New research teaches AI how people move with internet videos

September 14 2020, by Steve Crang

A cook prepares a dish in a user-uploaded video on the left, and the U-M team's neural network model identifies his position in each frame as represented by a 3-D mesh on the right. Credit: University of Michigan

New research at the University of Michigan can train neural network models to identify a person's position in videos where only a portion of their body is visible in the shot. This breakthrough opens up a huge library of video content to a new use—teaching machines the meaning behind people's poses, as well as the different ways they interact with their environment.

When we see a picture of a pair of hands typing at a desk, we can infer there's a person attached to those somewhere out of sight. Depending on the angle of their wrists and how tall the desk is, we might even be able to tell whether they're sitting or standing. These are the kinds of inferences neural network models haven't been great at to this point. If we ever want robots and AI that can tell what we're doing just by looking at us, they'll need a deeper understanding of how our bodies are positioned in different contexts and what different movements usually mean.

A field of study called human pose estimation focuses on teaching AI to identify how a person in an image or video is positioned, eventually enabling them to model those human positions independently. But current models are typically trained on very sanitized libraries of videos with an entire person fully in view, and perform very poorly on shots with only part of a person visible.

On top of that, the videos used are labeled—essentially, the videos come with the correct solution provided so that a neural network is able to check its work. In this case, that requires human workers to explicitly label the location of a displayed person's joints.

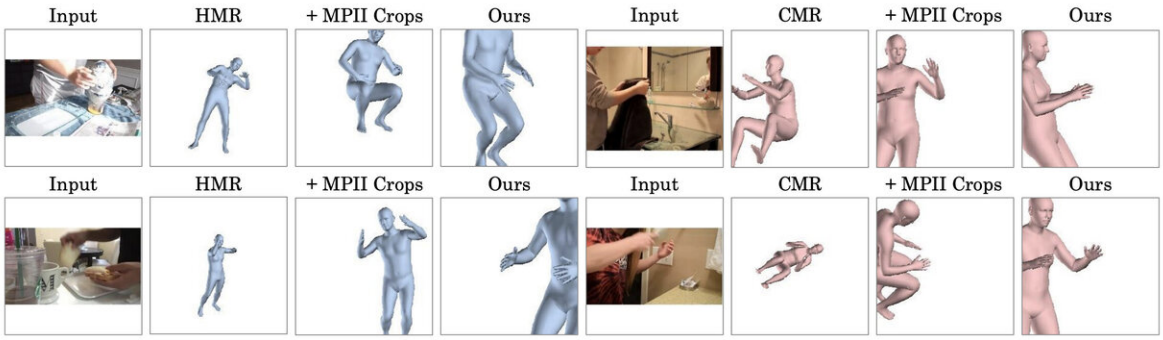
Prof. David Fouhey and Ph.D. student Chris Rockwell saw a problem with this practice—in the huge libraries of video content uploaded to public websites, only around 4% ever show an entire person, head to toe, squarely in the shot. And, of course, none of them come with labels. This means a whole world of video is virtually closed off to these old models, and new video would have to be tailor-made and manually labeled for neural networks to succeed.

"These datasets are somewhat normalized, where a person is of a certain height and they're in the center of the image," Rockwell says. "Stuff on the internet doesn't actually look like that."

To overcome these limitations, Rockwell and Fouhey came up with a pair of solutions, presented in a paper at the 2020 European Conference on Computer Vision. The techniques ultimately lead to two major breakthroughs in video training for neural network models: enabling models to make good predictions with only part of a person visible, and self-training of these models on unlabeled videos, wherein the model can smartly pick out good guesses without being told the solution.

First, they had to solve the issue of trained neural networks' poor performance on these truncated images of people. To do this, they took the intuitive step of cropping the networks' earlier training set to look more like the videos you'd find online. The duo took videos of a full-body activity and reduced them to just the torso, head, or arms. The existing models were re-trained on these crops, and could then produce more reasonable output with new data from internet videos.

The researchers point out that this technique's straightforward design makes the approach more flexible for widespread use.



A comparison of what trained neural nets see given a video with only partial view of the person as input. HMR and CMR, the original models, perform poorly; the output is closer to the target after initial training on the cropped dataset (labeled "+ MPII Crops"); and the final output (labeled "Ours") shows the

results of the model's self-training on the video. Credit: University of Michigan

"The method is deliberately simple so that it's as easy as possible to apply to multiple different methods," says Fouhey. "As work in more standard pose estimation progresses, it can be pretty easily adapted to Internet video with the use of this method."

The results were similarly intuitive—in experiments with two human 3-D mesh recovery techniques as comparison, their model gave significantly better approximations of the people's positions. Whereas before the results were often more or less random—a typical outcome when a neural net has low confidence in what it's seeing—the predictions made by the newly trained model gave a clear indication it could tell where the person was and roughly where their hands were.

Their second technique addressed the need for a neural network to be able to "train itself" on its own predictions, eliminating the need for people to label more videos. Instead, the model can make predictions on the person's position in the video, distinguish between the good and bad predictions, and label the video with the good. But in order to work, the model needs a way to decide which predictions to keep.

To do this, typically methods use "model confidence." When training a model to categorize objects in an image, for example, the limited set of categories provides an easy way for the model to express confidence by simply giving each object in the image a probability—80% chance this is a person, 40% this is a chair, and so on. When determining things like the angle of a person's elbow, it's harder to give a clear indication that the model knows what it's seeing.

To overcome this, the team adapted prior work in the field that involved

training the model on several iterations of each frame where each version was shifted slightly in different directions. The advantage of this arises from the same quality of neural nets—when they aren't confident, they're often very wrong, so predictions on a low-confidence frame can change drastically even with these small shifts. By gathering predictions from several near-identical frames, the researchers could aggregate confidence by comparing how similar the guesses were.

"If it knows what's going on the predictions will be really consistent," Rockwell explains. "We can identify some subset of the images that are pretty good and throw out a lot of the bad ones."

In the end, their models could determine automatically whether they had high confidence in their predictions about a video, and discard the worst guesses. Their combined methods allow a model to self-train on video frames without needing to label them.

Moving forward, Rockwell sees a lot of potential for continuing this work. Beyond improving the models' predictions, they can learn how to identify objects in the frame, guess what the person is doing with the object, and identify the object's size—ultimately allowing for a much greater contextual understanding of an environment through video. This could open up a number of other learning and identification tasks with video, as well.

"Modeling people is a step towards understanding them," says Rockwell, "and before this it was really tough to understand people in consumer videos. With these techniques we can much more readily recognize them."

More information: Rockwell et al., Full-Body Awareness from Partial Observations. arXiv:2008.06046 [cs.CV]. arxiv.org/abs/2008.06046

Provided by University of Michigan