Techniques from natural-language processing enable computers to efficiently search video for actions

May 13 2014, by Larry Hardesty

Jose-Luis Olivares/MIT Credit: video stills from Nesnad/Wikimedia Commons

With the commodification of digital cameras, digital video has become so easy to produce that human beings can have trouble keeping up with it. Among the tools that computer scientists are developing to make the profusion of video more useful are algorithms for activity recognition—or determining what the people on camera are doing when.

At the Conference on Computer Vision and Pattern Recognition in June, Hamed Pirsiavash, a postdoc at MIT, and his former thesis advisor, Deva Ramanan of the University of California at Irvine, will present a new activity-recognition algorithm that has several advantages over its predecessors.

One is that the algorithm's execution time scales linearly with the size of the video file it's searching. That means that if one file is 10 times the size of another, the new algorithm will take 10 times as long to search it—not 1,000 times as long, as some earlier algorithms would.

Another is that the algorithm is able to make good guesses about partially completed actions, so it can handle streaming video. Partway through an action, it will issue a probability that the action is of the type that it's looking for. It may revise that probability as the video continues, but it doesn't have to wait until the action is complete to assess it.

Finally, the amount of memory the algorithm requires is fixed, regardless of how many frames of video it's already reviewed. That means that, unlike many of its predecessors, it can handle video streams of any length (or files of any size).

The grammar of action

Enabling all of these advances is the appropriation of a type of algorithm used in natural language processing, the computer science discipline that seeks techniques for interpreting sentences written in natural language.

"One of the challenging problems they try to solve is, if you have a sentence, you want to basically parse the sentence, saying what is the subject, what is the verb, what is the adverb," Pirsiavash says. "We see an analogy here, which is, if you have a complex action—like making tea or making coffee—that has some subactions, we can basically stitch

together these subactions and look at each one as something like verb, adjective, and adverb."

On that analogy, the rules defining relationships between subactions are like rules of grammar. When you make tea, for instance, it doesn't matter whether you first put the teabag in the cup or put the kettle on the stove. But it's essential that you put the kettle on the stove before pouring the water into the cup. Similarly, in a given language, it could be the case that nouns can either precede or follow verbs, but that adjectives must always precede nouns.

For any given action, Pirsiavash and Ramanan's algorithm must thus learn a new "grammar." And the mechanism that it uses is the one that many natural-language-processing systems rely on: machine learning. Pirsiavash and Ramanan feed their algorithm training examples of videos depicting a particular action, and specify the number of subactions that the algorithm should look for. But they don't give it any information about what those subactions are, or what the transitions between them look like.

Pruning possibilities

The rules relating subactions are the key to the algorithm's efficiency. As a video plays, the algorithm constructs a set of hypotheses about which subactions are being depicted where, and it ranks them according to probability. It can't limit itself to a single hypothesis, as each new frame could require it to revise its probabilities. But it can eliminate hypotheses that don't conform to its grammatical rules, which dramatically limits the number of possibilities it has to canvass.

The researchers tested their algorithm on eight different types of athletic endeavor—such as weightlifting and bowling—with training videos culled from YouTube. They found that, according to metrics standard in

the field of computer vision, their algorithm identified new instances of the same activities more accurately than its predecessors.

Pirsiavash is particularly interested in possible medical applications of action detection. The proper execution of physical-therapy exercises, for instance, could have a grammar that's distinct from improper execution; similarly, the return of motor function in patients with neurological damage could be identified by its unique grammar. Action-detection algorithms could also help determine whether, for instance, elderly patients remembered to take their medication—and issue alerts if they didn't.

More information: PAPER: "Parsing videos of actions with segmental grammars" people.csail.mit.edu/hpirsiav/ … s/grammar_cvpr14.pdf

Provided by Massachusetts Institute of Technology