Teaching computers to see—by learning to see like computers

September 19 2013, by Larry Hardesty

With each of the raw images of the photos in color, today's state-of-the-art object-detection algorithms make errors — such as identifying a car (above) — that initially seem baffling. A new technique enables the visualization of a common mathematical representation of images (in black and white), which should help researchers understand why their algorithms fail. Credit: Courtesy of the researchers

Object-recognition systems—software that tries to identify objects in

digital images—typically rely on machine learning. They comb through databases of previously labeled images and look for combinations of visual features that seem to correlate with particular objects. Then, when presented with a new image, they try to determine whether it contains one of the previously identified combinations of features.

Even the best object-recognition systems, however, succeed only around 30 or 40 percent of the time—and their failures can be totally mystifying. Researchers are divided in their explanations: Are the learning algorithms themselves to blame? Or are they being applied to the wrong types of features? Or—the "big-data" explanation—do the systems just need more training data?

To attempt to answer these and related questions, researchers at MIT's Computer Science and Artificial Intelligence Laboratory have created a system that, in effect, allows humans to see the world the way an object-recognition system does. The system takes an ordinary image, translates it into the mathematical representation used by an object-recognition system and then, using inventive new algorithms, translates it back into a conventional image.

In a paper to be presented at the upcoming International Conference on Computer Vision, the researchers report that, when presented with the retranslation of a translation, human volunteers make classification errors that are very similar to those made by computers. That suggests that the learning algorithms are just fine, and throwing more data at the problem won't help; it's the feature selection that's the culprit. The researchers are hopeful that, in addition to identifying the problem, their system will also help solve it, by letting their colleagues reason more intuitively about the consequences of particular feature decisions.

Today, the feature set most widely used in computer-vision research is called the histogram of oriented gradients, or HOG (hence the name of

the MIT researchers' system: HOGgles). HOG first breaks an image into square chunks, usually eight pixels by eight pixels. Then, for each square, it identifies a "gradient," or change in color or shade from one region to another. It characterizes the gradient according to 32 distinct variables, such as its orientation—vertical, horizontal or diagonal, for example—and the sharpness of the transition—whether it changes color suddenly or gradually.

Credit: Researchers

Thirty-two variables for each square translates to thousands of variables

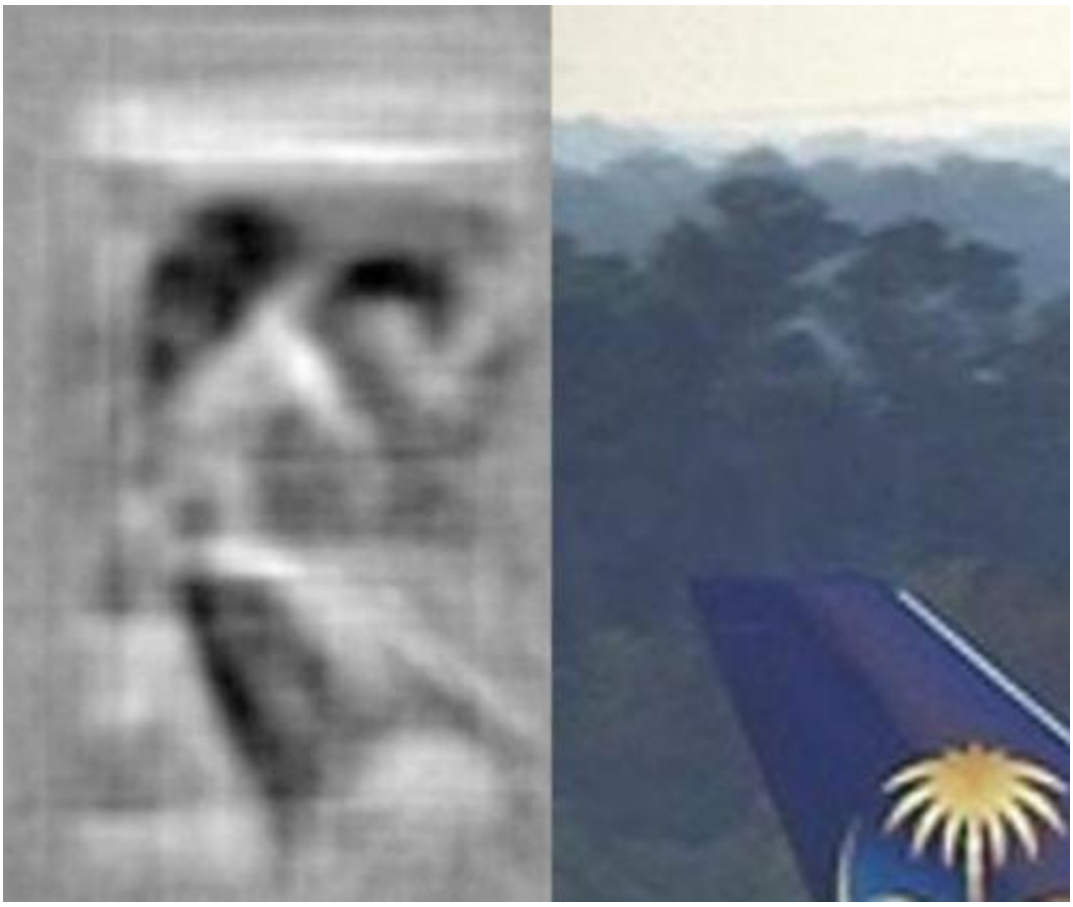
for a single image, which define a space with thousands of dimensions. Any conceivable image can be characterized as a single point in that space, and most object-recognition systems try to identify patterns in the collections of points that correspond with particular objects.

"This feature space, HOG, is very complex," says Carl Vondrick, an MIT graduate student in electrical engineering and computer science and first author on the new paper. "A bunch of researchers sat down and tried to engineer, 'What's the best feature space we can have?' It's very highly dimensional. It's almost impossible for a human to comprehend intuitively what's going on. So what we've done is built a way to visualize this space."

Vondrick; his advisor, Antonio Torralba, an associate professor of electrical engineering and computer science; and two other researchers in Torralba's group, graduate student Aditya Khosla and postdoc Tomasz Malisiewicz, experimented with several different algorithms for converting points in HOG space back into ordinary images. One of those algorithms, which didn't turn out to be the most reliable, nonetheless offers a fairly intuitive understanding of the process.

The algorithm first produces a HOG for an image and then scours a database for images that match it—on a very weak understanding of the word "match."

"Because it's a weak detector, you won't find very good matches," Vondrick explains. "But if you average all the top ones together, you actually get a fairly good reconstruction. Even though each detection is wrong, each one still captures the statistics of the original image patch."



Credit: Researchers

The reconstruction algorithm that ended up proving the most reliable is more complex. It uses a so-called "dictionary," a technique that's increasingly popular in computer-vision research. The dictionary consists of a large group of HOGs with fairly regular properties: One, for instance, might have a top half that's all diagonal gradients running bottom left to upper right, while the bottom half is all horizontal gradients; another might have gradients that rotate slowly as you move from left to right across each row of squares. But any given HOG can be represented as a weighted combination of these dictionary "atoms."

The researchers' algorithm assembled the dictionary by analyzing

thousands of images downloaded from the Internet and settled on the dictionary that allowed it to reconstruct the HOG for each of them with, on average, the fewest atoms. The trick is that, for each atom in the dictionary, the algorithm also learned the ordinary image that corresponds to it. So for an arbitrary HOG, it can apply the same weights to the ordinary images that it does to the dictionary atoms, producing a composite image.

Those composites are quite striking. What appears to be a blurry image of a woman sitting at a vanity mirror, for instance, turns out to be a reconstruction of the HOG produced by a photo of an airplane sailing over a forest canopy. And, indeed, a standard object-recognition system will, erroneously, identify a person in the image of the plane. It's a mistake that's baffling without the elucidation offered by the HOGgles.

To quantify the intuition that, given the representations of images in HOG space, object detectors' false positives are not as bizarre as they initially seem, the MIT researchers presented collections of their HOG reconstructions to volunteers recruited through Amazon's Mechanical Turk crowdsourcing service. The volunteers were slightly better than machine-learning algorithms at identifying the objects depicted in the reconstructions, but only slightly—nowhere near the disparity of 60 or 70 percent when object detectors and humans are asked to identify objects in the raw images. And the dropoff in accuracy as the volunteers moved from the easiest cases to the more difficult ones mirrored that of the object detectors.

The paper is titled "Inverting and visualizing features for object detection."

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

Provided by Massachusetts Institute of Technology