In Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV)*, 2005.

[5] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," in *In Proceedings of the Beyond Patches workshop,*

in conjunction with CVPR, 2006.

[6] S. Ekvall, P. Jensfelt, and D. Kragic, "Integrating active mobile robot object recognition and slam in natural environments," in *IEEE/RSJ International Conference on Robotics and Automation (IROS06)*. Beijing, China: IEEE, 2006.

[7] A. Ranganathan and F. Dellaert, "Semantic modeling of places using objects," in *In Proceedings of Robotics: Science and Systems (RSS)*, 2007.

[8] Website: http://www.semantic-robot-vision-challenge.org/.

[9] R. Sim and J. J. Little, "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)*, IEEE/RSJ. Beijing: IEEE Press, 2006.

[10] R. Rensink, "The dynamic representation of scenes," *Visual Cognition*, vol. 7, no. 1/2/3, pp. 17–42, 2000.

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

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

[13] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR07)*. IEEE Computer Society, June 2007.

[14] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[15] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *13th BMVC*, September 2002, pp. 384–393.

[16] P.-E. Forssén, "Learning saccadic gaze control via motion prediction," in *4th Canadian Conference on Computer and Robot Vision*. IEEE Computer Society, May 2007.

[17] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "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, 2003.

[18] H. Moravec and A. Elfes, "High-resolution maps from wide-angle sonar." in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, St. Louis, MO, USA, 1985, pp. 116–121.

[19] J. Borenstein and Y. Koren, "The vector field histogram – fast obstacle-avoidance for mobile robots," *IEEE Journal of Robotics and Automation*, vol. 7, no. 3, pp. 278–288, June 1991.

[20] B. Yamauchi, A. C. Schultz, and W. Adams, "Mobile robot exploration and map-building with continuous localization," in *IEEE Int. Conf. on Robotics and Automation*, Leuven, Belgium, 1998, pp. 2833–2839.

[21] S. E. Palmer, *Vision Science: Photons to Phenomenology*. MIT Press, 1999.

[22] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, June 2007.

[23] D. Lowe, "Distinctive image features from scale-invariant keypoints," in *International Journal of Computer Vision*, vol. 20, 2003, pp. 91–110.

[24] K. Murphy, A. Torralba, and W. Freeman, "Using the forest to see the trees: a graphical model relating features, objects and scenes." in *NIPS*, 2003.

[25] J. Vogel and K. Murphy, "A non-myopic approach to visual search," in *Proceedings of the Fourth Canadian Conference on Computer and Robot Vision CRV*, Montreal, Canada, May 2007.