AAAI Fellowship Application

John Oberlin
Brown University, 2014

1 Introduction

Imagine handing ten everyday objects to your robot assistant, who then proceeds to scan them and put them where they belong. Now think of a situation where a capable robot might improve someone's life by helping a disabled person prepare a meal, fetching a purse or storing a book for a bedridden patient, or handing ointment to a busy parent whose hands are full taking care of their child.

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-computer interaction. Additionally, although effective solutions exist for category level recognition, there is no decisive framework for handling instance level recognition.

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. An effective workflow for this task would be 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.

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.

2 Current Work, Future Progress and Broader Impact

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 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 locations spaced 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

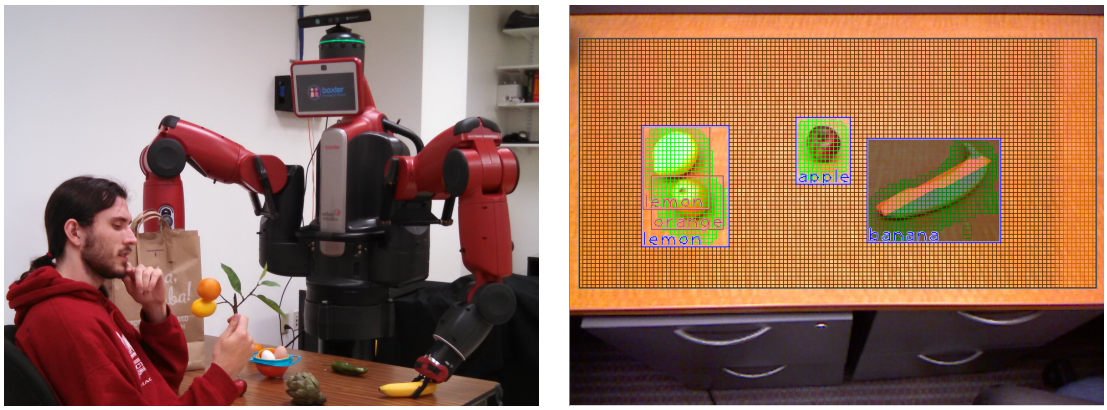

Figure 2.1: Teaching a robot to identify and manipulate objects can be as easy as bringing home a bag of groceries.

at the moment, but it would be natural to use a max-entropy approach such as herding [8] 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 in the image to which the point belongs.

In other words, $\mathcal{G}_f$ denotes areas in the image that probably belong to an object, $\mathcal{B}_f$ describes clusters of pixels that probably belong to the same object, 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 classification on crops C_B and C_R induced by $\mathcal{B}_f$ and $\mathcal{R}_f$ using k-nearest neighbors. We currently optimize the objective function for $\mathcal{R}_f$ using a variant of the wide-scale random noise algorithm [7].

Right now, detections and pose estimates are published in ROS for the C_B and C_R. 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 with the the Red Box optimization process 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. Using a POMDP [3] to explore the space of bounding boxes is a natural approach to investigate.

Using 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 system we have outlined abstracts the detector from the classifier, taking the notion of Objectness and extending it. As more sophisticated computer vision algorithms become tractable in real time or new data becomes available, they might be smoothly incorporated

into the system even as it is running. Automatic model validation can ensure that a new model outperforms the current model before the system switches it out with no interruption of service. This technology will enable robots to help people 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.

3 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 has so far prevented all but a handful of methods from spreading to other communities. At one time it was said that AI was "vision hard" and that solving vision would effectively solve AI. While that belief sweeps a bit under the rug, it is certainly true that in the past computer vision has been a strong bottleneck in the development of AI and robotics.

There are now many techniques in computer vision which are suited to fast and effective partial solutions of problems such as object detection but have been ignored in favor of much slower but slightly more effective state of the art approaches. Such techniques may be combined with the information available from extra sensors and the ability of real time systems to capture additional images of the same scene to form full solutions to multiple related problems (such as detection, segmentation, and pose estimation).

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 in a larger framework for planning in an interactive environment.

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.

4 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

5 Letter From Supervisor