

Curious George: The UBC Semantic Robot Vision System

Matthew Brown, Matthew Dockery, Scott Helmer, Mike Joya, James J. Little, David G. Lowe, Sancho McCann, David Meger, Marius Muja, Tristram Southey, Kevin Swersky, Pooja Viswanathan, Qingnan Zhou

Introduction

As an entrant in the 2008 Semantic Robot Vision Challenge (SRVC), UBC's Curious George robot will demonstrate that the combination of numerous state-of-the-art techniques from the current vision literature can produce a relatively robust visual search robot, with many of the competencies required, for example, for elder care. The SRVC task is to locate a set of unknown objects in an unknown environment, using only the World Wide Web as a source for specific training imagery. In 2007, Curious George placed first in the SRVC robot league, succeeding to fully or partially recognize 7 objects. The UBC team is committed to improving performance substantially for the 2008 contest, primarily by leveraging additional training information available on the Web, more sophisticated classifiers and by better exploiting the embodied vision system to gather cues for scene understanding. For example, semantic hierarchies of objects are available in various forms on the web. These can be used to filter inappropriate images and broaden visual representations by querying for images of closely related objects. Finally, active visual search can integrate recognition over frames as well as using appearance, structure and motion cues for mid-level vision tasks such as segmentation. The remainder of this abstract decomposes the SRVC into sub-tasks and describes our approach to each.

Training Data Acquisition

Our current acquisition system improves on the 2007 contest version by applying a more targeted approach for acquiring data about the appearance of objects from the web. Instead of just using Google Image Search as most teams did last year, we will also acquire images directly from vendor websites, such as Walmart. The images from these sites are of higher visual quality and are more accurately labelled than those from generic image search. Vendor sites also contain additional data such as size, weight and semantic hierarchical descriptions. Moreover, the images provide a mixture of clean product shots and examples of the product in use.

Utilizing images from on-line vendors is obviously not a replacement for Google Image Search since not all objects can be found on the these vendor sites. However, for the

2008 contest we plan to filter the results of each Google query in several ways. Our ideal set of training images is one where each photo is well focused, professionally lit, tightly framed, and does not have a distracting background. We have also found that it is important to avoid training object classifiers on non-photographic images, including cartoons, paintings, illustrations, computer rendered images and technical schematics, each of which may appear in a Google query. We have developed a system, based on the features proposed by Ke et al. in (Ke, Tang, & Jing 2006), to analyze the visual quality of an image and rate its suitability based on the above requirements. We will use this system to annotate Google Image Search results and focus training efforts on high quality images.

Object Appearance Learning

One weakness that the 2007 competition highlighted was that of detection of generic object classes with large intra-class variation. This year, we represent object appearance using a set of exemplars each in turn represented by a spatial pyramid description. This is similar to the work of (Chum & Zisserman 2007), with the exception that we intend to use a variety of feature types, including SIFT descriptors, contours, and colour. Moreover, to avoid having to run a sliding window over all test images at all locations and scales, we make use of as much contextual information as possible. The size of objects in a rough stereo segmentation can help constrain the search to portions of the image likely to actually contain the object of interest. We can then run the spatial pyramid detection as a cascade, using the coarsest (bag-of-features) level of the pyramid first, and progressing with more detailed levels of the pyramid only on regions that match coarser levels above a threshold.

One challenge with using the above approach is that it can be computationally intensive, and may not have high precision in some cases. As a result, we will also utilize SIFT matching using geometric constraints, similar to UBC's 2007 entrant. This approach typically has high precision for specific objects, such as a particular textbook, and is significantly faster than generic object classifiers. The intent here is that a matching produced by this approach is confident, and this can guide behaviour. With detection of a particular object, we can save computation by not running classifiers for this object on other images, and also inform

the visual search to not obtain additional views of this object.

Exploration and Search

The goal of the robot while it is searching the environment is to collect high resolution images of the objects, enabling recognition. We employ a hierarchical visual search strategy based initially on computationally inexpensive cues to focus recognition effort.

An initial geometric map is built in order to measure the extents of the contest environment and to determine likely object locations based on their structure and positioning. A Rao-Blackwellized particle filter integrates laser scans with odometry information as suggested by Montemerlo et al (Montemerlo *et al.* 2003), and frontier-based exploration (Yamauchi 1997) ensures all parts of the environment are included in the map.

Scene understanding processes visual appearance and dense stereo data to construct a 3-D geometric representation of the environment. In addition, potential object identification is performed so that detailed analysis can be focused towards promising regions of the environment. An essential step in this procedure is accurate figure-ground segmentation to separate an object from its surroundings. Depth and appearance information collected over numerous views are fused to estimate object boundaries and label 3-D structure.

Image data obtained from the internet often contains only single view of an object — often the canonical view as judged by human photographers — and the majority of object recognition algorithms will not successfully recognize the object from an image captured from a significantly different viewpoint. Our 2008 entry will address this issue more directly than was attempted in 2007: the robot executes a multiple viewpoint collection behaviour for each potential object, moving along a path that allows it to view the object from a variety of angles and collecting training data from each. Also, recent advances in 3-D object recognition such as (Toews & Arbel 2007; Savarese & Fei-Fei 2007) allow accurate recognition over a variety of viewpoints given suitable training data. The web imagery available for SRVC training lacks the variety of viewpoints needed to apply these methods directly, but we plan to adapt them to the extent possible into the 2008 SRVC entry.

Hardware System

The Curious George hardware system senses the world using two laser range finders and a directed peripheral-foveal vision system consisting of a binocular stereo camera and high-resolution digital still camera with zoom lens mounted on a pan-tilt unit. The second rangefinder has been added since 2007 and will be used as an obstacle avoidance sensor. This will allow our robot to avoid fairly small objects placed on the floor which otherwise create a dangerous situation for navigation. An onboard computer as well as several powerful laptops provide computation.

In 2007, our system was installed on a Powerbot robot base which provided excellent stability and power autonomy, but was somewhat limited in flexibility by its large size and weight. As of the writing of this abstract, the UBC team

is awaiting delivery of the much smaller Pioneer 3-AT platform, which will be used for the 2008 contest if it can be made ready in time. Please note that the Powerbot still appears in the Qualification Video, and that the final decision for robot platform is still pending.

Concluding remarks

The SRVC has been an excellent source of motivation and direction for numerous team members. Our involvement in last year's contest has led to numerous publications (Forssén *et al.* 2008; Meger *et al.* 2007; 2008) and our team is dedicated to making this year's entry a significant step towards robust embodied recognition.

References

- Chum, O., and Zisserman, A. 2007. An exemplar model for learning object classes. In *CVPR* (2007).
- Forssén, P.-E.; Meger, D.; Lai, K.; Helmer, S.; Little, J. J.; and Lowe, D. G. 2008. Informed visual search: Combining attention and object recognition. In *Proceedings of ICRA (to appear)*.
- Ke, Y.; Tang, X.; and Jing, F. 2006. The design of high-level features for photo quality assessment. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 419–426. Washington, DC, USA: IEEE Computer Society.
- Meger, D.; Forssén, P.-E.; Lai, K.; Helmer, S.; Sancho McCann, T. S.; Baumann, M.; Little, J. J.; Lowe, D. G.; and Dow, B. 2007. Curious george: An attentive semantic robot. In *IROS 2007 Workshop: From sensors to human spatial concepts*. San Diego, CA, USA: IEEE.
- Meger, D.; Forssén, P.-E.; Lai, K.; Helmer, S.; McCann, S.; Southey, T.; Baumann, M.; Little, J. J.; Lowe, D. G.; and Dow, B. 2008. Curious george: An attentive semantic robot. *Robotics and Autonomous Systems Journal*. Accepted.
- Montemerlo, M.; Thrun, S.; Koller, D.; and Wegbreit, B. 2003. 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)*, 1151–1156. Acapulco, Mexico: IJCAI.
- Savarese, S., and Fei-Fei, L. 2007. 3d generic object categorization, localization and pose estimation. In *In proceedings of ICCV*.
- Toews, M., and Arbel, T. 2007. Detecting and localizing 3d object classes using viewpoint invariant reference frames. In *In proceedings of ICCV 3DRR workshop*.
- Yamauchi, B. 1997. A frontier based approach for autonomous exploration. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*.