Viewpoint Detection Models for Sequential Embodied Object Category Recognition

David Meger, Ankur Gupta and James J. Little

***Abstract*— This paper proposes a method for learning viewpoint detection models for object categories that facilitate sequential object category recognition and viewpoint planning. We have examined such models for several state-of-the-art object detection methods. Our learning procedure has been evaluated using an exhaustive multiview category database recently collected for multiview category recognition research. Our approach has been evaluated on a simulator that is based on real images that have previously been collected. Simulation results verify that our viewpoint planning approach requires fewer viewpoints for confident recognition. Finally, we illustrate the applicability of our method as a component of a completely autonomous visual recognition platform that has previously been demonstrated in an object category recognition competition.**

I. INTRODUCTION

When a human attempts to identify what they are looking at, they may often pick objects up to rotate them or move their head from side to side in order to obtain a variety of viewpoints. In some cases this behavior allows a "canonical" viewpoint of the object to be obtained (e.g. the label on a bottle) and in other cases, the movement may allow disambiguation between similar items (e.g. searching for the logo to identify the brand of car being viewed). Humans integrate information over the numerous viewpoints they see without effort, and can rapidly decide where to move next to gather the most information. In contrast, the analogous scenario remains a challenge for a visually guided mobile robot.

For robots that attempt to interact naturally with humans in home environments, the majority of tasks require semantic knowledge about the category labels of objects. Hence, the problem of "where to look" and how to integrate information from multiple views, so easily solved by humans, is a vital requirement. Existing active vision methods are primarily focused on specific objects with easily described appearances. Multiview object recognition techniques from the Computer Vision community have also recently shown strong performance on recognizing specific objects, but these do not generalize to many of the object categories found in a typical home. Category recognition has mostly been studied in the context of labeling a single image from a database, which ignores several aspects facing a robot system such as viewpoint. Note that we distinguish between specific instances such as "Norco Launch 2002 Mountainbike" and generic object categories such as "bicycle".

All authors are with the Department of Computer Science, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4 {dpmeger,ankgupta,little@cs.ubc.ca}

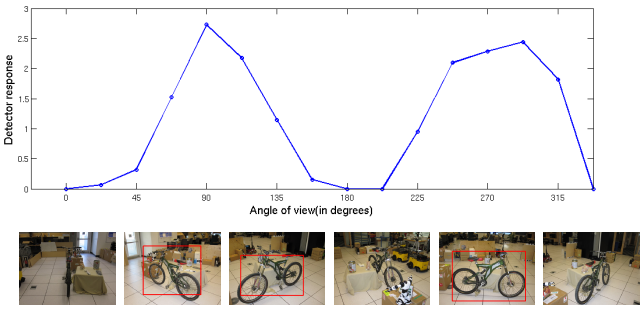

Fig. 1. The response of the *deformable parts model* detector from [1] on images of a bicycle from numerous viewpoints. Images shown below align with datapoints, and bounding boxes drawn in images represent detector responses that exceed a threshold pre-calibrated to balance precision and recall.

This paper proposes a sequential category recognition (SCR) solution centered on learning models of a category detector's response with respect to viewing direction using training data from a multiview category database. For example, Figure 1 shows the detection responses of a state-of-the-art category recognizer on a number of views of a single bicycle. The model learning procedure described below summarizes the responses of a detector across numerous instances and to capture its dependence on viewpoint. These learned models allow for fusion of the information from a sequence of images of the same object using sequential Bayesian estimation. Also, informative viewpoints can be chosen based on the current estimate and viewpoint model, which allows an active system to recognize an object with fewer views.

We have constructed our SCR solution in the context of an integrated visual search robot system named Curious George [2]. This system has previously been evaluated in the Semantic Robot Vision Contest (SRVC) [3], a competition amongst completely autonomous object category recognition platforms. SRVC requires systems to use Internet imagery for learning visual models (no human annotation), to autonomously explore a realistic environment, and to use the learned models to visually identify the presence of instances from object categories, placed by the organizers. Curious George placed first in the robot category of the SRVC for 2007 and 2008, recognizing roughly half of the test objects However, all contestants in SRVC mainly recognized those objects with specific appearances, and rarely the instances of truly generic categories (e.g. the robots always recognize Coke cans and never vacuums). We have observed that a

primary challenge in recognizing generic categories during the SRVC contest or in any quasi-realistic home scenario is that appearance differs drastically across viewpoints and state-of-the-art recognizers are not well suited to model this fact.

We do not refer to the method described in this paper as pose estimation. Although accurate pose inference for category recognition would be extremely useful, it is beyond the state-of-the-art for methods in visual modeling, for all but the most geometrically consistent categories. This is in contrast to pose estimation for a specific instance where methods in automated feature matching and geometric inference (techniques similar to those used for Visual SLAM) allow highly accurate solutions. The instances that share semantic labels (category members) often have drastically different geometry, and the models suitable for capturing their variation are inexact in nature. Several recent methods include [4], [5], [6]. Our viewpoint-dependent models of detector response can be seen as a soft form of pose estimation and are inspired by the approaches listed.

The next section will discuss related work in Active Vision and multiview object category recognition. Following, our SCR method is presented, along with our strategy for learning viewpoint detection functions, and an entropy minimization planning algorithm. Finally, we will present results for evaluation of the system on a simulator and with a physical robot platform.

II. RELATED WORK

Embodied object recognition systems, and in particular those aimed towards home robotics, often consider similar problems to those addressed in this paper (e.g. [7], [8], [9]). In particular, Ye *et al.* [10] have considered modeling the variation in viewpoint when observing a specific object and learning this model from training data. We have been inspired by this approach, and have performed a similar analysis for the response of object category detectors over many instances per class. More recently Sjo *et al.* have constructed a highly capable recognition system [11] but note explicitly that object viewpoint is not modeled in their work at present.

The problem of moving a camera through the world to aid in inference is typically referred to as Active Vision. The idea of minimizing the entropy of an estimator based on camera motion has been studied in the Active Vision community, notably by [12], [13]. Various authors (e.g. [13]) have previously suggested the use of a generative model of object appearance conditional on the object label and other confounding variables such as pose and lighting $p(A|o, \theta, l)$, along with a sequential Bayesian update strategy in order to solve this problem. However, these models have typically been associated with systems performing relatively simple visual tasks such as recognition of specific instances of objects annotated with identifiable markers. This paper studies a method for Active Vision during category recognition.

Several authors have recently considered building object category recognizers that perform well over all viewpoints [4], [5], [6]. These methods typically require annotated data from a semi-dense sampling of viewing directions and in some cases require additional information such as a video sequence [4]. Several authors have also explored the variation of object category models with respect to viewing direction, similar to our work (e.g. [14], [15]). While multiview category recognition is a promising direction, it is unlikely that truly viewpoint invariant category recognition is possible due to the extreme intra-category appearance variation for some objects (e.g. the back sides of monitors). In fact, for some categories, human observers have difficulty in recognizing certain views, which leads to the behavior of turning the object with the hands or moving the head and eyes to see a different viewpoint.

In order to learn the viewpoint detection function for an object recognizer, validation data containing multiple viewpoints of numerous instances of each category is required. Many image databases containing multiple viewpoints of objects have recently been made available, however, we found that many of them did not fit our purposes. For example, Viksten *et al.* [16] collected a database with fine-grained viewpoint sampling for each object, but only a single instance of each category is present, as their efforts have been targeted towards grasp planning for industrial applications. The use of the Internet as an online forum for users to annotate data has been used to produce very large labeled databases such as LabelMe [17]. Also, online task auction sites are suitable for dataset construction and have been used in ImageNet [18] and also by [19]. These large category datasets have so far not been annotated with image viewpoint. A dataset collected by Savarese *et al.* [20] contains 72 views (8 azimuth angles, 3 heights and 3 scales) of each of 10 instances for 10 common object categories. While containing far fewer object instances than some other resources, the precise viewpoint labels associated with each image make this dataset suitable for evaluation of multiview techniques, and it will be used to construct our viewpoint detection response functions in the next section.

III. SEQUENTIAL CATEGORY RECOGNITION

We define the active sequential category recognition problem as inferring the category label of an object based on a series of images collected over time and from various viewpoints, as well as choosing new viewpoints at each timestep (path planning). However, without factoring the problem, an SCR solution would be required to plan in an extremely high dimensional search space formed by every control action of the robot and its camera. To focus the discussion, in this paper we consider a subset of this problem by assuming several visual processing tasks can be completed efficiently outside the scope of this work (in our case, we defer to existing system components of Curious George [2]). In particular, we assume that locations of the world have been identified as potential object candidates (proto-objects), for example by a mid-level visual attention system that chooses targets and segments potential objects from the world. This leaves the task of choosing the next viewing angle from which to observe one of the candidate

objects. Additionally, for the purposes of this paper, we do not choose between multiple objects, but consider the scenario where a single new object has been encountered and the robot is tasked to verify its identity before moving on to the next task.

In short, we consider the restricted variant of SCR where the robot must infer $p(o = x|f_1...f_N)$, the likelihood that the proto-object has category label x, for all categories conditioned on the classification responses received on N images so far. The system must choose a sequence of viewpoints $(\theta_1, ...\theta_N)$ from which to observe the object - that is it has some ability to actively select which data to examine. Our solution follows a similar approach to that taken in [13] to infer the category of the object being considered as well as the pose of the object. Specifically, we have trained a number of generative detector models: $p(f(A(V_\theta))|o, \theta)$, where $f(A(V_\theta))$ represents the response of a detector f evaluated on an image with appearance $A(V_\theta)$, for a given view V_θ. o is a variable representing the category label and θ represents the pose of the object. The image appearance obtained from a viewpoint, $A(V_\theta)$ is a complicated function depending on an object's appearance and many factors in the environment such as lighting. For simplicity in much of the discussion, we will describe a detector's response as f_i, indexing only by i, the order in which the image was taken – the reader is asked to remember that the detector's response is a function of the viewpoint and environmental factors.

A. Learning a Viewpoint Function

As mentioned, the score of an object recognizer trained on a single viewpoint of each object is likely to be biased towards that viewpoint. Correctly modeling this fact will allow a visual search system to correctly infer the state of the world, and so we set out to model the detection response as a function of viewpoint for several state-of-the-art object recognizers trained on a variety of datasets. In particular, we have examined three object recognition approaches that are currently used heavily in the Computer Vision community:

1) *SIFT matching* is an algorithm based on the observation that local image features can be reliably detected and described in a fashion that is largely invariant to changes in scale, lighting and in-plane rotation [21] (N.B. the list of invariances does not include viewpoint changes, although invariance over a small range of views is possible, as discussed in [22]). In particular, we have implemented image matching based on SIFT features with RANSAC to fit a fundamental matrix to a candidate set of point matches in order to discard outliers and return highly confident match results.
2) *Bag-of-Features Matching* is equivalent to SIFT-matching without checking of the geometric consistency between feature matches. This allows the method to generalize better across intra-category variation in geometry and makes the approach more suitable for category recognition. Note, for clarity, that we have not utilized vector-quantized features or an SVM for classification as has been attempted by [23] and is often also refered to as "Bag-of-Features".
3) *Deformable parts model* is an algorithm that combines several feature types and jointly infers parts and object labels with an SVM. This method was selected due to its strong performance on the recent Pascal Visual Object Categories competition [24]. We have used the author's implementation for this method [1].

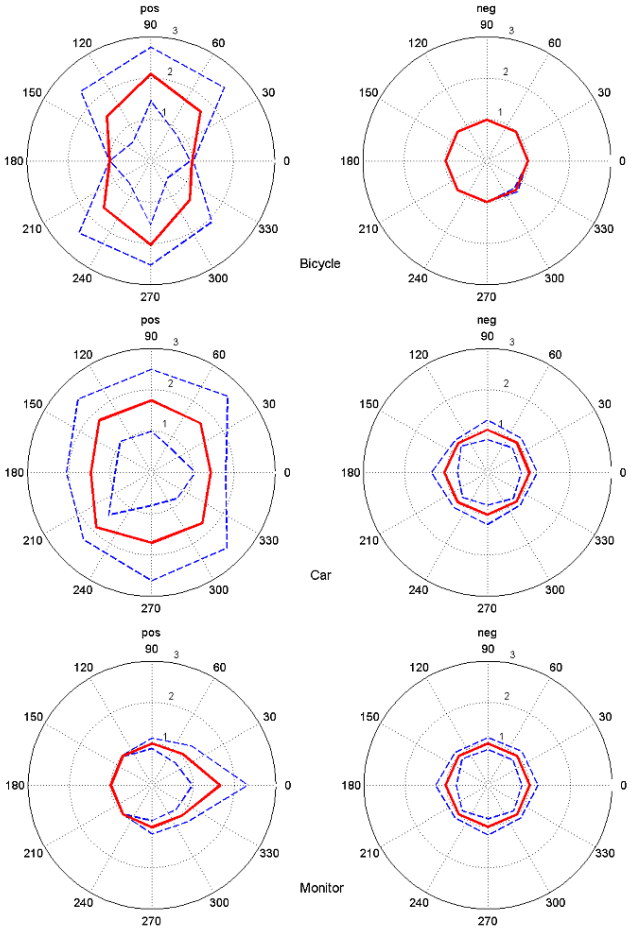

Fig. 2. Example viewpoint detection functions of the *deformable parts model* detector for classes: (top) bicycle, (middle) car, and (bottom) monitor. The radial coordinate represents the detector response to positive(left) and negative(right) samples. The solid red line is the expected value and dotted blue lines depict the uncertainty in the response.

Each of the three methods was evaluated across a large number of views drawn from the object category dataset which has recently been collected by Savarese *et al.* [20] described above. The detector results over this set characterize the distribution of responses over viewpoints. We modeled the empirical distribution of detector responses obtained over the dataset with a univariate normal per $\{o, \theta\}$ pair. This produces a viewpoint likelihood function: $p(f|o, \theta)$ which can be evaluated for each detector response and integrated into the SCR framework as will be shown below.

Several viewpoint detection models for the *deformable parts model* are displayed in Figure 2. Each row in the image represents the response given for a different category:

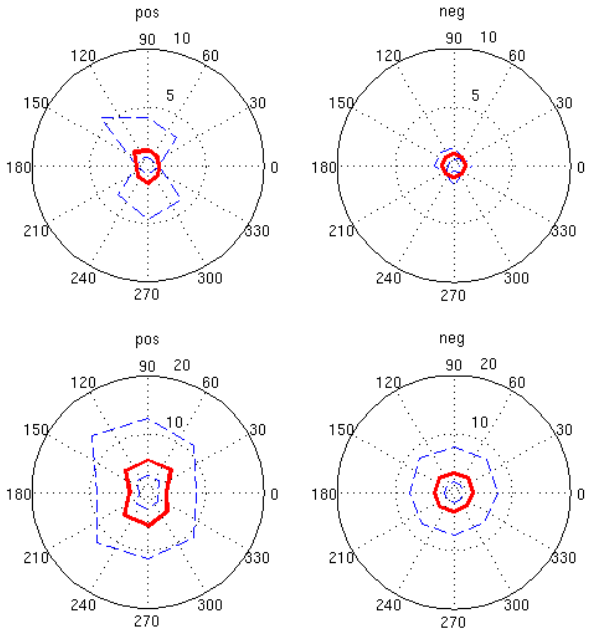

Fig. 3. Viewpoint detection function for the (top)*SIFT matching* and (bottom) *Bag-of-Features* detectors. The radial coordinate represents the detector response to positive(left) and negative(right) samples. The solid red line is the expected value and dotted blue lines depict the uncertainty in the response.

bicycle, car and monitor. Some notable structure is present in each: responses for the bicycle category show clear symmetries, and, as was clear in Figure 1, the front and back views give much lower detector responses than views from close to the side; responses for cars have a similar shape, but the front and the back views are somewhat more recognizable due to a car's greater width and identifiable features such as headlights; and finally, the response function for monitors is the canonical single-viewpoint recognition scenario as monitors only demonstrate a reliable appearance from straight-on front views.

Figure 3 shows the viewpoint detection models of the *SIFT matcher* and *Bag-of-Features* approaches when trained to recognize bicycles. The viewpoint profile of the responses for both methods are similar to those observed in the previous figure, which adds support to the observation that side views of bicycles are more readily distinguishable than front and rear views. In contrast to the *deformable parts model*, however, we found that the detectors' response functions for negative instances (images that do not contain bicycles) were nearly as strong as those for the positive instances (images containing bicycles) over most of the viewpoint range. This is due to the fact that the feature matching step in both of these approaches returned a small number of features for instances of the category not present in the training set. That is, the local object appearance varied too greatly for correct matching. This is can be seen in the figure in that the mean values for both positive and negative responses are similar. For this reason, we have primarily focused on the *deformable parts model* in the rest of the results given in this paper. Integrating a specific view recognizer such as the *SIFT matching* approach with a general category recognizer is left for future work.

B. Multiview Sequential Bayesian Estimation

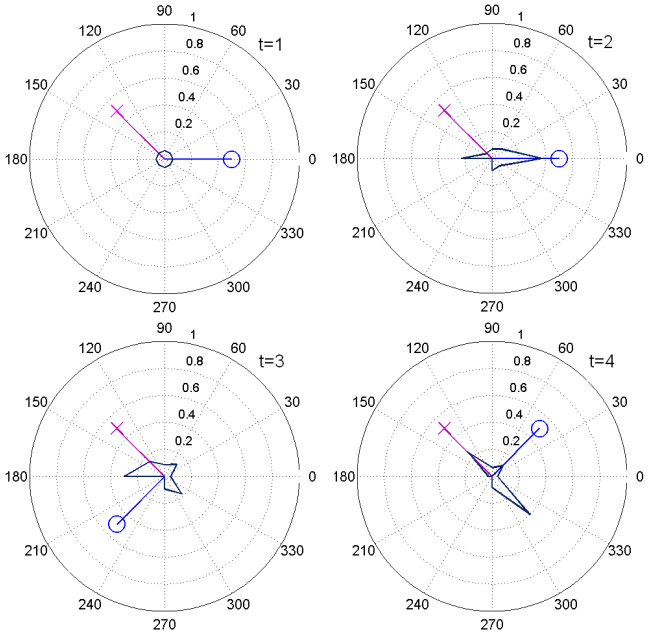

Fig. 4. The posterior distribution over object presence and pose is updated as each image is collected. This is demonstrated for 4 steps of one robot recognition simulation trial. The graphs display: the prior top-left $p(o = x, \theta)$, the posterior after one image top-right, $p(o = x, \theta | f_1)$, and so on. In each graph, the radial coordinate represents the belief probability for the object occuring and having the pose indicated by the angular coordinate. This trial is evaluation of the category label "car" and the true world state is that a car is present with pose $135°$. The magenta "x" shows the pose of the object and the blue circle shows the robot's pose at each time step.

This section describes our approach to integrating the scores of classifiers over images of an object from multiple viewpoints. We build upon the viewpoint detection models described previously. Consider inferring $p(o, \theta | f_1 ... f_N)$, the probability that an object is present at a given viewpoint based on n responses. This can be easily derived using Bayes' Rule:

$$p(o, \theta | f_1 ... f_N) = \frac{p(f_1 ... f_N | o, \theta) p(o, \theta)}{p(f_1 ... f_N)} \quad (1)$$

$$= \frac{p(f_1 ... f_N | o, \theta) p(o, \theta)}{\sum_{o_i \epsilon \{t, f\}} \sum_{\theta_j \epsilon [0, 2\pi)} p(f_1 ... f_N | o_i, \theta_j)} \quad (2)$$

We make the standard Naive Bayes assumption, that each pair of classifiers is conditionally independent given the object label and viewpoint, so the expression becomes:

$$p(o,\theta|f_1 ... f_N) = \frac{p(o,\theta)\prod_{k=1}^{N} p(f_k|o,\theta)}{\sum_{o_i \epsilon \{t,f\}} \sum_{\theta_j \epsilon [0,2\pi)} \prod_{k=1}^{N} p(f_k|o_i,\theta_j)} \quad (3)$$

We have learned generative viewpoint detector models for $p(f_k|o,\theta)$ as described previously, and we use an uniform prior for $p(o,\theta)$ in our work. In extensions to integrated systems, it is likely to be beneficial to use domain knowledge to specify an informative prior such as the likelihood of each type of object occurring in each room of a house. This is left for future work. Also, please note that we have excluded a model for robot motion in this work. For simplicity, we assume that the robot's motion is known exactly. While this is not true in general, our work makes a very coarse discretization of angle into 8 bins, and so it is likely that we can correctly determine the correct bin for the robot's position a large fraction of the time from odometry or SLAM position estimates.

Figure 4 illustrates the posterior evolving over each time step for the object category "car". As a new observation is made, the updated posterior function becomes narrower and eventually aligns with the actual pose of the object. This corresponds to probabilistic estimation of pose and category.

C. Viewpoint Planning

The active component of our SCR system requires a decision making strategy to control the position of the camera in the world – the viewpoint from which objects are observed. The choice of camera motions allows numerous views to be collected so that, for example, the canonical viewpoint present in the training data can be observed, or a view can be obtained that allows objects with similar appearances to be disambiguated. We employ entropy as a measure to determine the confidence of the recognition system in its belief about the presence (or absence) of the object. Entropy is defined as follows:

$$H(p(x)) = -\sum_i p(x_i)\log(p(x_i)) \quad (4)$$

For random variable x. For the viewpoint planning problem, we attempt to minimize the entropy of the posterior belief by selecting the next viewpoint V_θ as follows:

$$V_\phi^* = argmin_\phi H(p(o,\theta|f_1 ... f_N, f(A(V_\phi)))) \quad (5)$$

Search for the minimizing view requires evaluation of equation (3) for each viewpoint, which is not trivial because it depends on the next detector response that will be obtained – a quantity that cannot be known exactly until after the planning action has been executed. Integration over all possible detector responses (a continuous variable) is computationally expensive. It can be avoided by computing the expected classifier response, but this produces a biased estimate for the entropy. So, instead we draw a number of samples for the value of f from $p(f|o,\theta)$ and compute the minimum averaged over these samples.

IV. EXPERIMENTAL RESULTS

A. Simulated Multiview Recognition

We have constructed a simulated recognition environment to test our SCR approach. The simulator models a robot's position with respect to an object, and returns a pre-collected image from the simulated robot's viewpoint, in place of an image that would be acquired by a physical robot's camera. The pre-collected images were drawn from a hold-out portion of the Savarese *et al.* dataset used during training. We evaluated a variety of detectors on each image and used the responses to update the recognition system's posterior belief about the object's presence and viewpoint.

We compare our method with a non-adaptive viewpoint selection strategy that chooses a random previously unseen view at each timestep. This method has been a favorite approach for contestants in the SRVC contest, and was suggested in [2] as an approach that obtains coverage of viewpoints while reducing viewpoint overlap early in the search process. Compared to other non-adaptive strategies, the random approach may find interesting views faster, at the cost of additional robot motion.

We compared each planning strategy using our viewpoint simulator by repeatedly simulating detection results and allowing the planner to view the result and choose a new robot position. For statistical significance, 160 trials were conducted. Between each trial a different object instance is chosen at random from the validation set. Also a random initial viewing angle is chosen from one of the 8 azimuth angles available in the Savarese dataset. The object's identity and initial viewpoint are hidden from the planning approaches, so the situation is a realistic approximation to the situation where the robot segments a proto-object from the world, has no prior knowledge about the category label or viewpoint of the object, and must infer these quantities by collecting and analyzing images.

Figure 5 summarizes the results of the simulation trials. The results demonstrate that planning to reduce entropy allows the recognition system to confidently infer the category label from fewer test images, since it is able to use the history of detector responses to determine the viewpoints that are most likely to discriminate the object. As more and more views are collected, the probability that the random strategy finds these views increases also, and once each method has exhausted the available viewpoints, performance is identical. A similar result is shown on the right of the figure. In this case, the rapid initial decrease in posterior entropy results from the planner discovering discriminative views, and the subsequent small increase in entropy results from the fact that we force the planner to continue even after it has essentially converged on its decision about the object, so it encounters the viewpoints that are difficult to discriminate later in the recognition process. In both cases,

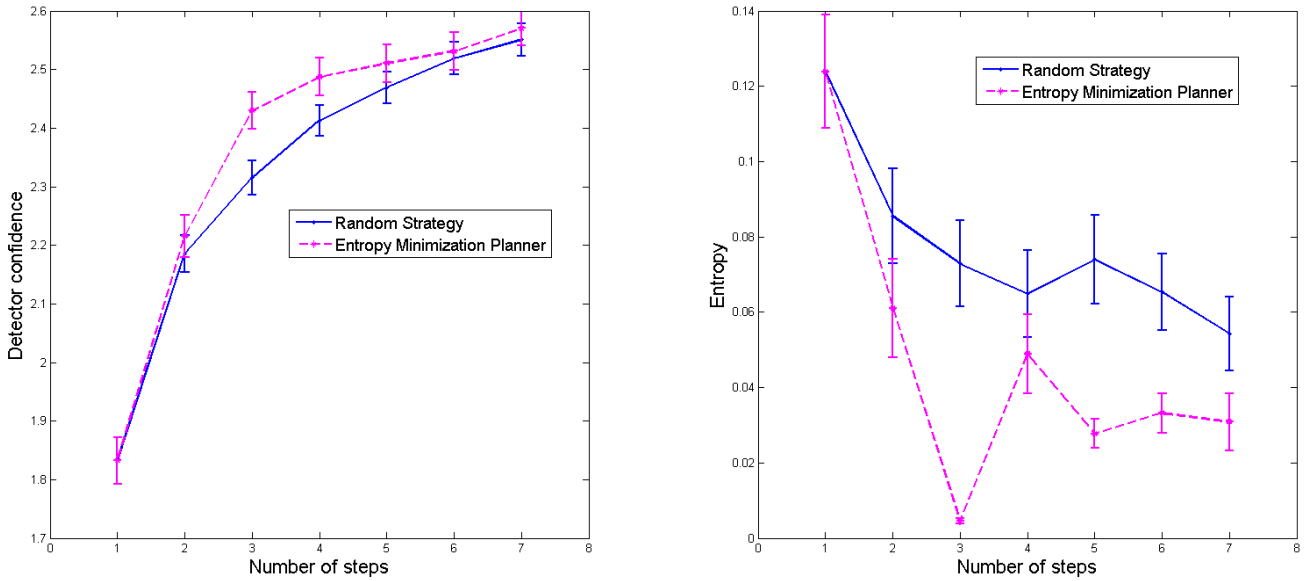

Fig. 5. A comparison of detection results between a system using entropy minimization planning and a system which uses a random planning strategy. The graph on the left shows the sum of detector responses for true positives minus the sum of responses for true negatives, a summary statistic for classification performance. The graph on the right shows the entropy of the marginal $p(o = x|f_1...f_N)$, the detector's belief in the true category label x. All results are averages over 160 random selections of an object instance and starting viewpoint.

the results demonstrate that adaptive, entropy minimization planning aids in the sequential object recognition process.

B. Visual Search with the Curious George Platform

The Curious George platform uses visual saliency and depth cues to locate possible objects in the environment. As mentioned above, these mid-level vision techniques limit the search space which includes infinite locations and point of views. Figure 6 shows a sample scenario where robot has identified a proto-object in its view. The bicycle is correctly segmented based on depth and visual saliency features, in realtime, and this candidate object is passed to the SCR system for evaluation. The viewpoint planning method described above is integrated with this pre-existing feature of the robot. We have previously applied a planning algorithm which weighs between multiple objectives such as map building, coverage of the environment and certainty of object labels, and the SCR method described here is an additional component within this framework, which will be evaluated during the upcoming SRVC contests.

V. CONCLUSION AND FUTURE WORK

This paper has outlined an active multiview framework that can be used by an embodied visual searcher to infer the identity of a target object being considered. We have demonstrated the dependence of state-of-the-art object recognizers on the viewpoint from which an object is seen. This relationship is always likely to be present given the wide variety of appearance amongst category members for some viewpoints. We have learned viewpoint detection models for a number of detectors, and demonstrated that the sequential Bayesian estimation approach is capable of leveraging these models to provide improved recognition performance when compared to single-view strategies. Our method has been evaluated on a simulator based on a dataset of challenging images and its applicability has been illustrated for a physical embodied platform, Curious George.

There are several natural extensions to the current work. In this paper we have evaluated three object detection algorithms, but have chosen the one which performed best overall to use in all cases. Instead, a visual search planner could be given the opportunity to integrate information from all detectors, or better yet, the visual searcher could choose which method to run at each viewpoint, prioritizing computation towards detection results that are likely to be informative. Also, we have focused our analysis to the visual search problem involving only a single target object. In a home environment, a robot is faced with a large number of potential targets, and it may also be tasked with exploring new regions to discover new objects. In this case, a visual search platform must choose between numerous potential objects as well as between the viewpoints for each object. This is a challenging problem, but solving it will produce an active visual search robot capable of determining the semantic categories of objects within a home and subsequently performing useful tasks for the human inhabitants.

REFERENCES

[1] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *In Proceedings of the IEEE CVPR*, 2008.

[2] 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 Special Issue on From Sensors to Human Spatial Concepts*, vol. 56(6), pp. 503–511, 2008.

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

Fig. 6. A sample scenario. Curious George looks at a bicycle and segments it from the background using its visual attention system.

[4] L. F.-F. Hao Su, Min Sun and S. Savarese, "Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009.

[5] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, and L. V. Gool, "Using multi-view recognition and meta-data annotation to guide a robot's attention," *International Journal of Robotics Research*, 2009.

[6] J. Liebelt, C. Schmid, and K. Schertler, "Viewpoint-independent object class detection using 3d feature maps," in *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[7] M. Schlemmer, G. Biegelbauer, and M. Vincze, "Rethinking robot vision - combining shape and appearance," *International Journal of Advanced Robotic Systems*, vol. 4, pp. 259 – 270, 2007.

[8] G. Medioni, A. R. Franois, M. Siddiqui, K. Kim, and H. Yoon, "Robust real-time vision for a personal service robot," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 196 – 203, 2007, special Issue on Vision for Human-Computer Interaction.

[9] P.-E. Forssn, D. Meger, K. Lai, S. Helmer, J. J. Little, and D. G. Lowe., "Informed visual search: Combining attention and object recognition," in *In proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2008.

[10] Y. Ye and J. K. Tsotsos, "Sensor planning for 3d object search," *Computer Vision and Image Understanding*, vol. 73, p. 145168, 1999.

[11] K. Sjo, D. G. Lopez, C. Paul, P. Jensfelt, and D. Kragic, "Object search and localization for an indoor mobile robot," *Journal of Computing and Information Technology*, 2008.

[12] P. Whaite and F. Ferrie, "Autonomous exploration: Driven by uncertainty," McGill U. CIM, Tech. Rep. TR-CIM-93-17, March 1994.

[13] C. Laporte and T. Arbel, "Efficient discriminant viewpoint selection for active bayesian recognition," *International Journal of Computer Vision*, vol. 68, pp. 1573 – 1405, 2006.

[14] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *In Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2004.

[15] A. Opelt, A. Pinz, and A. Zisserman, "Incremental learning of object detectors using a visual shape alphabet," in *In Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2006.

[16] F. Viksten, P.-E. Forssen, B. Johansson, and A. Moe, "Comparison of local image descriptors for full 6 degree-of-freedom pose estimation," in *In proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[17] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, pp. 157–173, 2008.

[18] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei, "Construction and Analysis of a Large Scale Image Ontology." Vision Sciences Society, 2009.

[19] S. Vijayanarasimhan and K. Grauman, "Whats it going to cost you? : Predicting effort vs. informativeness for multi-label image annotations," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, June 2009.

[20] S. Savarese and L. Fei-Fei, "3d generic object categorization, localization and pose estimation," in *IEEE Intern. Conf. in Computer Vision (ICCV)*, Brazil, October 2007.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[22] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, pp. 43–72, 2005.

[23] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features." in *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Beijing, China, October 2005.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results," http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.