STAIR: The STanford Artificial Intelligence Robot project

Andrew Y. Ng, Stephen Gould, Morgan Quigley,
Ashutosh Saxena and Eric Berger
Computer Science Department
Stanford University
Stanford CA 94305
{ang,sgould,mquigley,asaxena,eberger}@cs.stanford.edu

We describe an application of learning and probabilistic reasoning methods to the problem of having a robot fetch an item in response to a verbal request. Done on the STAIR (STanford AI Robot) platform, this work represents a small step towards our longer term goal of building general-purpose home assistant robots.

Having a robot usefully fetch items around a home or office requires that it be able to understand a spoken command to fetch an item, that it can navigate to the location of the object (including opening doors), find and recognize the object it is asked to fetch, understand the 3d structure and position of objects in the environment, and be able to figure out how to physically pick up the object from its current location, so as to bring it back to the person making the request. By tying together different algorithms for carrying out these tasks, we recently succeeded in having the robot fetch an item in response to a verbal request.

We describe below some of the key components integrated together to build this application.

Probabilistic multi-resolution maps. For a robot to navigate indoor environments and open doors, it must be able to reason about about maps on the scale of 10s of meters, as well as perform manipulation that is accurate to millimeters to use door-handles. Building on the work of [2], a unified probabilistic representation is described in [6] that allows our robot to coherently and simultaneously reason using both course grid-maps of a building discretized at 10cm intervals, and very accurate models of doors that are accurate to the 1mm level. The key idea is a single representation that allows us to compute the probability of any sensor measurement (laser scan reading) given a map that has both high- and low-resolution portions. This allows our robot to navigate indoors and open doors. By extending these ideas to use computer vision to recognize doors and doorhandles, we are further able to open and manipulate previously unknown doors—even ones of designs different from those in the training set—with 91% success rate.

Learning to grasp unknown objects. Even though robots today can carry out tasks as complex as assembling a car, most robots are hopeless when faced with novel objects and novel environments. However, the STAIR robot must be able to grasp even previously unknown objects, if it is to be able to fetch items that did not appear in the training set (such as a stapler, coffee mug, etc. of novel shape and/or appearance). Using sensors such as stereo vision, it is extremely difficult for a robot to build an accurate 3d model of an object that it is seeing for the first time, because of occlusion, etc. However, given even a single monocular image, it is possible to obtain a rough estimate of the 3d shape of a scene. [7, 1, 4] Using multiple monocular cues, [8] describes an algorithm for finding a strategy for grasping novel objects. By further incorporating a learning algorithm for selecting grasps even in the presence of nearby obstacles, our algorithms are often able to grasp objects even in the presence of nearby clutter.

Foveated vision. There are many reasons, such as context [9], that human object recognition is far superior to robotic vision. However, one reason that has been little exploited in computer object recognition is that humans use a fovea (a high resolution, central part of the retina) to obtain high resolution images of objects that they are trying to recognize. And, object recognition

is far easier from high resolution images than low resolution ones. For example, a coffee mug at 5m distance is a mere 8x8 pixels (using a standard webcam); it is no surprise that recognition from such tiny images is hard. When doing vision on a physical robot (as opposed to downloading images off the internet), this is easily rectified. Using a pan-tilt-zoom camera that can zoom into any part of the image, we can easily obtain high resolution images of objects we are trying to recognize. Using a learned strategy [3] for choosing what part of an image to zoom into so as to maximize expected information gain, this immediately gives a significant increase to the accuracy of object recognition on STAIR.

These components are further driven by an MDP-based spoken dialog system (see also [5]). A video of the robot fetching an item is at

http://stair.stanford.edu/multimedia.php .

The robot begins its task when its spoken dialog system understands the spoken command to fetch a stapler. It navigates to the lab area from where it is asked to fetch the stapler, and uses a steerable fovea (a pan-tilt-zoom camera that can turn and "look around") to find the object. Lastly, despite never having been trained before to grasp a stapler, it uses its learned grasping strategy to correctly pick up the stapler, and bring it back.

Since the birth of AI in 1956, the AI dream has been to build systems that exhibit broad-spectrum competence and intelligence. Today, AI has, however, fragmented into many different sub-fields. By developing and integrating onto a single robot platform tools drawn from all areas of AI including learning, vision, navigation, manipulation, planning, and speech/NLP, STAIR represents our attempt to revisit the AI dream. Over the long term, we envision a single robot that can perform tasks such as tidying up a room, using a dishwasher, fetching and delivering items, and preparing meals.

Preference: Oral
Presenting author: Andrew Ng
Category: Learning Algorithms

References

[1] Erick Delage, Honglak Lee, and Andrew Y. Ng. Automatic single-image 3d reconstructions of indoor manhattan world scenes. In *Proc. International Symposium on Robotics Research (ISRR)*, 2005.

[2] A. Eliazar and R. Parr. DP-SLAM 2.0. In *Proceedings of the International Conference on Robotics and Automation*, 2004.

[3] Stephen Gould, Joakim Arfvidsson, Adrian Kaehler, Benjamin Sapp, Marius Meissner, Gary Bradski, Paul Baumstarck, Sukwon Chung, and Andrew Y. Ng. Peripheral-foveal vision for real-time object recognition and tracking in video. In *IJCAI*, 2007.

[4] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005.

[5] Filip Krsmanovic, Curtis Spencer, Daniel Jurafsky, and Andrew Y. Ng. Have we met? MDP based speaker ID for robot dialogue. In *Proceedings of the Ninth International Conference on Spoken Language Processing (InterSpeech–ICSLP)*, 2006.

[6] Anya Petrovskaya and Andrew Y. Ng. Probabilistic mobile manipulation in dynamic environments, with application to opening doors. In *IJCAI*, 2007.

[7] Ashutosh Saxena, Sung Chung, and Andrew Y. Ng. Learning depth from single monocular images. In *NIPS 18*, 2005.

[8] Ashutosh Saxena, Justin Driemeyer, Justin Kearns, Chioma Osondu, and Andrew Y. Ng. Learning to grasp novel objects using vision. In *Int'l Symposium on Experimental Robotics (ISER)*, 2006.

[9] Antonio Torralba, Kevin Murphy, and William Freeman. Contextual models for object detection using boosted random fields. In *NIPS 17*, 2004.