Robotics and Autonomous Systems 56 (2008) 503–511

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

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

Department of Computer Science, University of British Columbia, Vancouver B.C. V6T 1Z4, Canada

ARTICLE INFO

Article history:
Available online 29 March 2008

Keywords:
Object recognition
Visual attention
Saliency
Semantic robot vision
Object permanence

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 that 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 that 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 a demonstration of semantic mapping. We also empirically verify that the attention system outperforms an undirected approach even with a significantly lower number of foveations.

1. Introduction

A driving motivation behind much of cognitive robotics research today is the notion of a personal robot companion 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. In particular, a robot with the ability to visually identify objects of interest will have much of the information necessary for successful operation. A human-like visual 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 [1–5]. 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 spirit to 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 that is capable of performing real-world object recognition in realistic scenarios.

Our efforts have been motivated and directed by the Semantic Robot Vision Challenge (SRVC) [8], held at the Association for the Advancement of Artificial Intelligence (AAAI) conference in 2007. 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

[☆] This work was in part supported by the Natural Sciences and Engineering Research Council of Canada and by the Swedish Research Council through a grant for the project *Active Exploration of Surroundings and Effectors for Vision Based Robots.*

* Corresponding author. Tel.: +1 604 827 3987; fax: +1 604 822 5485.
E-mail address: dpmeger@cs.ubc.ca (D. Meger).

0921-8890/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2008.03.008

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 4 uses a laser range finder. Visual mapping, such as the method of Sim et al. [9], would be preferable in terms of cost and safety. Since it would provide the ability to sense obstacles outside the plane of the range finder.

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 2 describes the hardware system. Section 3 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 4 and then by a brief description of the visual object recognition approach in Section 5. Section 6 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.

2. 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. 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 Fig. 1.

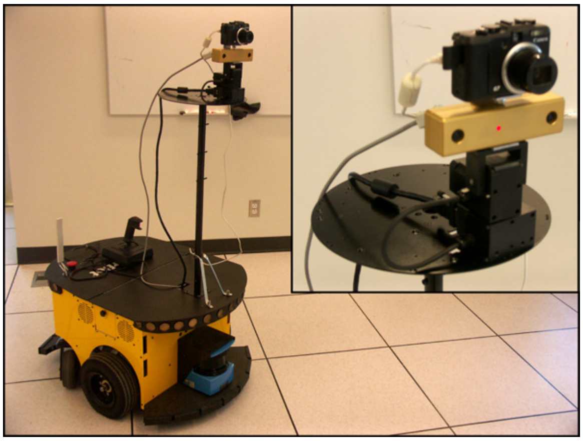

Fig. 1. The "Curious George" robot platform.

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× optical zoom which allows for high resolution imaging of tightly focused regions.

3. 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 objects. Our system makes use of multiple cues to solve this problem. Specifically, we obtain depth from stereo to determine structures that 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.

3.1. 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 Fig. 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.

3.2. 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

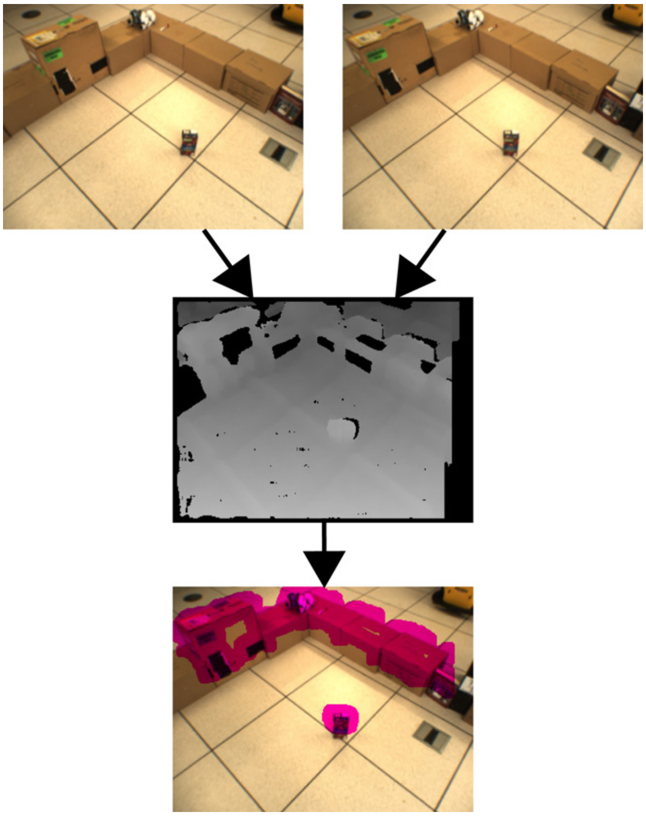

Fig. 2. Stereo computation. Top to bottom: Left and right input images, disparity map, and obstacle map superimposed on right input image.

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]. 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 Fig. 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 3.1) by 20%. Regions which pass these restrictions are shown in green in Fig. 3.

Compared to [14], which can be considered state-of-the-art in saliency detection, the above described detector offers three advantages:

(i) The use of spectral saliency and the MSER detector makes the algorithm an order of magnitude faster. (0.1 instead of 3.0 s in our system).

(ii) 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.

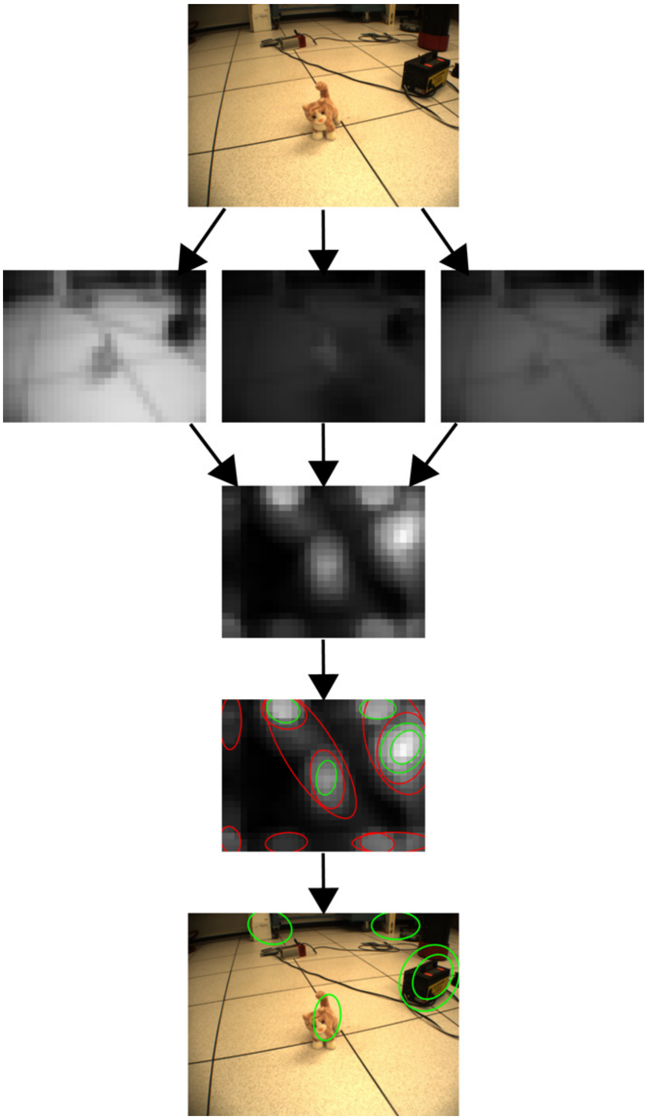

Fig. 3. Saliency computation. Top to bottom: Input image, colour opponency channels (int, R-G, Y-B), spectral saliency map, detected MSERs, and MSERs superimposed on input image. Figure best viewed in colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(iii) The use of an average intensity-related threshold allows us to adapt the number of salient regions reported, depending on the image structure. In particular, this thresholding technique will report that there are no salient regions when analysing a highly uniform image such as a wall or floor. 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.

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.

3.3. Gaze control

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

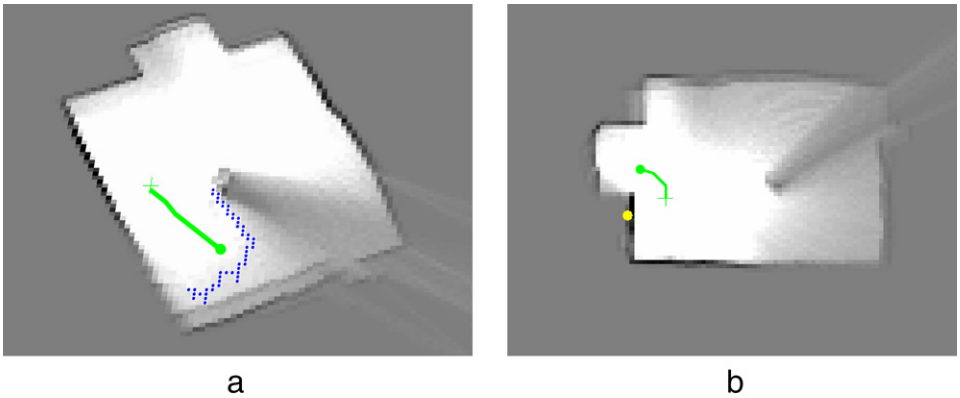

Fig. 4. Paths are planned to achieve numerous goals. (a) Path towards frontier of unexplored space (indicated by blue dots) allow for exploration. (b) A path to another clear view of an object (indicated by a yellow dot) can be used to obtain multiple views. Legend: + start of path. ● end of path. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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.

4. 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.

4.1. 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.

4.2. 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 Fig. 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.

4.3. 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 that 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.

4.4. 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 perform successful recognition using such limited training data, we attempt to collect numerous views of each potential object by looking back to the same locations as the robot moves. This requires awareness of an object's location even when it is not in the visual field, an ability known as *object permanence*. The behaviour of looking back from many views increases the likelihood that one of the collected images is taken from a similar view to that of the training data. 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

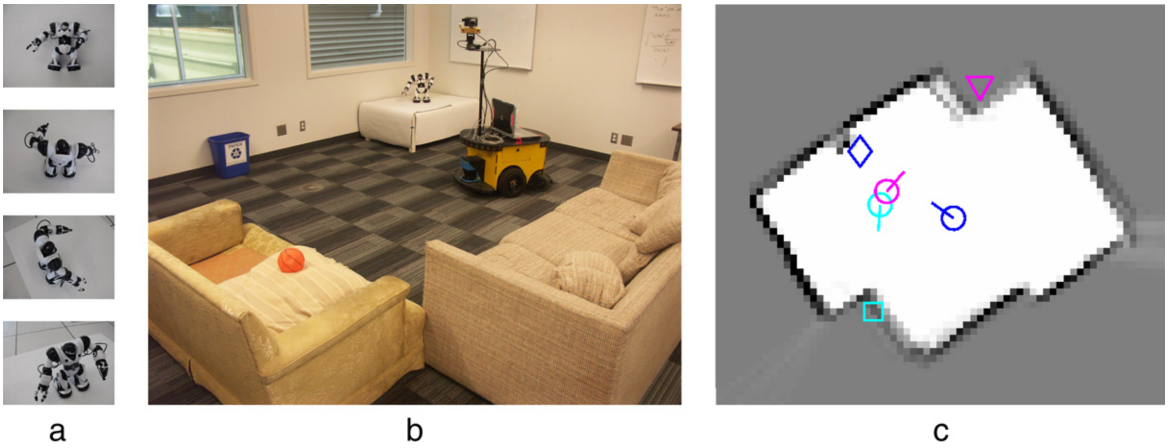

Fig. 5. Combining the spatial awareness provided by SLAM with object recognition, meaningful object labels can be assigned to locations in the map. (a) Training data for object "robosapien". (b) Overview photo of the room the robot is exploring. (c) The map with three objects, and the locations from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

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. Fig. 4(b) shows an example of a path produced during this behaviour.

5. Object recognition

While the focus of this article is robot exploration and image collection, our system also includes a 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. Our system collects its training data by submitting text-based queries to internet image search engines and storing the collections of images returned. The results of a typical query include numerous images containing the desired object, as well as some mislabelled images, cartoon representations of the object, and extensive clutter. It is a significant challenge to construct an object appearance model from such unstructured data, particularly when coupled with the time constraints of the competition. We address this challenge by *heuristic re-ranking of the images* to focus time and attention on more useful training images and also by employing a simple and robust *local feature matching* approach for recognition. Each of these components will be discussed in detail in the following sections.

5.1. Image re-ranking

We re-rank the input training imagery using a number of intuitively useful cues. First, we analyse the information within *single images*, using colour histogram analysis to demote images with few colours, which are likely to be an artist's renderings and colour image segmentation to promote uncluttered images with homogeneous backgrounds. Next, we search for consistency across *multiple images* by computing pairwise similarity between each training image and searching for large cliques in the induced graph. Since the images in such cliques strongly agree upon a common visual appearance for the category, they are unlikely to be noise, and thus are ranked highly.

5.2. Local feature matching

Even after re-ranking, numerous distracting and cluttered images remain. Recognition approaches that seek to build a generic category appearance model such as Zhang et al. [5] were found to be ineffective in the face of such data. Instead, direct image matching based on local feature patches, as described by Lowe [22], proved more successful. In particular, our training phase consists of computing SIFT features for each training example from the Web. The classification procedure extracts the same type of features from the images collected by the robot and matches features between each training and test image. Feature matches between a pair of images are verified with a geometric consistency check. That is, each match suggests a 2D *similarity transform* between images and two matches that agree upon a transformation increase the likelihood that the object is present in the image. The number of these consistent feature matches provides a measure of confidence in the detection. 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.

6. Experimental results

6.1. Semantic mapping

The combination of techniques described in the previous sections endows 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, Fig. 5(b) and (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 Fig. 5(a) for an example. Each object shown was identified by the attention system and observed from various locations, giving several pieces of information about its position, and allowing for collection of numerous views 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.

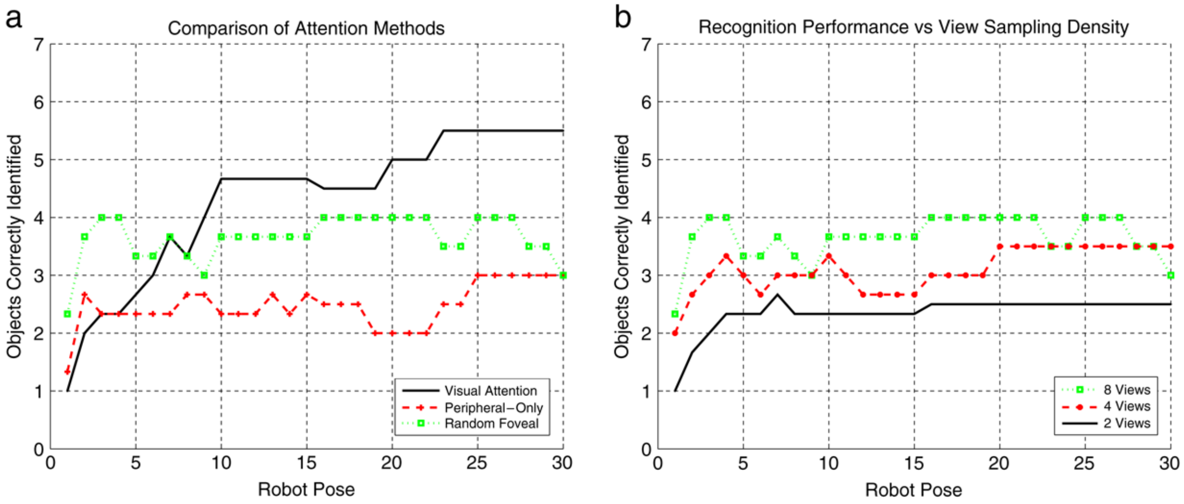

Fig. 6. Attention methods were compared in a number of ways. (a) Shows recognition results for the three classes of approaches. (b) Demonstrates that for the random view selection approach, recognition performance increases with sampling density. Each result is averaged over 3 separate runs of the robot.

6.2. Comparison of attention approaches

To validate the effectiveness of the saliency and structure based attention systems described in Section 3, we compared its performance against two other methods for selecting foveal views which will be described shortly. In order to ensure a fair comparison, the remaining components of our system were held constant. This suggests the following decomposition of our system into three parts:

(i) (Identical for each method) Robot motion to a location that allows coverage of the environment and collection of a peripheral image of a large region at low resolution.
(ii) (Three different methods compared) Selection of a number of sub-regions and collection of foveal images.
(iii) (Identical for each method) Classification of the collected foveal images by the object recognition system.

The three attention methods evaluated were the visual saliency and structure approach described in Section 3 and two comparative methods:

(i) *Peripheral view only*. This method took only one image at each robot pose, simulating the lack of a foveal vision system. The image covered the entire peripheral region at a wide zoom setting. Recognition results from this approach should be viewed as a baseline for any more selective attention system.
(ii) *Random view selection*. This method sampled from sub-regions of the peripheral view by randomly selecting n pan-tilt and zoom values from the view-cone visible in the peripheral camera, where n is a tuneable parameter. The space of possible images collected by this method is the same space in which the guided attention system searches. As such, there is some likelihood that these samples are identical to the interesting views obtained by the guided attention system, or are even better views. So, the recognition results from this approach can be used to evaluate whether or not our guided system is better than chance at selecting interesting views.

To additionally enforce fairness of comparison, we ran each of the three attention methods from identical robot locations. That is, once the robot had moved to a point, each one of the three attention methods was executed in sequence. The pan-tilt unit and zoom settings were reset to their defaults between each method, and the robot base was kept stationary during the process.

Fig. 6(a) displays the evolution of recognition performance over time for each of the 3 approaches. In all trials attempted, the approach based on visual saliency was able to recognise more objects by the end of the trial, even when the view sampling approach took many more images. In this particular plot, the random approach uses 8 images at each location, whereas the attention method uses 3 on average. We additionally varied the parameter n for the random view selection strategy, to observe its effect on recognition performance. Fig. 6(b) demonstrates that for all of 2, 4, and 8 views per pose, the collection of more data generally increases performance. However, even with twice as many view samples, the random approach does not perform as well as our guided attention measure. This is strong evidence that the visual saliency method is indeed guiding the robot to obtain promising views of objects, and that it is performing well in realistic scenarios.

6.3. 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. Fig. 7 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 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 3 and 4 were sufficiently successful at exploring and determining the locations of interesting objects to deal with these challenges.

7. Conclusions

We have 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

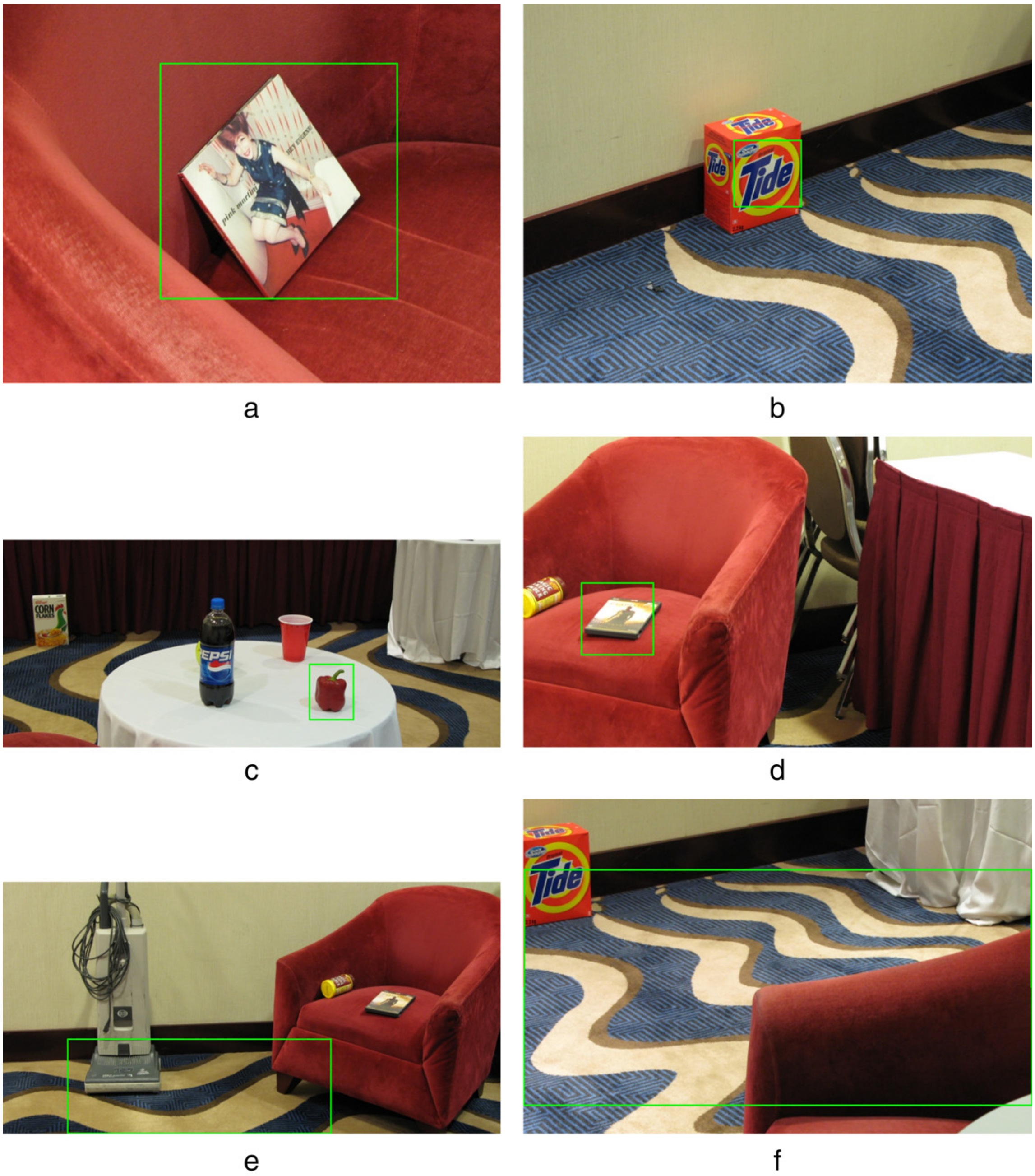

Fig. 7. Recognition results recorded during the official run of the 2007 SRV Contest. (a)–(d) High quality views obtained by the focus of attention system, allowing for correct recognitions. (e)–(f) The system's best guesses at objects for which no good views were obtained — these are clearly incorrect.

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 current object permanence ability has proven useful to establish the identity of objects which cannot be recognised from a particular view, but it could be extended to further leverage the available information. For example, the results of object recognition on the current set of collected views should impact the number of additional views and the locations from which these are taken. Once the system is nearly certain of the identity of a particular object, it would be better served by moving on to objects that it is less confident about. In the future we hope to implement an integrated decision making system which would be able to capture this intuition, perhaps by extending an approach similar to [23].

Online and life-long learning will both be essential for a truly useful home companion which is able to enter a person's dwelling and continue to succeed there for a long period. 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). We believe many components in our method are well suited to such a scenario.

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 [23].

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.

References

[1] S. Helmer, D. Lowe, Object recognition with many local features, in: Proceedings of Generative Model Based Vision (GMBV) (Workshop at CVPR), Washington, DC, USA, 2004.

[2] K. Mikolajczyk, B. Leibe, B. Schiele, Multiple object class detection with a generative model, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, CVPR, 2006, pp. 26–36.

[3] D. Hoiem, A. Efros, M. Hebert, Putting objects in perspective, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, CVPR, 2006, pp. 2137–2144.

[4] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from googles image search, in: Proceedings of the 10th IEEE International Conference on Computer Vision, ICCV, 2005, pp. 1816–1823.

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

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

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

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

[9] R. Sim, 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, IEEE Press, Beijing, 2006, pp. 2082–2089.

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

[11] T. Southey, 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, 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, L. Zhang, Saliency detection: A spectral residual approach, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR07, IEEE Computer Society, 2007.

[14] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Networks 19 (9) (2006) 1395–1407.

[15] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: 13th BMVC, 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, 2007, pp. 44–54.

[17] M. Montemerlo, S. Thrun, D. Koller, 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), IJCAI, Acapulco, Mexico, 2003, pp. 1151–1156.

[18] H. Moravec, 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, Y. Koren, The vector field histogram – fast obstacle-avoidance for mobile robots, IEEE Journal of Robotics and Automation 7 (3) (1991) 278–288.

[20] B. Yamauchi, A.C. Schultz, 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] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 20 (2003) 91–110.

[23] J. Vogel, K. Murphy, A non-myopic approach to visual search, in: Proceedings of the Fourth Canadian Conference on Computer and Robot Vision CRV, Montreal, Canada, 2007, pp. 227–234.

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

David Meger is a Computer Science Ph.D. student at the University of British Columbia. His research interests include embodied object recognition and decision making under uncertainty. He has previously completed an M.Sc. degree at McGill University on Calibration of Camera Sensor Networks by a Mobile Robot. He was a member of the robot team for UBC's 2007 SRVC team.

Per-Erik Forssén received his Master of Science in Computer Science and Engineering from Linköping University, Sweden in 1998. He obtained his Ph.D. at Linköping University in 2004, based on research on robust visual feature extraction and machine learning. He is currently a Postdoctoral fellow at UBC, working on Cognitive Vision for Robots, and Active Learning for Gaze Control.

Kevin Lai is a senior Bachelors student at the University of British Columbia (UBC). As a member of UBCs 2007 Semantic Robot Vision Challenge team, he was responsible for the agents use of sensory information in mapping, planning and directing its behaviours. Kevin intends to begin graduate studies in 2008 and continue his research on the interaction between perception and action.

Scott Helmer is a Ph.D. student at the University of British Columbia focusing on the problem of object classification using computer vision and machine learning techniques. He received a B.Sc. at the University of Toronto in 2002 and a M.Sc. at the University of British Columbia in 2004. He was one of the architects of the object recognition system of UBC's 2007 SRVC team.

Sancho McCann is a Ph.D. student at the University of British Columbia focusing on object recognition and scene understanding. He holds a Bachelor of Computer Science (Honours) degree from the University of Manitoba. He was a member of the vision sub-team of UBC's 2007 SRVC team, specializing in interest point detection and keypoint extraction.

Tristram Southey is a Ph.D. student in the Computational Intelligence Lab at the University of British Columbia. He also completed his M.Sc. at UBC focusing on multi-cue integration for object discovery. His areas of research include robotics, vision, object modelling and recognition and spatial learning and awareness. In the SRVC team he was the team leader for the Web Data Acquisition team.

Matthew Baumann is a Computer Science M.Sc. student at UBC. His current research is in vision-guided industrial robots for bin-picking and object manipulation tasks, as part of the Collaborative Advanced Robotics and Intelligent Systems (CARIS) lab, in association with the Laboratory for Computational Intelligence (LCI).

James J. Little is a professor in Computer Science at the University of British Columbia. He received his Ph.D. from the University of British Columbia in 1985. He has been a research analyst at the Harvard Laboratory for Computer Graphics and Spatial Analysis (1972–1975), a research associate at Simon Fraser University (1975–1978), and a research scientist at the MIT Artificial Intelligence Laboratory (1985–1988). His research interests include early vision, understanding image sequences, surface representation, and visually-guided mobile robotics. He was the group leader of UBC's 2007 SRVC team.

David G. Lowe is a professor of Computer Science at the University of British Columbia and a Fellow of the Canadian Institute for Advanced Research. He received his Ph.D. in computer science from Stanford University in 1984. David Lowe contributed to the object recognition approaches used in the Semantic Robot Vision Challenge.