Informed Visual Search: Combining Attention and Object Recognition

Per-Erik Forssén, David Meger, Kevin Lai, Scott Helmer, James J. Little, David G. Lowe

Abstract— This paper studies the sequential object recognition problem faced by a mobile robot searching for specific objects within a cluttered environment. In contrast to current state-of-the-art object recognition solutions which are evaluated on databases of static images, the system described in this paper employs an active strategy based on identifying potential objects using an attention mechanism and planning to obtain images of these objects from numerous viewpoints. We demonstrate the use of a bag-of-features technique for ranking potential objects, and show that this measure outperforms geometric matching for invariance across viewpoints. Our system implements *informed visual search* by prioritising map locations and re-examining promising locations first. Experimental results demonstrate that our system is a highly competent object recognition system that is capable of locating numerous challenging objects amongst distractors.

Fig. 1. The handcrafted experimental setup for accurately assessing the informed visual search technique. The robot has models of a number of objects, and searches for them in a previously unknown environment during a limited time.

I. INTRODUCTION

Finding and recognising common objects in a visual environment is easy for humans, but remains an ongoing challenge for mobile robots. Applications that require *human–robot interaction* would benefit greatly from an object finding capability, since people communicate largely using concepts that relate to visible objects. For example, the command "Robot, bring me my shoes!" is more natural for a person than guiding a robot based on a geometric map. This area has recently attracted significant interest within the robot vision field. At least two robot competitions have been undertaken: the *Semantic Robot Vision Challenge* (SRVC) [1], and *RoboCup@Home* [2]. Both competitions require participants to design a platform capable of autonomously exploring a previously unknown environment and locating specific objects, based on training data collected by the robot online or from Internet image searches. The results of these competitions have demonstrated that, while many of the component behaviours are available, the design of a robust recognition system in a quasi-realistic scenario remains a significant challenge. An earlier version of the system described in this paper placed first in the SRVC which was held in July 2007. This paper describes that base system, as well as several significant technical enhancements which together produce a highly capable object recognition platform.

Our system attempts to solve the *lost-and-found problem*, which consists of finding a set of objects present in a scene, like the one shown in figure 1. To achieve this, the system relies on many recently published robot navigation techniques to move through the environment safely and efficiently. As it moves, the robot will use its peripheral

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

cameras to collect low-resolution visual imagery and employ an attention scheme that allows it to identify interesting regions, that correspond to *potential objects* in the world. These regions will be focused on with the foveal camera, to obtain high-resolution information, which may be of sufficient quality to allow the objects contained in the region to be recognized. Due to sparse training data, and the dependence of object recognition on viewpoint, however, objects may not be recognized from the first image in which they are seen (e.g., if all training images view a shoe from the side, and it is seen from the front by the robot). Our system uses top-down information to rank the potential objects it has identified, and proceeds to actively collect images of these objects from different viewpoints, so that a useful view will eventually be obtained.

Our sequential object recognition approach is inspired by the model of the human visual attention system proposed by Rensink [3], where so-called *proto-objects* are detected sub-consciously in the visual periphery, and attention shifts between these to allow more detailed consideration. This approach is well suited to environments cluttered with distracting objects, because it allows for computation to be focused on promising regions first, while the remainder of the visual field can be ignored immediately.

Section II summaries related work in object recognition and active vision. The remainder of this paper describes our sequential object recognition system. The overall system architecture, based on peripheral foveal vision hardware with active viewpoint control, as well as a potential object selection strategy for visual attention is presented in section III. Section IV describes a method for ranking potential objects using top-down information. The active viewpoint

collection portion of our system is described in section V. Results based on a fairly general visual search task using a mobile robot are presented in section VI. Finally, conclusions and future work will be presented.

II. PREVIOUS WORK

Computer Vision researchers have made impressive progress on object recognition from still imagery during recent years [4], [5], [6], [7], but an autonomous robot performing search in the real world faces numerous additional challenges. For example, in order to get a clear view, or to disambiguate an object from distractors with similar appearance, the robot may often be required to observe the object from a particular viewpoint. Obtaining these views has been studied by the Active Vision community. For example, Laporte et al. [8] and Paletta et al. [9] describe planning algorithms for obtaining views which would maximally discriminate objects of various classes and at different orientations. These systems can be shown to improve classification performance and to allow classification with a smaller number of views when compared to an uninformed image collection approach; however, both require numerous training images annotated with orientation information (which in their setup is available from the image collection apparatus), and require significant computation which may not be suitable for real-time systems. In addition, Walther et al. [10] have studied the use of an attention mechanism as a pre-processing step for object recognition. They demonstrate that small attended regions are indeed easier to recognize than large cluttered scenes but use completely bottom-up information so that recognized regions may or may not actually be objects of interest.

Several authors have previously considered performing object recognition on robot platforms. Kragic and Björkman [11] employ a peripheral-foveal vision system that uses bottom-up visual saliency and structure from stereo in order to identify objects; however, this approach does not use top-down information which would allow it to search for specific objects. That is, their attention system is not tuned to particular objects of interest. Gould et al. [12] have built a system which does perform targeted object detection; however, their system emphasises tracking of previously recognised objects and depends on reasonably reliable recognition in its peripheral view. Ekvall et al. [13] is the most similar previous system to our own. Their system utilises top-down information to guide their visual attention system, by training a saliency-like visual map to fire at specific objects. They also utilise a foveal object recognition system based on SIFT features. Our system extends these previous approaches by integrating into an autonomous robotic system, both explicit viewpoint planning, and the use of top-down information to revisit the most promising potential objects first.

III. SYSTEM CAPABILITIES

Our robot system combines numerous components and capabilities which allow it to successfully recognise objects in realistic scenarios. This section will give a high-level overview of the system as a whole, and will provide brief descriptions of the capabilities which are not the focus of the paper (more details on these parts of the system can be found in [14], [15] and [16]). Capabilities that represent novel contributions will be described in more detail after this section.

A. Overview

Our system makes use of a planar laser rangefinder to map its environment, and to do real-time obstacle avoidance. During mapping, plans are made and executed to explore unseen regions in the map. The system has been designed to recognise objects in a potentially cluttered and challenging environment, based on training data which may contain only a single, or a sparse set of views of the objects. As such, we adopt an approach where a peripheral stereo camera is used to identify potential objects that are then imaged in greater detail by the foveal zoom camera. These images are then passed to the object recognition system, which tries to determine whether they are likely to contain any of the sought objects. Plans are regularly executed to actively move to locations where promising potential objects can be viewed from novel directions.

B. Peripheral-Foveal Camera System

We employ a peripheral-foveal vision system in order to obtain both the wide field-of-view needed for visual search and simultaneously the high resolution required to recognise objects. The cameras are mounted on a pan-tilt unit that allows them to quickly change gaze direction. The attention system identifies *potential objects* in the peripheral stereo camera, and changes the pan, tilt and zoom to focus on each of these objects and collect detailed images of them.

C. Visual Attention

The potential objects are selected in the peripheral cameras based on depth from stereo, and *spectral residual saliency* [17]. Spectral residual saliency gives an output that is similar to state of the art multi-scale saliency measures, such as [18], as implemented in [19], but is an order of magnitude more efficient (0.1 instead of 3sec run-time on our system). We compute saliency on intensity, red-green, and yellow-blue channels. The saliency maps obtained from the three channels are then summed, and regions are detected in their sum using the *Maximally Stable Extremal Region* (MSER) detector [20]. In contrast to most segmentation algorithms, MSER outputs nested regions, and these come in a wide range of sizes. The different region sizes map nicely to different zoom settings on our $6\times$ zoom foveal camera. The detected MSERs are additionally pruned by requiring that they have a depth different from that of the ground plane. We have previously shown [15] that this attention system is superior to using a wide field-of-view only, and to random foveations, even when these are on average three times as many.

D. Gaze planning

In order to 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 have asked it to 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 largest zoom for which the MSER still fits inside the image of the still image camera is chosen.

E. Geometric Mapping

As the robot moves from one location to another, it builds a geometric map of the environment that encodes navigability and facilitates trajectory planning. In addition, the positions of potential objects are recorded within the map, so that these locations can be re-examined from different viewpoints (as will be discussed later). Our geometric mapping approach is an implementation of the FastSLAM [21] algorithm. That is, the robot's position, as well as the locations of map quantities are inferred efficiently from noisy laser range data and odometry measurements by factoring the joint density into robot's path and the feature positions. Using this factorisation, map quantities are conditionally independent given the robot's path, which can be estimated by sampling.

F. Navigation

High-level system behaviours, which will be described later, provide goal locations from which the robot can collect information (e.g., laser scans of previously uncharted territory, or an image of a potential object). Two tiers of lower level navigation enable the robot to move through the environment and arrive at these locations. The first tier is A^*-search through the occupancy grid map. In order to keep the robot at a safe distance from obstacles, occupied cells are dilated and blurred to produce a soft weight function before planning paths. The second tier uses an implementation of the Vector Field Histogram algorithm described by Borenstein et al. [22] to directly control the robot's velocity and avoid dynamic and previously unrecorded obstacles.

IV. OBJECT RECOGNITION

In order to rank the potential objects detected from bottom-up visual attention, our system makes use of the top-down information gathered from training images. It is common to construct top-down saliency measures to accomplish this task, see e.g., [23], [24], [25]; however, in its most common form, top-down saliency is a way to weigh the different channels used to compute saliency. That is, it can be used to say that red is more important than blue, or that horizontal texture is more important than vertical. While being useful for explaining how people find, e.g., a green shape among a set of red shapes, this is of limited use for object recognition in general. Instead, when looking for a specific object, we note that we have a much more powerful source of top-down information than this, namely the object model itself. In this section we will describe our approach to using the local appearance part of an object model as a top-down information source to rank the potential objects.

A. Bag-of-features and Geometric matching

Most of the successful object recognition systems in use today employ local invariant features, such as SIFT [5] and MSER [7]. Depending on the object in question, there are two main categories of object recognition systems that use local features. One category is known as bag-of-features, see e.g., [26], [27]. Here local features are matched between an image and memory. To determine how good a match is, the count of feature matches relative to the total number of features is used. That is, all geometric structure between the features is disregarded, except the fact that they exist in the same image; hence the name bag-of-features. This class of methods is successful at recognising objects from large viewpoint changes, and under moderate shape deformations. Consequently, it is popular in object class recognition problems.

In our implementation of the bag-of-features approach, SIFT features [5] from a training image are first matched to the test image in a nearest neighbour approach. We then prune these matches by thresholding the ratio of the closest distance to a feature in the training set, and the closest distance to a feature in a background dataset, as suggested by Lowe [5]. Features with a ratio above 0.8 are then discarded. The *bag-of-features score* is now computed as the number of remaining features normalised by the number of features in the training image.

The methods that do make use of geometric structure, e.g., in the form of a 2D *similarity transform* [28], [29] require more training data to work well, in order to capture appearance changes due to view changes or deformations. They are, on the other hand, more reliable once a match has been found, for the simple reason that they make use of more information.

For the *geometric score*, we take the accepted matches from the bag-of-features matching and find a similarity transformation that best aligns the two images based on these matches. The score is a measure of how much each training feature agrees with the transformation. More specifically, each SIFT feature has a location in the image, a scale, and a rotation. How much a training feature agrees with a transformation is a Gaussian weighted function of how far away the test feature is in location, scale and rotation once the similarity transformation has been applied.

B. View invariance and ranking experiments

Under object rotation, the features on an object will move about in the image according to their 3D location relative to the centre of rotation, and not according to a 2D similarity transform. It would thus be reasonable to assume that as

the object is rotated, bag-of-features would exhibit a more gradual drop of the match score than methods that use geometric structure. To test this, we collected 64 views of an object at constant distance, rotated between $0°$ and $360°$, see figure 2. We then computed match scores for five test images, one containing the object, and four containing other objects, see figure 3. This was done for all the 64 training views, to produce the curves shown in figure 4. As can be seen, the drop in the score as one moves away from the peak is quite similar for the two measures. Note however that the bag-of-features score does have one significant advantage: it produces a higher signal to noise ratio in regions between the peaks.
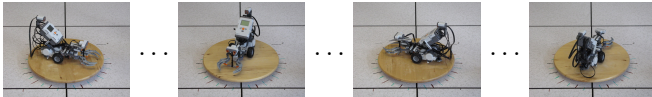
Fig. 2. Training images. 64 different views of a LEGO Mindstorms model.

Fig. 3. Test images. The first image is an inlier (the Mindstorms model), the rest are outliers (a stapler, a toy cat, an Aibo, a telephone).
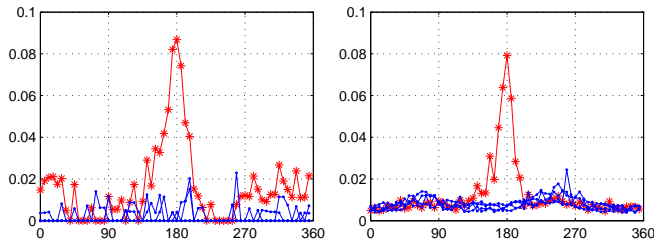

Fig. 4. Bag-of-features and geometrically constrained matching scores, as function of viewing angle (64 different angles). Left: Bag-of-features score, Right: Geometric score. Red ∗ curves are match scores for the test image containing the correct object, blue · curves are scores for other test images, see figure 3.

The above result is in agreement with recent results by Perona and Moreels [30] for feature matching on 3D objects. They found that for non-planar objects, it is the features themselves that produce the rapid drop of the score, since even for affine-invariant features, the range of rotations for which a feature can be recognised is typically only $\pm 15°$ [30].

In the following section, we will describe a method for selecting potential objects based on their rank-order. To evaluate the ability of various methods to select objects correctly with this strategy, we have designed a second test. Instead of using all 64 training views, we now pick just four views consecutive in angle. For testing, we use 10 object categories, with 10 views of each. For each test view, the match scores are computed and sorted in descending order. The result of this is shown in figure 5. As can be seen in this

plot, the bag-of-features score is slightly better at moving the inliers to the front of the list.
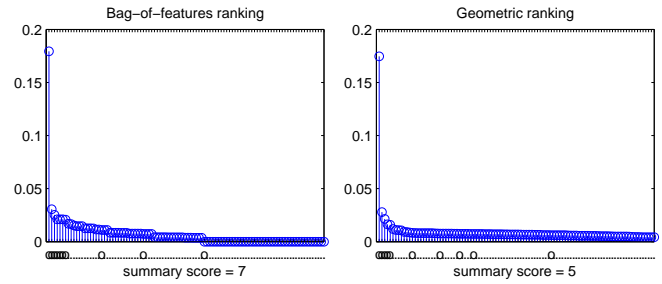

Fig. 5. Ranking performance for bag-of-features and geometric scores. Left: Bag-of-features rankings, Right: Geometric score rankings. Both use 4 equidistantly spaced training views. Circles below the plot correspond to inliers, and dots correspond to outliers.

As a criterion for evaluating the score we will use the number of inliers that are at the top of the list because those are the ones that would be chosen for closer study by a robot. For the example in figure 5, this score would be 7 for the bag-of-features ranking, and 5 for the geometric ranking. If we average this score over all consecutive groups of 4 training views, we get 4.02 and 2.73 for the two scores respectively. There is naturally some variation in scores, since some views contain more features than others, and are thus better at characterising an object. Such views are known as *characteristic views* [31]. The highest scores for the two methods are 8 and 6 respectively.

C. Using both bag-of-features and geometric approaches

Motivated by the results in the previous section, we will make use of a bag-of-features score computed on SIFT features [5] for deciding which of the observed potential objects are currently the most promising. Since a high score from the geometric matching method is in general more reliable, we will use it as a criterion for when an object ceases to be seen as a potential object, and becomes a trusted recognition.

V. INFORMED VISUAL SEARCH

Based on the system architecture and analysis described in the previous sections, we now present a sequential object recognition strategy which identifies a number of potential objects in an environment and explicitly collects images from numerous viewpoints for those objects which appear promising based on the bag-of-features ranking. Our approach is a combination of three behaviours: (1) an *exploration behaviour* that moves the robot toward unexplored regions of the environment in order to sweep out new areas with its range sensor; (2) a *coverage behaviour* that explores the environment with the peripheral camera, applies visual attention, and identifies potential objects; and (3) a *viewpoint selection behaviour*, which acquires additional images of promising objects from novel perspectives. Each of these three behaviours will be described in detail in the following sections.
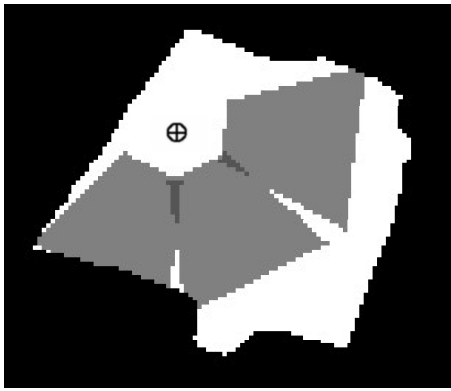
Fig. 6. An illustration of the coverage map which is used to record regions of the environment which have already been observed with the peripheral cameras. The interior of the black region represents the reachable map area, and gray shading indicates a camera field of view.

A. Exploration Behaviour

Initially, the robot's goal is to produce a geometric map of the nearby traversable regions. This is accomplished using a Frontier-based exploration approach, as described by Yamauchi et al. [32].

B. Potential Object Detection

Once a geometric map is available, the robot attempts to locate potential objects, and associate these with regions of the map. Potential objects are discovered by applying the visual attention algorithm previously described in section III-C; however, in order to be detected, an object must be located in the field of view of the camera, and at a suitable scale. So, it is important to make sure the environment has been covered by the camera. Figure 6 provides a representation of camera coverage that was achieved during a typical run of the system. Once the environment has been fully covered, numerous potential objects will likely have been identified, and the system can proceed to perform recognition to identify the true semantic labels for each of these (if any).

C. Multiple Viewpoint Collection Behaviour

Many sources of training imagery, such as the internet image searches used in the SRVC [1], contain only sparse representative views of an object. Most such object views can only be recognised from a narrow slice of viewing directions. In order to successfully recognise objects in the face of this challenge, our system attempts to obtain images from a wide variety of viewpoints for the most promising N potential objects from each class, according to the bag-of-features score, where N is a system parameter. This behaviour is achieved by storing the geometric locations, visual appearance and previous viewing angles of each potential object and commanding the robot to locations from which it can obtain novel views of as many of those objects as possible. This section will describe the implementation of this behaviour in detail.

In order to keep the geometric locations of potential objects consistent even in the case of small drift in mapping,

object locations are recorded relative to the robot's pose at a given instant. This pose is kept up-to-date by smoothing the trajectory periodically, which keeps the pairwise error between any two poses small. This ensures that the information is as accurate as possible. Note that object position information is stored as an annotation to the map, rather than as a position measurement for the SLAM algorithm. This is because the accuracy of potential object locations is orders of magnitude worse than the accuracy of the laser rangefinder, and so it would provide no gain in mapping accuracy. The system stores views from which it has seen a potential object in an angular histogram. Currently, we discretise horizontal viewing angle into 36 histogram bins, so that each bin spans $10°$. Each time a new view is obtained, the corresponding bin is incremented.

New potential objects are identified each time the visual attention algorithm is evaluated on a peripheral image. This means that, especially in a cluttered environment, the list of potential objects may grow very large, and a method is needed for selecting only those which are truly likely to be objects of interest. The bag-of-features ranking algorithm is an effective means of performing this selection, due to its ability to identify images of an object category over distracting objects and background, as was demonstrated in the previous section of this paper. Our goal is to ensure that the list of objects considered by the robot at any instant remains bounded, so we choose the N images with the highest bag-of-features score for each object category. A more adaptive strategy which selects the potential objects based on the numerical values of the scores is an area of potential future improvement.

We attempt to select a goal location for the robot which will provide a novel view for as many of the promising potential objects as possible. Each object votes for cells in the occupancy grid map based on the following conditions: (1) the robot will be physically able to collect an image of the object from that location; (2) that the cell is reachable by the robot, this implies that obstacles always have a value of 0; and (3) the view has not been previously collected. The third criterion is computed from the angular histograms by applying a soft Gaussian blur weight function over a range of cells around previously collected views. An example of the combined scoring function produced by two objects which have each been seen from a single view is shown in figure 7. In this case, the scores are highest in locations where both objects can be re-observed from roughly the opposite side, as the scores are combined in these regions.

Once the grid of novel view values has been constructed, it can be used in a number of ways. For example, one could normalise the values and draw random locations weighted by the values, or simply choose the cell with maximum score. One of the goals of our system is to recognise objects quickly, and therefore we make use of a `value/cost`-type function, where the value is the novel view value for a cell. The cost should ideally be the time required to reach the cell plus the estimated time the robot will spend at the new location. In practice we use a cost proportional to the length
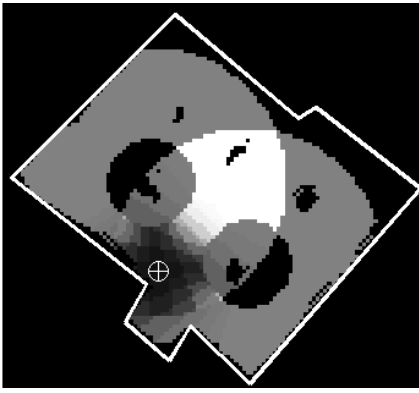
Fig. 7. An example novel view scoring function for two objects with one view each. The robot's location is shown with a ⊕ and the borders of the map are represented with a white line. Inside the map, black represents zero scores, which result from uninteresting or geometrically impossible views, or the presence of obstacles in the map. Lighter colours represent more favourable locations.

of the path from the robot's current location to the cell, plus an offset. The planner then greedily selects the cell with the greatest score.

D. Decision Making Strategy

During the visual search task, the robot must periodically decide which objects to look back at, and from what location. At some point it should also decide that it has collected sufficient information. Sequential decision making is a well studied problem and principled methods exist for deciding when to stop collecting information about the label of an object, see for example [33], [34], [35]. In our target scenario, it is difficult to construct such strong probabilistic models accurately, so we rely on several task-dependent heuristic decision-making policies. For example, the coverage behaviour is iterated until a sufficient ratio of the reachable space has been observed. Similarly, when solving the lost-and-found problem, we continue to look back at the top N promising objects for each class until one of the images scores above a threshold on the geometric matching test. This is considered a strong match, and the system moves on to search for objects of other categories. For data collection tasks, the stopping criteria can be based on sufficiently dense coverage of the viewing direction histogram.

VI. EXPERIMENTS

The system has been evaluated experimentally in several different scenarios in order to demonstrate its capacity as a general object recognition system, and to analyse the behaviour of the informed visual search procedure described in this paper. The SRVC contest provided a carefully refereed, public, and challenging scenario, and will be described in section VI-A. A slightly more engineered setup was constructed in our laboratory to enable collection of additional ground truth information, and to carefully examine the effects of the informed visual search behaviours. These experiments will be described in section VI-B. We have also recently [15]

published an evaluation of the attention system described in section III-C.

A. SRVC Performance

The Semantic Robot Vision Challenge was held at the Association for the Advancement of Artificial Intelligence (AAAI) in July 2007. At the start of the competition, robots were given a text file containing a list of object category names. The robots were allowed to connect to the Internet in order to gather training data from image searches. Using the images gathered, the robots were required to autonomously search for the objects inside a contest environment which contained the objects from the list, as well as numerous distracting objects, in addition to tables and chairs. Finally, the robots reported their results as a set of bounding-box-annotated images and the results were compared to a human's interpretation of the location of the objects within the images.

As with any task requiring object recognition in cluttered natural scenes, confounding factors such as figure/ground segmentation, viewpoint, navigation, and lighting had to be dealt with. Unlike many scenarios, this task was made significantly more difficult due to the presence of mislabelled images, cartoon depictions, and other distracting factors which are present in the imagery that can be obtained from the Internet.

An earlier version of the system described in this paper competed in the SRVC [1] and successfully recognised 7 out of 15 objects within this challenging scenario. (Compare this to 2 and 3 objects for the second and third ranked teams.) For example, our system correctly recognised the objects "Tide box", "red pepper", and "Gladiator DVD" during the official contest run.
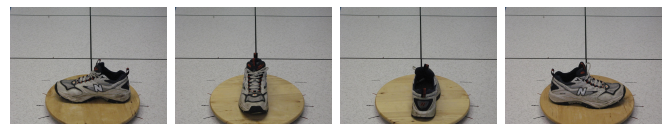

Fig. 8. Example of training data for robot experiment. The four views that were used of the object 'shoe'.

B. Informed Visual Search

In order to evaluate the performance of the informed visual search technique, we constructed a testing area containing numerous evenly spaced objects in an enclosed area in our lab. Figure 1 gives an overview of this experimental setup. We chose to space the objects evenly so that we could clearly identify which objects the robot was attempting to re-examine during the multiple viewpoint collection behaviour, and provide accurate ground-truth.

The system was provided with 4 uniformly-spaced training views of each object, so that the viewpoint invariance of the potential object ranking algorithm could be evaluated. That is, when an object was observed from an intermediate viewpoint (up to $45°$ from the closest training view) it might still rank highly and be re-examined. Eight object categories were specified for search: 'phone', 'mindstorm',

'book', 'aibo', 'keg', 'toy cat', 'shoe', and 'robosapien'. See figure 8 for several examples of the training data used. We had the robot execute informed visual search, with $N = 2$ as the limit on potential objects for each category. We manually examined the resulting images and recorded their correct labels for comparison. We also recorded the robot's state (i.e., the instantaneous list of best potential objects). This allowed us to determine whether or not the robot had decided to look back at the correct potential object for each class at discrete points during the trial.

Figure 9 summarises the results of one trial of this experiment. The horizontal axis of this plot represents the image sequence number. A solid green bar displayed in a row indicates that the top-ranked potential object was in fact a view of the correct category for the time indicated. Note that we should not expect the system to be correct for all timesteps, because this would imply that the system has achieved perfect recognition even from poor viewpoints. Rather, we would hope that the number of correct objects increases as more views are collected, as this demonstrates that our bag-of-features score is able to correctly rank objects once suitable views are available.

The top line in the graph displays the frequency at which pictures taken with the foveal camera actually contain the desired objects. The fact that this line is densely populated, especially after the robot begins to look back at objects indicates that the viewpoint collection behaviour is operating successfully. At the end of the experiment, 6 out of 8 of the objects have correctly top-ranked views (as can be seen at the right of the graph), which indicates the robot will be obtaining many novel views. The top-ranked objects upon completion of the experiment are shown in figure 10.

The final evaluation method used was to examine the goal locations produced by the multiple viewpoint collection behaviour, and verify that these did indeed allow for novel views to be collected. Figure 11 is one example of a path generated by this method. In this case, the end of the path provides a significantly different viewpoint for many objects, as desired.

VII. Performance Issues

Our current implementation collects images at the rate of approximately 5 per minute during coverage and 2 per minute during the look-back behaviour. If desired, the system could be made to run much faster. For example, the matching of features in captured images with memory is done using an exhaustive search. This search could be speeded up by orders of magnitude by organising the features in memory in a tree structure. E.g. kD trees [36], ball-trees and k-means trees [37] can be used here. These algorithms have access times that are logarithmic in the memory size, an essential property if a method is to scale to memories consisting of thousands of objects.

VIII. Concluding Remarks

We have demonstrated that actively collecting images from various viewpoints enhances object recognition. Regions
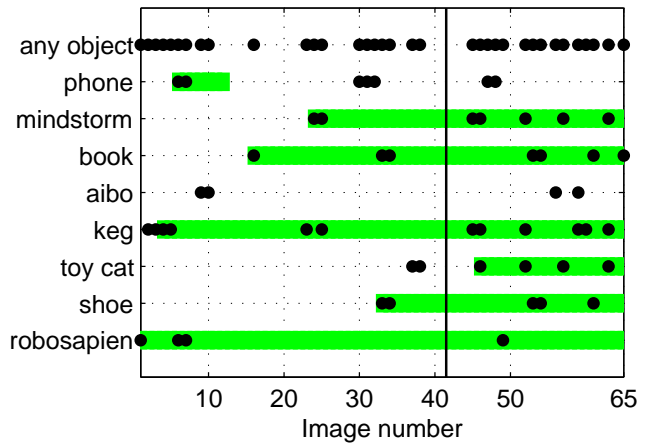

Fig. 9. Evaluation of look-back behaviour. Each black dot signifies that the corresponding image contains the object in this row. E.g. image number 33 contains both the book and the shoe, and thus has black dots in two rows. The green bars signify that the robot's currently best guess for a particular object is correct. The black vertical line indicates the switch between the coverage behaviour and the look-back behaviour, which occurred when the coverage threshold reached 0.9 in this experiment.
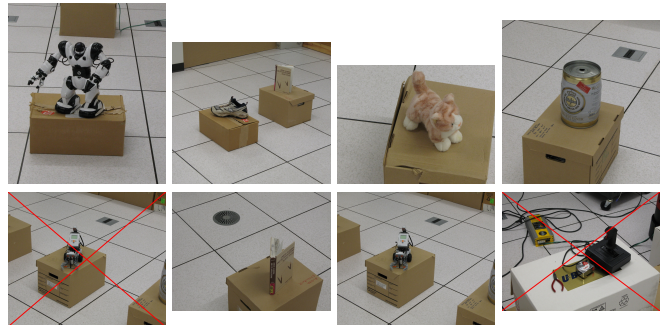

Fig. 10. The robot's best guesses at the end of the experiment. Crossed out images are incorrect.

identified by bottom-up saliency are ranked using top-down information, in the form of bags of features from the models of the sought objects. In the future, we plan to make even better use of the high level information, by using it to improve object segmentation. Also, we have shown some initial results on obtaining useful viewpoints, but there is more work to be done in this area. For example, we are currently only planning for horizontal view coverage, but plan to also add vertical view planning. The approach described in this paper is already an important step towards solving the lost-and-found problem. With future enhancements, it could form the basis for a domestic robot companion capable of responding to "Robot, fetch me my shoes!"
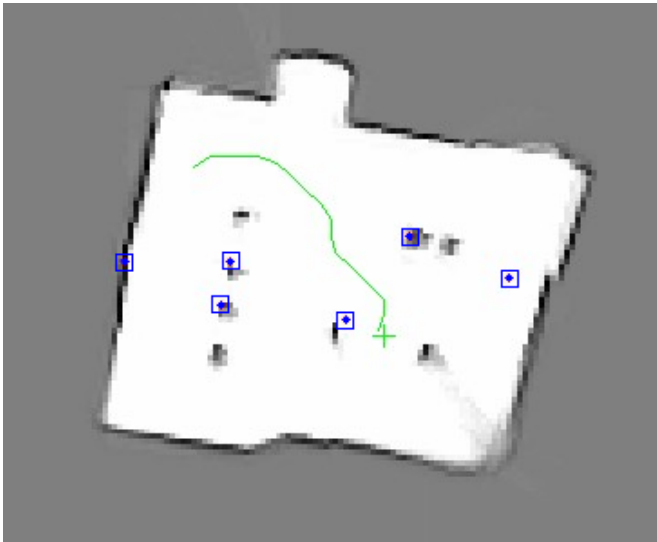
Fig. 11. An example of a path generated to obtain novel object views. The locations of the highest ranking potential objects are displayed as squares, with the robot's current position as a cross. The robot's path (shown as a green line where colour is available) moves it such that the objects can be observed from another direction.

REFERENCES

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

[2] Website: http://www.robocupathome.org/.

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

[4] 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.

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

[6] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137 – 154, 2004.

[7] S. Obdržálek and J. Matas, "Sub-linear indexing for large scale object recognition," in *BMVC 2005: Proceedings of the 16th British Machine Vision Conference*. London, UK: BMVA, September 2005, pp. 1–10.

[8] C. Laporte and T. Arbel, "Efficient discriminant viewpoint selection for active bayesian recognition," *International Journal of Computer Vision*, vol. 68, no. 3, pp. 267–287, 2006.

[9] L. Paletta and A. Pinz, "Active object recognition by view integration and reinforcement learning," *Robotics and Autonomous Systems*, vol. 31, pp. 1–18, 2000.

[10] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Computer Vision and Image Understanding*, vol. 100, pp. 41 – 63, June 2005.

[11] 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.

[12] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Meissner, G. Bradski, P. Baumstarck, S. Chung, and A. Ng, "Peripheral-foveal vision for real-time object recognition and tracking in video," in *Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*, 2007.

[13] 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.

[14] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, T. S. Sancho McCann, M. Baumann, J. J. Little, D. G. Lowe, and B. Dow, "Curious george: An attentive semantic robot," in *IROS 2007 Workshop: From sensors to human spatial concepts*. San Diego, CA, USA: IEEE, November 2007.

[15] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, D. G. Lowe, and B. Dow, "Curious george: An attentive semantic robot," *Robotics and Autonomous Systems Journal*, 2008, accepted.

[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] 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.

[18] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

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

[20] 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.

[21] 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.

[22] 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.

[23] B. Rasolzadeh, M. Björkman, and J.-O. Eklundh, "An attentional system combining top-down and bottom-up influences," in *International Cognitive Vision Workshop (ICVW), ECCV'06*, 2006.

[24] S. Mitri, S. Frintop, K. Pervölz, H. Surmann, and A. Nüchter, "Robust object detection at regions of interest with an application in ball recognition," in *IEEE International Conference on Robotics and Automation*. Barcelona, Spain: IEEE, April 2005, pp. 125–130.

[25] J. K. Tsotsos, S. M. Culhane, W. Y. Kei, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via slective tuning," *Artificial Intelligence*, vol. 78, pp. 507–545, 1995.

[26] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV International Workshop on Statistical Learning in Computer Vision*, Prague, 2004.

[27] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Weak hypotheses and boosting for generic object detection and recognition," in *ECCV*, vol. 2, 2004, pp. 71–84.

[28] B. Leibe and B. Schiele, "Interleaved object categorization and segmentation," in *British Machine Vision Conference (BMVC'03)*, Norwich, UK, Sept. 2003, pp. 759–768.

[29] 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.

[30] P. Moreels and P. Perona, "Evaluation of features detectors and descriptors based on 3D objects," *IJCV*, vol. 73, no. 3, pp. 263–284, July 2007.

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

[32] 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.

[33] A. Wald, *Sequential Analysis*. New York: Dover, 1947.

[34] J. Matas and J. Sochman, "Wald's sequential analysis for time-constrained vision problems," in *In Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2007.

[35] Timothy H. Chung and Joel W. Burdick, "A Decision-Making Framework for Control Strategies in Probabilistic Search," in *Intl. Conference on Robotics and Automation*. ICRA, April 2007.

[36] J. Beis and D. Lowe, "Shape indexing using approximate nearest-neighbour search in highdimensional spaces," in *Conference on Computer Vision and Pattern Recognition*, Puerto Rico, 1997, pp. 1000–1006.

[37] K. Mikolajczyk and J. Matas, "Improving sift for fast tree matching by optimal linear projection," in *IEEE International Conference on Computer Vision*, vol. CFP07198-CDR. Rio de Janeiro, Brazil: IEEE Computer Society, October 2007.