

Curious George: An Integrated Visual Search Platform

David Meger, Marius Muja, Scott Helmer, Ankur Gupta, Catherine Gamroth,
Tomas Hoffman, Matthew Baumann, Tristram Southey, Pooyan Fazli,
Walter Wohlkinger, Pooja Viswanathan, James J. Little and David G. Lowe
The University of British Columbia
Laboratory for Computational Intelligence
2366 Main Mall, Vancouver, BC, Canada

James Orwell
Kingston University
Faculty of Computing, Information Science and Mathematics
Penrhyn Road, Kingston-upon Thames Surrey KT1 2EE UK

Abstract

This paper describes an integrated robot system, known as Curious George, that has demonstrated state-of-the-art capabilities to recognize objects in the real world. We describe the capabilities of this system, including: the ability to access web-based training data automatically and in near real-time; the ability to model the visual appearance and 3D shape of a wide variety of object categories; navigation abilities such as exploration, mapping and path following; the ability to decompose the environment based on 3D structure, allowing for attention to be focused on regions of interest; the ability to capture high-quality images of objects in the environment; and finally, the ability to correctly label those objects with high accuracy. The competence of the combined system has been validated by entry into an international competition where Curious George has been among the top performing systems each year. We discuss the implications of such successful object recognition for society, and provide several avenues for potential improvement.

1 Introduction

Humans interact with their world each day largely based on visual understanding. The human visual system is a highly capable modelling and inference device. It quickly learns the appearances for new objects that are encountered, combines weak sources of information from multiple views, attends only to the most useful regions, and integrates numerous priors. As a result, humans can form highly accurate

representations of the world.

The analogous abilities, scene understanding and object recognition, are longstanding, but currently largely unachieved goals in Artificial Intelligence research. Computer Vision researchers have recently begun to make significant progress on the problem of recognizing objects in single images. For example, the best performing methods on the Pascal Visual Object Categories (VOC) challenge [3] are increasing recognition performance each year, and methods such as that of Felzenszwalb *et al.* [4] can now correctly classify objects over half of the time on average, for some object categories, when labeling images contained within an annotated image database. This performance has rarely been replicated by an integrated system such as a mobile robot that can translate successful recognition on such a hand-crafted scenario into real-world performance.

In particular, few robot systems have demonstrated the ability to recognize more than a one or two specific objects within realistic environments such as homes and offices. There are several significant challenges in applying an object recognition approach successfully on a physical platform. Pictures taken by a robot can often have significantly different properties, both in terms of image quality and viewing geometry, when compared to those taken by a human. In addition, objects present in a realistic environment are varied and constantly changing (unlike the static list of categories that is attempted year after year for the VOC, for example). While numerous data sets have been developed to validate the performance of object recognizers that label single images, there are very few such resources suitable for evaluating robot platforms.

A recent international competition, the Semantic Robot Vision Challenge (SRVC), has been developed in order

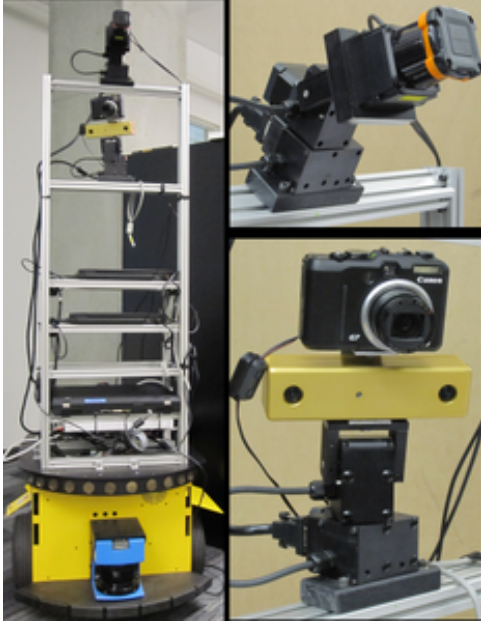


Figure 1. The robot component of our semantic environment mapping system.

to encourage development of robot recognition systems. Briefly, the SRVC is composed of a training phase where each competitor is required to train visual classifiers for previously unknown object categories based on images downloaded from the Internet or previously prepared image databases. This is followed by an environment exploration phase where robots must search a realistic environment created by the organizers in order to locate instances of the object categories listed during training. Successful recognition in this scenario is a strong indicator for good general performance since there is little possibility to tailor systems to a specific set of categories.

Please note that we distinguish between object categories such as “bottle” and specific objects such as “500 ml Diet Coke bottle”. Specific objects can often be recognized by direct image matching. Keypoint-based methods such as [10] have been extremely successful in recognizing specific objects. However, the recognition accuracy of state-of-the-art methods for object categories is generally much lower. In this paper, we consider a robot capable of the *category recognition* task. This differentiates our approach from many previous visual robotic platforms.

We have designed our robot platform, Curious George (shown in Figure 1), to compete in SRVC, and thus be competent in object localization. Curious George has placed first in the robot league of SRVC in 2007 and 2008 and first in the software league in 2009. Building on successes, each

successive Curious George system has become increasingly more capable of solving many of the challenges facing robot recognition systems that were listed above. This paper describes the system components that make up the “Curious George 3” platform – a system that has recognized 13 out of 20 objects during the 2009 SRVC, and includes techniques to successfully deal with many of the challenges facing embodied recognition platforms that were mentioned earlier. Curious George autonomously learns appearance models, navigates safely in and takes high quality pictures of its environment, and recognizes objects in the environment with state-of-the-art accuracy. The remainder of this paper will describe Curious George in detail.

The next section of this paper will discuss related work on robot recognition platforms. In the following section we will describe the components of the Curious George system. This paper will conclude by presenting numerous results obtained with the platform and by discussing future directions that will allow for additional improvements.

2 Background

Embodied object recognition systems, and in particular those aimed towards home robotics, often consider similar problems to those addressed in this paper (e.g. [17, 11, 13]). Notably, Ye *et al.* [22] have considered modelling the variation in viewpoint when observing a specific object and learn this model from training data. More recently Sjo *et al.* have constructed a highly capable recognition system [18]. Also, the Stanford Artificial Intelligence Robot (STAIR) [5] has been developed in parallel to the system that we describe. The systems are similar in that they both employ visual attention to guide the robot’s camera and a tilting laser to recover accurate 3D information. A previous version of the Curious George system was described by Meger *et al.* [12]. The current paper describes significant recent advances in this system since previous publication.

3 System Description

3.1 Overview

Inspired by the SRVC, Curious George has the ability to perform object recognition based on autonomous training of classifiers using Internet imagery. The robot autonomously explores a previously unseen environment in order to locate objects. A geometric map is constructed and when sensing occurs, this map is augmented to record the regions covered by the sensors. Several path planning algorithms allow the robot to cover the entire environment as well as to focus sensing efforts on areas likely to contain objects. An attention system based both on 3D structure and visual imagery

guides the robot’s camera, which can pan, tilt, and zoom to collect high quality images. This attention system allows images of interesting objects to be collected with little unnecessary background (i.e. floor and walls). The attention system is described in section 3.6. A number of object classifiers search the collected imagery for the presence of objects. Section 3.7 discusses these classifiers and preliminary investigation of integrating the results from various classifiers. We will now begin to discuss each system component in detail.

3.2 Hardware Components

The current implementation of the Curious George platform is comprised of a Powerbot from Mobile Robots Inc.¹ which provides mobility, a SICK² laser rangefinder mounted parallel to the floor at a height of 15 cm for basic obstacle avoidance and to perform Simultaneous Localization and Mapping (SLAM), a Canon G7 consumer digital camera for high resolution imaging, a Point Grey Research Bumblebee stereo camera to obtain stereo depth reconstructions, and a Hokuyo³ UTM laser rangefinder mounted on a tilt unit for more highly reliable sensing of three dimensional structure. In order to mount each of these sensors in the most useful configuration and to give the entire setup a degree of stability, we have constructed a reconfigurable but rigid tower based on aluminum profile components. Figure 1 illustrates the sensing setup.

The combination of a high resolution camera (the Canon G7) with a zoom lens and a pan-tilt unit enables the collection of high quality imagery of many areas in the environment with minimal robot motion, and allows imaging of regions that are inaccessible to the robot. This flexibility in imaging has proven extremely useful in the somewhat adversarial environments constructed by the organizers of the SRVC contest.

In addition to mobility and sensing, Curious George is enabled with significant computation capability. In the version used for the 2009 SRVC contest, Curious George included 6 unique computation units: the on-board processor used for navigation and simple sensory processing, as well as 5 laptops of various computational ability. Computation was shared between these systems using the open-source Robot Operating System (ROS) architecture as is described in section 3.3.

3.3 Software Architecture

As we have re-designed Curious George several times in preparing for the 3 years of participation in the SRVC con-

¹<http://www.activmedia.com/>

²www.sick.com

³<http://www.hokuyo-aut.jp/>



Figure 2. A visualization of the software architecture for Curious George. Each oval represents a single distributed process and arcs between ovals represent data topics that allow the processes to communicate.

test, we have explored a wide range of tools for robotic system integration including hardware drivers for the robot’s various sensors and middle-ware for distributing computations. The most successful solution so far has been “an open-source Robot Operating System” (ROS) [14]. ROS is a robot-specific middle-layer solution for distributed computation and message passing. It allows easy integration of sensor drivers and data processing components including both off-the-shelf and in-house components.

During the 2009 SRVC contest, approximately 50 independent processes were simultaneously executing on Curious George’s computational units. The distributed nature of ROS allows each independent component to function with some degree of independence and facilitates extensibility. Figure 3.3 illustrates the connectivity between the set of active components. The region of the graph which is contained within the red line represents the portion of Curious George software responsible for visual processing and object classification, as an example.

3.4 Web Image Download and Classifier Training

Curious George is able to interface to a number of sources in order to obtain visual imagery to train appearance models. This includes several public web-based data sources (namely Google Image Search and the Walmart product database), which have an exhaustive quantity of images for a large number of object categories, but contain a

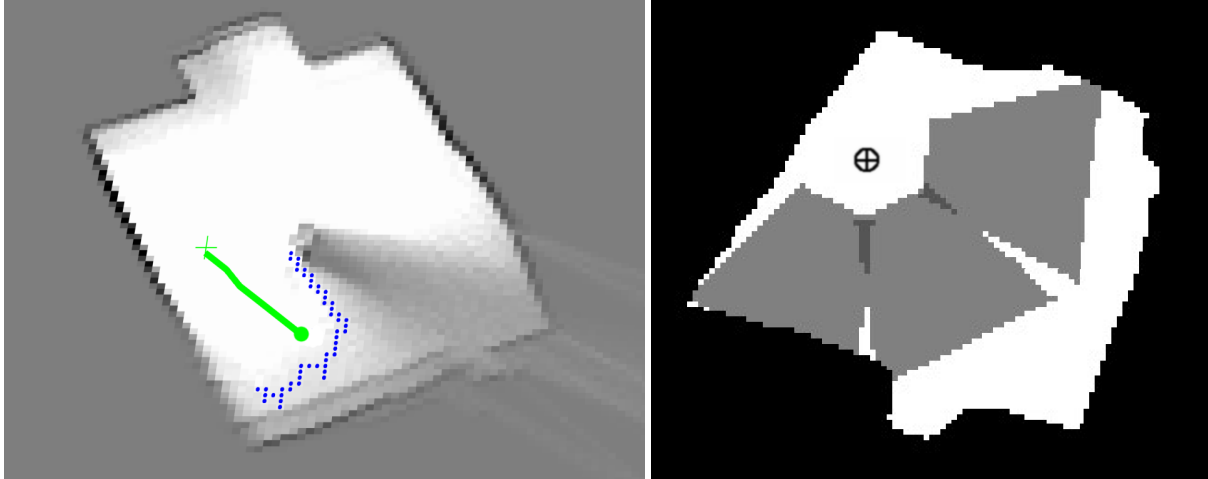


Figure 3. We have implemented a number of exploratory behaviours. The left image shows a frontier navigation goal. The right image is an illustration of the coverage map used to record regions of the environment already observed with the peripheral cameras. The interior of the black region represents the reachable map area, and grey shading indicates a camera field of view.

high level of noise in the image labellings since each image has presumably not been screened by a human at any point in the preparation of the search tool. Curious George also interfaces with several databases that have been specifically created for the Computer Vision community (namely the LabelMe database [15] and ImageNet [2]), which have a smaller number of total images, but with higher average quality since each image has been annotated by a human with the specific purpose of creating correct object labeling. Each of these data sources has proven useful for different tasks.

We have relied upon Google Image Search as our primary source of imagery for specific objects as it contains examples of nearly every object that can be imagined. To recognize object categories, a system must be able to handle a much larger degree of intra-category variation between the viewed instances and so the large ratio of mis-labeled images present in Google Image searches is unacceptable. In this case, we have primarily utilized the fewer but better labeled images available in Computer Vision databases.

3.5 Exploration and Mapping

In order to ensure the entire environment has been explored and to subsequently model environment geometry in a consistent fashion, Curious George attempts to construct a complete environment map. The map construction process involves SLAM to integrate laser and odometry data and form an occupancy grid map. We have employed the GMapping (see [6]) algorithm for SLAM which produces

occupancy grid maps. In order to guide the robot during map building, we have implemented a variant of the frontier-based exploration strategy proposed by Yamauchi *et al.* [21].

For visual search, a 2D occupancy representation is insufficient to represent the naturally 3D object positions. In particular, the plane on which the occupancy grid lies is embedded as a plane with height 15 cm. Objects are located at a variety of heights, and so a more complete spatial representation is clearly required. Therefore, after completing the construction of an occupancy-grid, Curious George employs a number of behaviours based on the tilting laser rangefinder in order to determine the positions of useful surfaces in the environment. In particular, we have employed a ROS package known as *table_object_detector* [16], which has been written by Radu Rusu in order to find horizontal surfaces in the environment. These surfaces are likely to be the tops of furniture such as tables and chairs, and are therefore likely locations for objects. The robot is actively guided through the room when searching for tables with a combination of a coverage behaviour and a procedure to determine likely furniture locations in the occupancy grid.

3.6 Attention

For a robot equipped with a pan-tilt-zoom enabled camera, there are an enormous number of potential views of the environment – far too many for any system to consider processing each view in order to locate objects. However, many of these views are highly redundant and others can be

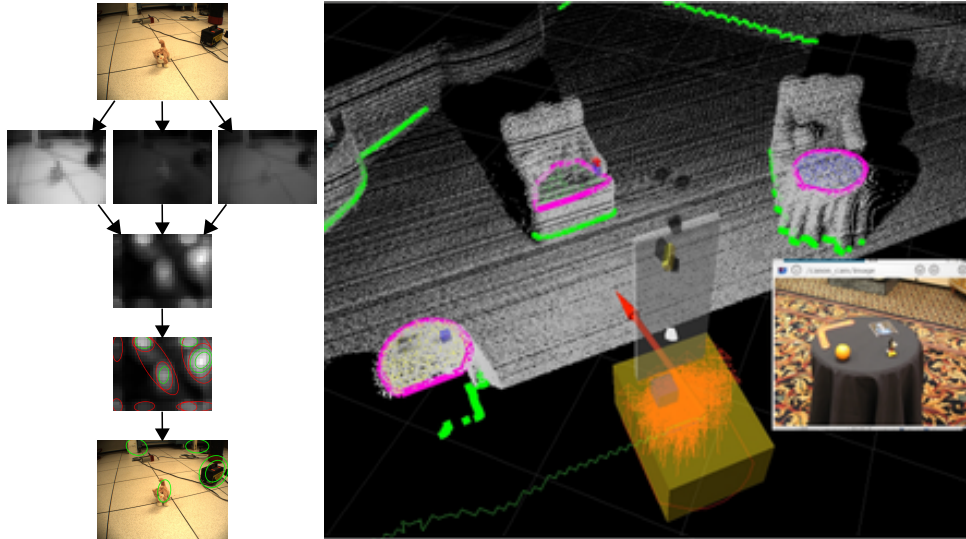


Figure 4. Two approaches that have been used to form Curious George’s attention system. (Left) Visual saliency computation. Top to bottom: Input image, colour opponency channels (int,R-G,Y-B), spectral saliency map, detected MSERs, and MSERs superimposed on input image. (Right) Structure information is used to determine the likely locations of furniture and objects of interest. Figure best viewed in colour.

deemed likely uninteresting based on simple cues. The goal of an attention system is to select all interesting views that are likely to be useful in later stages of processing while keeping the set of selected views small enough that later stages of processing are feasible. Curious George has a variety of approaches for selecting interesting regions, based both on visual appearance, as well as environment structure.

Visual saliency as a driving element of human attention has a long history in the neuroscience and vision science. For example, Treisman *et al.* [19, 1] demonstrated the importance of several low-level visual features for visual attention. Itti, Koch *et al.* [9] proposed an intermediate representation between various types of features and the attention system called the “saliency map”. A saliency map is roughly registered to the incoming visual field and encodes the sum “interesting-ness” of the region centered at each location. The use of saliency maps to segment interesting objects is a concept easily implemented for Computer Vision tasks and a well-known toolbox has been developed by Walther *et al.* [20]. After initial investigation, we have abandoned the use of the Walther toolbox for real-time robotics due to its relatively high computational cost and because there are several difficulties in tuning the scale parameters between various feature channels. We have instead adopted a variant of the Spectral Residual Saliency method developed by Hou *et al.* [8]. This is a computationally efficient approach for producing a saliency map based

on Fourier analysis. The method exploits the well known result that natural images are continuous in the log power spectrum on average. Regions of the image that do not obey this statistic are assigned high saliency.

Structure is also a powerful cue to determine the likely locations of objects in an environment. Two separate approaches can be considered: first, scene decomposition where priors such as objects appearing on top of furniture are used in a top-down fashion to prioritize regions based on their context in the structure of the environment; and second, local structure constraints such as the size of each particular object and object-specific priors such as the flatness (or lack thereof) of each particular object which can be used in a scanning type approach to rank each location individually. We have attempted each one of these approaches. The first has been mainly implemented by adapting the *table_object_detector* ROS package that clusters structure above furniture surfaces into potential objects. The local structure constraints have been explored only briefly as a secondary filtering loop on the salient regions proposed by visual attention. This method has attempted to discard regions that are clearly not promising because they are, for example, too large or small to be any of the objects of interest. This filtering process would be more effective if it were based on more informative 3D priors about the shape of each object (such as a template 3D model or set of 3D descriptors), and this is discussed as future work.

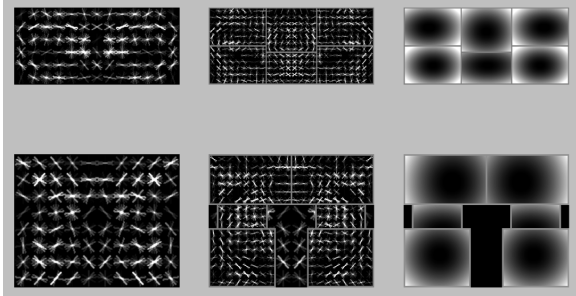


Figure 5. A visualization of the DPM detector’s underlying edge template representation of a frying pan. This edge kernel is applied to the gradient responses for each sub-window as a sliding window and the best matching windows are returned as candidate frying pans.

3.7 Visual Classification

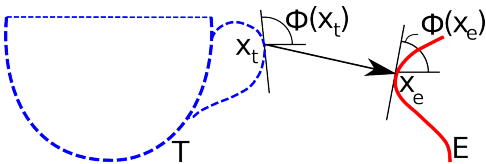


Figure 6. Contour matching based upon oriented chamfer-matching. The difference between two contours is the sum total of the minimum distance between each edgel in the model contour to its closest edgel in the image contour with the addition of an orientation difference between matching edgels.

Curious George currently relies upon three classification techniques to locate objects. First, a direct image matching technique based on Scale-Invariant Feature Transform (SIFT) features, similar to that described in [10] is utilized. Here, a set of keypoints are detected and the surrounding region encoded into descriptors, each containing a histogram of gradients. These locations can be reliably detected, and the descriptors are somewhat invariant to rotation, lighting, and minor changes in scale, and translation. Thus determining the presence of a known object in a new image comes down to finding similar features in a similar geometric arrangement. This approach is particularly effective for specific objects with texture, as it can reliably locate these objects in cluttered scenes at arbitrary scale or rotation. How-

ever, this approach is less effective for categories, or objects with little texture or more defined by their external shape.

To tackle objects defined primarily by shape, we also utilize a contour matching method based on edge detection, see Figure 3.7. Here, the external contours of the objects in our exemplar images are extracted using homogeneous background subtraction. The exemplar contours are then used to find similar contours via chamfer-matching at multiple scales, as in [7]. This approach is effective for cases of shape based objects for which we have few training images, but can have a high false positive rate in cluttered scenes. This can be mediated with structural information, or with the use of scale priors as in [7], but this is left for future work.

Finally, the system includes the Deformable Parts Model (DPM) classifier developed by [4], who also released the source code. This method is among the state-of-the-art for category recognition, placing highly in the recent Pascal Visual Object Categories (VOC) challenge [3]. This model is a mixture of a root filter and deformable parts based upon histograms of gradients, as seen in Figure 3.7, which is searched for in an image using a sliding window approach.

In order to combine the numerous classifiers that are evaluated on each image taken by the robot, as well as to fuse information between different viewpoints, it is essential to have a noise model for each detector’s response. For the 2009 SRVC contest, this was done by evaluating each classifier on an annotated validation image set, and the most confident classifier’s response was accepted for each category. More sophisticated viewpoint integration has not been applied during any SRVC contest, but this has the potential to improve the system’s accuracy, and will be discussed as future work.

4 Results

Using the techniques described above, Curious George is often able to recognize objects correctly in realistic environments. Figure 7 shows a sample of the results from the 2009 SRVC contest. The SIFT detector located almost all specific objects with high confidence. Results for the category recognition are mixed, but encouraging. Several object categories had been announced before the contest, and for these we were able to pre-train DPM detectors based on high-quality human-annotated imagery. Of these categories, Curious George correctly identified instances of “bottle” and “frying pan” but missed “toy car” and “laptop”. 4 additional object categories were revealed only at the beginning of the contest. For these categories, appearance models were trained based on Internet imagery – a less reliable description. The system correctly recognized an instance of “orange” and did not recognize “ping-ping paddle”, “pumpkin” and “white soccer ball”. The appearance

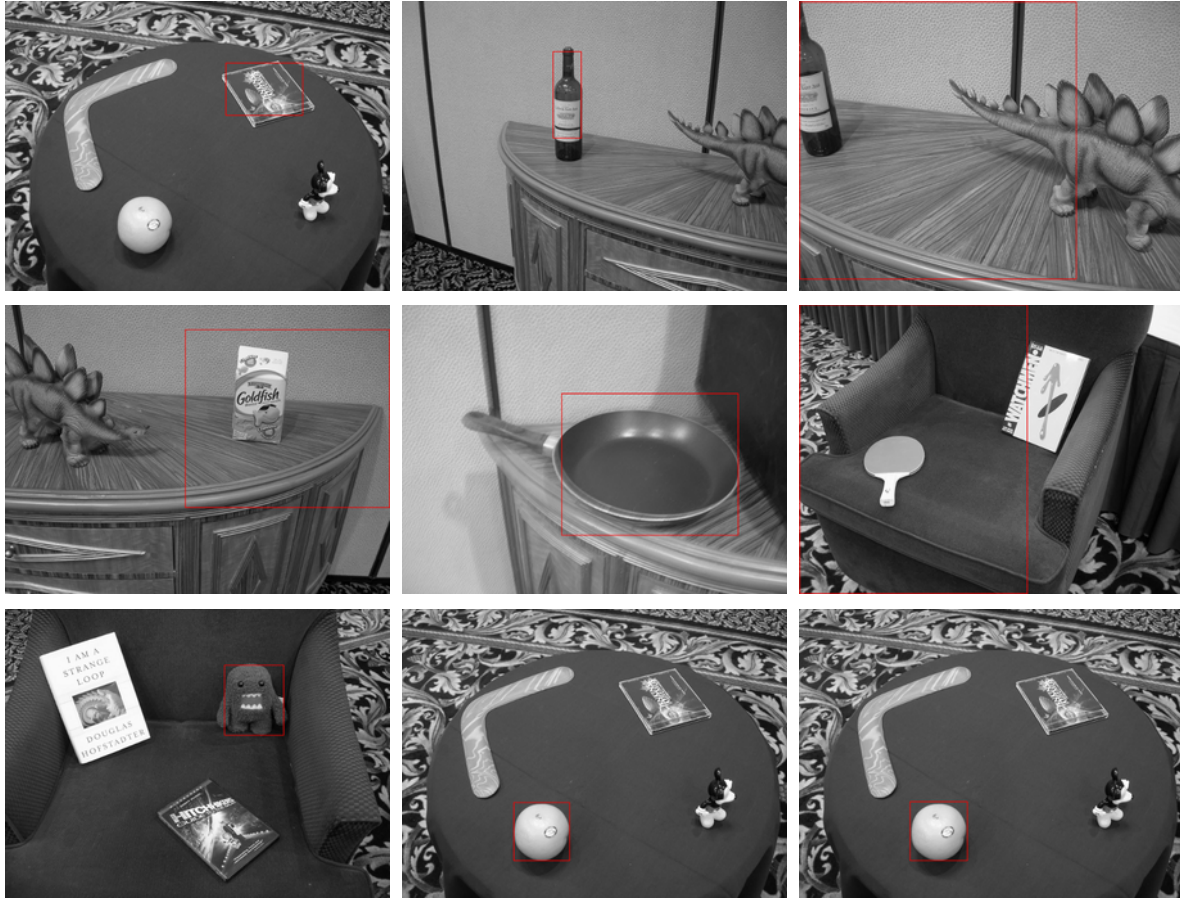


Figure 7. A sample of results from the 2009 SRVC contest. The first column shows correct detections for specific objects, the second column shows correct category objects, and the third column shows incorrect category guesses. From left to right and top to bottom, the system’s labels for the images are: 1) Karl Jenkins CD, 2) Bottle, 3) Laptop, 4) Goldfish Crackers, 5) Frying Pan, 6) Dinosaur, 7) Toy Domo Kun, 8) Orange, 9) White Soccer ball.

models for these categories were highly similar, and so as shown in the bottom row of figure 7, the category hypotheses were confused. With a proper scale prior, we would have easily recognized the soccer-ball and pumpkin, as we have found in later experiments.

These results support the observations made earlier in this document – that specific objects are well-recognized given good viewpoints and that the failures for object categories are often due to clutter effects which suggest unreasonable 3D regions or inter category confusion. In addition, the object category recognition techniques we utilized are not invariant to viewpoint, so categories such as laptop and table tennis racket are particularly challenging to recognize. Improvement in category recognition can additionally be improved by using spatial reasoning to filter incorrect hypotheses and allow the classifier’s lower scores on the

correct object to be returned more often. In general, these results demonstrate a remarkable ability to locate both specific and category objects within an unknown environment.

5 Conclusions

We have presented a robotic system, known as Curious George, that has demonstrated state-of-the-art performance on the task of recognizing objects in its environment. The ability to learn visual appearance models from Internet training data, and the wide variety of classification techniques used in our system provides for generalization to many object categories. The attention system and distributed architecture of our system allows the scene to be surveyed efficiently and for that visual survey to be labeled

with the present objects in a scalable fashion in relation to scene size.

We believe that continued efforts in robot object recognition will produce increasingly competent approaches. In particular, the integration of non-visual information that is available to a robot, such as proprioception and sensed 3D structure, have great potential to aid in the recognition process. Another promising area is the use of priors from higher level scene understanding, such as place recognition and surface reasoning, which can augment the object recognition process. For example, a robot system should be aware that a refrigerator is likely to occur in the kitchen, and so this object should likely not be reported in the bathroom.

In the near future, continued success in this domain will enable robots to perform a variety of object-centric tasks such as home assistance and food delivery. While still far from human-level, the visual understanding now possible by mobile robots has the potential for greatly enriching the lives of those in our society.

References

- [1] A. A. Treisman. Features and objects: the fourteenth bartlett memorial lecture. In *Quarterly Journal of Experimental Psychology*, volume 40A, pages 201–236, 1988.
- [2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge (VOC) Results. <http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html>, 2009.
- [4] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *In Proceedings of the IEEE CVPR*, 2008.
- [5] 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.
- [6] G. Grisetti, C. Stachniss, and W. Burgard. Improved techniques for grid mapping with rao-blackwellized particle filters. In *IEEE Transactions on Robotics*, 2006.
- [7] S. Helmer and D. G. Lowe. Using stereo for object recognition. In *Accepted to appear in the proceedings of the IEEE International Conference of Robotics and Automation (ICRA)*, 2010.
- [8] 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.
- [9] 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*, 20(11):1254–1259, Nov 1998.
- [10] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2003.
- [11] 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*, 108(1-2):196 – 203, 2007. Special Issue on Vision for Human-Computer Interaction.
- [12] 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*, 56(6):503–511, 2008.
- [13] Per-Erik Forssén, David 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.
- [14] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. B. Foote, J. Leibs, R. Wheeler, and A. Y. Ng. Ros: an open-source robot operating system. In *International Conference on Robotics and Automation*, Open-Source Software workshop, 2009.
- [15] 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*, 77:157–173, 2008.
- [16] R. B. Rusu, A. Holzbach, M. Beetz, and G. Bradski. Detecting and segmenting objects for mobile manipulation. In *ICCV, S3DV Workshop*, 2009.
- [17] M. Schlemmer, G. Biegelbauer, and M. Vincze. Rethinking robot vision - combining shape and appearance. *International Journal of Advanced Robotic Systems*, 4:259 – 270, 2007.
- [18] 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.
- [19] A. M. Treisman and G. Gelade. A feature-integration theory of attention. In *Cognitive Psychology*, volume 12, pages 97 – 136, 1980.
- [20] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.
- [21] B. Yamauchi. A frontier based approach for autonomous exploration. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, Monterey, CA, 1997.
- [22] Y. Ye and J. K. Tsotsos. Sensor planning for 3d object search. *Computer Vision and Image Understanding*, 73:145168, 1999.