Novel optimization tool allows for better video motion estimation

October 10 2023, by Tom Fleischman

We present a new method for estimating full-length motion trajectories for every pixel in every frame of a video, as illustrated by the motion paths shown above. For clarity, we only show sparse trajectories for foreground objects, though our method computes motion for all pixels. Our method yields accurate, coherent long-range motion even for fast-moving objects, and robustly tracks through occlusions as shown in the dog and swing examples. For context, in the second row we depict the moving object at different moments in time. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2306.05422

Cornell researchers have developed a new optimization tool to estimate motion throughout an input video, which has potential applications in video editing and generative AI video creation.

The tool, called OmniMotion, is described in a paper, "Tracking Everything, Everywhere, All at Once," presented at the International Conference on Computer Vision, Oct. 2–6 in Paris.

"There are these two dominant paradigms in motion estimation—optical flow, which is dense but short range, and feature tracking, which is sparse but long range," said Noah Snavely, associate professor of computer science at Cornell Tech and in the Cornell Ann S. Bowers College of Computing and Information Science. "Our method allows us to have both dense and long-range tracking across time."

OmniMotion uses what the researchers term "a quasi-3D representation"—a relaxed form of 3D that retains important properties (such as tracking pixels when they pass behind other objects) without the challenges of dynamic 3D reconstruction.

"We found a way to basically have it estimate more qualitative 3D," Snavely said. "It's saying, 'I don't know exactly where these two objects are in 3D space, but I know that this one is in front of that one.' You can't look at it as a 3D model, as things will be distorted, but it captures the ordering relationships between objects."

The new method takes a small sample of frames and motion estimates to create a complete motion representation for the entire video. Once optimized, the representation can be queried with any pixel in any frame to produce a smooth, accurate motion trajectory across the full video.

This would be useful, Snavely said, when incorporating computer-generated imagery, or CGI, into video-editing.

"If I want to place an object—say a sticker—on a video, then I need to know where it should be in every frame," he said. "So I place it in the first frame of the video; to avoid having to edit every subsequent frame

in a painstaking way, it'd be nice if I could just track where it should be in every frame—as well as if it shouldn't be there, if there's something occluding it."

OmniMotion could also help inform algorithms in generative text-to-video applications, Snavely said.

"Often these text-to-video models aren't very coherent," he said. "Objects will change size over the course of the video, or people move in uncanny ways, and that's because they're just generating the raw pixels of a video. They don't have any notion of the underlying dynamics that would result in pixel motion.

"We're hoping that by providing algorithms for estimating motion in videos, we can help improve the motion coherence of generated videos," he said.

Qianqian Wang, a postdoctoral researcher at the University of California, Berkeley, and a research scientist at Google Research, was lead author. Other co-authors were Bharath Hariharan, assistant professor of computer science at Cornell Bowers CIS; doctoral students Yen-Yu Chang and Ruojin Cai; and Aleksander Holynski, postdoctoral researcher at Berkeley and a scientist at Google Research; and Zhengqi Li of Google Research.

Also at the conference, Cai presented "Doppelgangers: Learning to Disambiguate Images of Similar Structures," which uses a massive dataset of image pairs to train computer vision applications to distinguish between images that look the same but are not, like different sides of a clock tower or building.

For Doppelgangers, Snavely and his team show how to use existing image annotations stored in the Wikimedia Commons image database to

automatically create a large set of labeled image pairs of 3D surfaces.

Doppelgangers comprises a collection of internet photos of landmarks and cultural sites that exhibit repeated patterns and symmetrical structures. The dataset includes a large number of image pairs—each labeled as either positive or negative matching pairs.

"Big Ben or the Eiffel Tower—they kind of look the same from different sides," Snavely said. "Computer vision just isn't good enough to tell the sides apart. So we invented a method to help tell when two things look similar but are different, and when two things really are the same."

In Doppelgangers, a neural network is trained to assess the spatial distribution of key points in an image, to differentiate between pairs of images that look similar but are different—like two different faces of Big Ben—from images of actual identical scene content. This would be useful in 3D reconstruction technology, Snavely said.

"The network likely learns things like whether the backgrounds are the same or different, or if there are other details that differentiate them," he said. "Then it outputs a probability: Are these really matching, or do they just look like they're matching? Then we can integrate that with 3D reconstruction pipelines to make better models."

More information: Qianqian Wang et al, Tracking Everything Everywhere All at Once, *arXiv* (2023). DOI: 10.48550/arxiv.2306.05422

Ruojin Cai et al, Doppelgangers: Learning to Disambiguate Images of Similar Structures, *arXiv* (2023). DOI: 10.48550/arxiv.2309.02420

Provided by Cornell University