AAAI Fellowship Application

John Oberlin
Brown University, 2014

1 Dissertation Abstract

Robots have the potential to help people by making changes to the physical world. A robot assistant could improve someone's life by helping a disabled person prepare a meal, fetching a purse for a bedridden patient, or handing ointment to a busy parent whose hands are full taking care of their child. As a matter of course, if a robot is going to affect the world it must have an understanding of the positions of objects. Computer vision is an appealing route for obtaining this understanding, as RGB-D sensors are cheaper and more familiar than most alternatives.

State of the art techniques in object detection and pose estimation are powerful and general but usually run at a rate much less than 1 Hz and require time and expertise to build, maintain, and operate. The high demands of modern systems can make it difficult to employ such techniques in real-time human-robot interaction. Although effective solutions exist for category recognition, there is no off-the-shelf framework for object detection, segmentation and pose estimation that allows new objects to be quickly and easily added by non-experts.

Popular techniques for real-time detection use modified deformable part models (DPMs) [10] [5], and sometimes exploit different channels of data [6] [9]. These approaches have seen success in their target domains, but are too technical for general application and need to be integrated with interactive systems. A computer vision system that is simple, reliable, and easy to use with ROS (Robot Operating System) would benefit many researchers.

The aim of my dissertation at Brown is to enable a robot to autonomously scan 10 novel objects in order to construct robust models for detection, pose estimation, grasping, and manipulation of those objects during collaborations with a human operator. We propose the following workflow for this task: 1.) A human operator provides the robot with a box of objects. 2.) The robot picks up each object and scans it from many different perspectives to collect appearance data. 3.) The robot trains classifiers to recognize each object. 4.) The robot collects labels and metadata for the objects from Amazon Mechanical Turk. 5.) The robot can detect the objects in the environment and accurately respond to an operator's request to fetch an object or put it away.

Our prototype system detects and estimates poses of objects in RGB-D video taken with a Kinect and runs at a frequency of 2 Hz. Our detection framework calculates three quantities for each frame f which facilitate planning and reasoning.

The first quantity is a subset $\mathcal{G}_f$ of pixel locations, which induces a grid graph whose nodes correspond to points spaced $s = 5$ pixels apart in the input image f and whose edges connect each pixel to its four cardinal neighbors in the grid. A pixel g is included in $\mathcal{G}_f$ if the local average Objectness (see below) exceeds a threshold. We manually tune the threshold at the moment, but it would be natural to use a max-entropy approach such as herding [8]

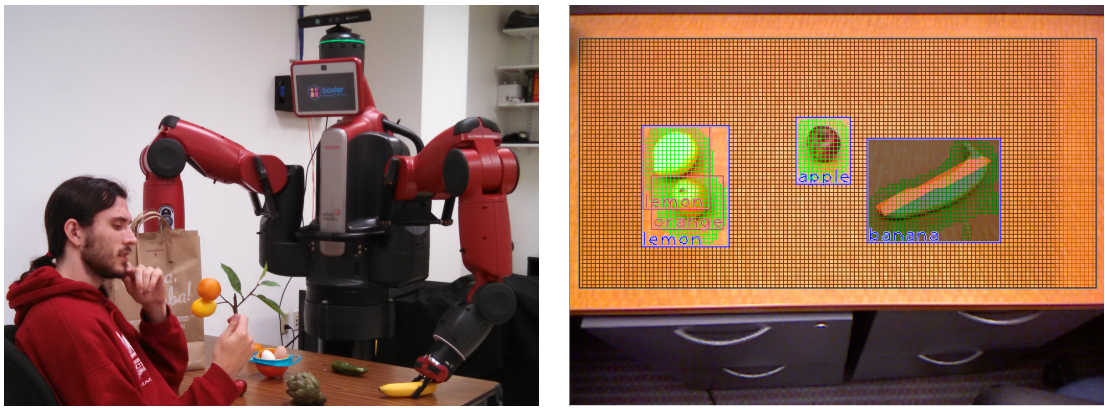

Figure 1.1: Teaching a robot to identify and manipulate objects can be as easy as bringing home a bag of groceries.

to adaptively control it. The second quantity is $\mathcal{B}_f$, a list of the connected components of the graph induced by $\mathcal{G}_f$. The third quantity is $\mathcal{R}_f$, a labeling which associates with each point $g \in \mathcal{G}_f$ a token corresponding to the object instance depicted at that pixel location.

In other words, $\mathcal{G}_f$ denotes areas in the image that probably belong to an object, $\mathcal{B}_f$ encodes clusters of pixels which likely correspond to individual objects, and $\mathcal{R}_f$ segments $\mathcal{G}_f$ into object instances, which may involve breaking apart the clusters in $\mathcal{B}_f$.

We determine $\mathcal{G}_f$ using an Objectness measure [1], extract features with SIFT [4] and k-means, and perform k-nearest neighbors classification on crops C_B and C_R induced by the bounding boxes of components in $\mathcal{B}_f$ and $\mathcal{R}_f$. We currently optimize the objective function for $\mathcal{R}_f$ using a variant of the wide-scale random noise algorithm in [7]. Our system publishes detections and pose estimates for the crops C_B and C_R on separate topics in ROS.

To provide accurate help to a human partner, the robot needs to know how to manipulate objects and understand their typical uses. Our system will address this issue by inferring object affordances and inducing an OO-MDP [2] that the robot can use for robust planning and object manipulation.

We might achieve better performance when determining $\mathcal{R}_f$ by using more sophisticated techniques that incorporate feedback with the robot's OO-MDP planner in order to collect additional, specific data as it is needed. Employing a POMDP [3] to explore the space of bounding boxes in an image is a natural approach to investigate.

With our prototype system, a computer vision researcher can train accurate models for 5 objects in 15 minutes. Our aim is to integrate a robot into the workflow so that a novice user can train 10 objects with 5 minutes of human input.

The framework we have outlined abstracts the detector from the classifier by taking a local notion of Objectness and extending it to a global model in a way that allows new detectors for local properties to be incorporated seamlessly and without down time. Our technology will enable robots to help humans by robustly sensing and manipulating the objects that people value most.

References

[1] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, *Bing: Binarized normed gradients for objectness estimation at 300fps*, in IEEE CVPR, 2014.

[2] C. Diuk, A. Cohen, and M. L. Littman, *An object-oriented representation for efficient reinforcement learning*, in Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 240–247.

[3] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, *Planning and acting in partially observable stochastic domains*, Artificial intelligence, 101 (1998), pp. 99–134.

[4] D. G. Lowe, *Distinctive image features from scale-invariant keypoints*, International journal of computer vision, 60 (2004), pp. 91–110.

[5] M. A. Sadeghi and D. Forsyth, *30hz object detection with dpm v5*, in Computer Vision–ECCV 2014, Springer, 2014, pp. 65–79.

[6] Y. Sun, L. Bo, and D. Fox, *Learning to identify new objects*, ICRA, (2014).

[7] P. Valiant, *Distribution free evolvability of polynomial functions over all convex loss functions*, in Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ACM, 2012, pp. 142–148.

[8] M. Welling, *Herding dynamical weights to learn*, in Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 1121–1128.

[9] S. S. J. Xiao, *Sliding shapes for 3d object detection in depth images*, ECCV, (2014).

[10] M. Zhu, N. Atanasov, G. J. Pappas, and K. Daniilidis, *Active deformable part models inference*, in Computer Vision–ECCV 2014, Springer, 2014, pp. 281–296.

2 Attendance Statement

Work in computer vision has largely centered around specific problems concerning single images. This focus has resulted in significant progress on such tasks, but the challenges of engineering real-time systems have so far prevented all but a handful of methods from spreading to other communities.

My aim is to bridge this gap by bringing state-of-the-art computer vision techniques to robotics and AI in order to create a robot that can engage in collaborative perception with its human partner and infer actions that are beneficial to the pair.

I recently joined a robotics lab, so despite my strong training in computer vision, there are still a few gaps in my background that I would like to fill. By attending AAAI I will be able to better provide computer vision capabilities in a way that is compatible with AI formalisms, researchers, and systems.

Recent advances in object detection on large data sets suggest that we are ready to move beyond attacking vision in an isolated setting and begin integrating it with a larger framework for planning in an interactive environment. The information available from extra sensors and the ability of real-time systems to capture additional images of the same scene will faciliate efficient, joint solutions to multiple related problems (such as detection, segmentation, and pose estimation).

When combining techniques and information, a well organized framework will help keep a system simple, efficient, and growing. To those ends I would learn more about knowledge representation and probabilistic reasoning. In particular, it is natural to apply planning techniques to the tasks of searching within an image and searching within a physical space.

Finally, a system which can maintain the attention of the operator stands a better chance of useful employment. I will pursue that goal by better understanding game playing, interactive entertainment, and general human-robot interactions.

3 Curriculum Vitae

Education

BS in Math, Florida State University, 2003-2006
MA in Math, UC Berkeley, 2006-2008
PhD program in Computer Science at U Chicago, 2010-2011
PhD program in Computer Science at Brown University, 2011-Present

Employment

Developer Support Engineer, Havok, 2008-2009

Conference Papers

S. Naderi Parizi, J. Oberlin, P. Felzenszwalb.
Reconfigurable Models for Scene Recognition.
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012

P. Felzenszwalb, J. Oberlin.
Multiscale Fields of Patterns.
To Appear, 2014

Conferences Attended

CVPR 2011
CVPR 2012

Teaching Experience

Being a TA in the UCB Math Deparment can be close to a full time teaching position, including quiz design, office hours, and grading in addition to the intense recitation sections (which more closely resembled lectures).

Calculus 1 TA
UC Berkeley, 2006-2007
Two sections in the Fall, three sections in the Spring, 30 students in each section, each section met with me two or three times a week for a total of about three hours a week.

Linear Algebra and ODE TA
UC Berkeley, 2007-2008
Same configuration as calculus.

Algorithms Grad TA
Brown University, 2012-Present

I am currently the Grad TA for the Algorithms class at Brown for the fourth semester. During this time the class has been between 45 and 100 students each iteration. I have been responsible for administering oral exams, holding office hours, lots of grading (most of our problems are proof based), and managing a group of at least 6 Undergraduate TA's each semester.

Departmental Service

Graduate Student Orientation Leader
Brown CS Department, Fall 2012

Computer Vision Reading Group Coordinator
Brown CS Department, 2012

Department Tea Organizer
Brown CS Department, 2012-Present

Misc Research

Master's Thesis in Mathematics
UC Berkeley, Written 2006-2008

Research Experience for Undergraduates in Mathematics
Oregon State University, Summer 2005

Undergraduate Research Program in Physics
Florida State University, Summer 2004

Research Assistant in Molecular Biology
Florida State University, Summer 2003

Hobbies

Gardening
Ballroom Dance
Martial Arts
Blacksmithing
3D Printing