New algorithm could enable household robots to better identify objects in cluttered environments

January 12 2015, by Larry Hardesty

Credit: Christine Daniloff and Jose-Luis Olivares/MIT

For household robots ever to be practical, they'll need to be able to recognize the objects they're supposed to manipulate. But while object recognition is one of the most widely studied topics in artificial intelligence, even the best object detectors still fail much of the time.

Researchers at MIT's Computer Science and Artificial Intelligence Laboratory believe that household robots should take advantage of their mobility and their relatively static environments to make [object recognition](#) easier, by imaging objects from multiple perspectives before making judgments about their identity. Matching up the objects depicted in the different images, however, poses its own computational challenges.

In a paper appearing in a forthcoming issue of the *International Journal of Robotics Research*, the MIT researchers show that a system using an off-the-shelf algorithm to aggregate different perspectives can recognize four times as many objects as one that uses a single perspective, while reducing the number of misidentifications.

They then present a new algorithm that is just as accurate but that, in some cases, is 10 times as fast, making it much more practical for real-time deployment with household robots.

"If you just took the output of looking at it from one viewpoint, there's a lot of stuff that might be missing, or it might be the angle of illumination or something blocking the object that causes a systematic error in the detector," says Lawson Wong, a graduate student in electrical engineering and [computer science](#) and lead author on the new paper. "One way around that is just to move around and go to a different viewpoint."

First stab

Wong and his thesis advisors—Leslie Kaelbling, the Panasonic Professor of Computer Science and Engineering, and Tomás Lozano-Pérez, the School of Engineering Professor of Teaching Excellence—considered scenarios in which they had 20 to 30 different images of household objects clustered together on a table. In several of the scenarios, the

clusters included multiple instances of the same object, closely packed together, which makes the task of matching different perspectives more difficult.

The first algorithm they tried was developed for tracking systems such as radar, which must also determine whether objects imaged at different times are in fact the same. "It's been around for decades," Wong says. "And there's a good reason for that, which is that it really works well. It's the first thing that most people think of."

For each pair of successive images, the algorithm generates multiple hypotheses about which objects in one correspond to which objects in the other. The problem is that the number of hypotheses compounds as new perspectives are added. To keep the calculation manageable, the algorithm discards all but its top hypotheses at each step. Even so, sorting through them all, after the last hypothesis has been generated, is a time-consuming task.

Representative sampling

In hopes of arriving at a more efficient algorithm, the MIT researchers adopted a different approach. Their algorithm doesn't discard any of the hypotheses it generates across successive images, but it doesn't attempt to canvass them all, either. Instead, it samples from them at random. Since there's significant overlap between different hypotheses, an adequate number of samples will generally yield consensus on the correspondences between the objects in any two successive images.

To keep the required number of samples low, the researchers adopted a simplified technique for evaluating hypotheses. Suppose that the algorithm has identified three objects from one perspective and four from another. The most mathematically precise way to compare hypotheses would be to consider every possible set of matches between

the two groups of objects: the set that matches objects 1, 2, and 3 in the first view to objects 1, 2, and 3 in the second; the set that matches objects 1, 2, and 3 in the first to objects 1, 2, and 4 in the second; the set that matches objects 1, 2, and 3 in the first view to objects 1, 3, and 4 in the second, and so on. In this case, if you include the possibilities that the detector has made an error and that some objects are occluded from some views, that approach would yield 304 different sets of matches.

Instead, the researchers' algorithm considers each object in the first group separately and evaluates its likelihood of mapping onto an object in the second group. So object 1 in the first group could map onto objects 1, 2, 3, or 4 in the second, as could object 2, and so on. Again, with the possibilities of error and occlusion factored in, this approach requires only 20 comparisons.

It does, however, open the door to nonsensical results. The algorithm could conclude that the most likely match for object 3 in the second group is object 3 in the first—and it could also conclude that the most likely match for object 4 in the second group is object 3 in the first. So the researchers' algorithm also looks for such double mappings and re-evaluates them. That takes extra time, but not nearly as much as considering aggregate mappings would. In this case, the algorithm would perform 32 comparisons—more than 20, but significantly less than 304.

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Source: Massachusetts Institute of Technology