Wang, Zheng (404855295)

TA: Guangyu Zhou

COM SCI 35L

Nov. 25th, 2018

Turning 2D Video into 3D Motion Sculpture—MoSculp Method

Ever since the maturity of video recording technology, this technology has in many ways changed the way mankind interpret motion—from professional athletes who try to learn counterstrokes of their opponents to high school students who attempt to teach themselves tennis, they all have tried to interpret the motion recorded in the video. It is not hard to justify this trend —in comparison to the classical way of interpreting motion by mimicking antecedents, the video recording appears to be a much more economical, approachable, and reproducible way. Unfortunately, video recordings also have several fatal drawbacks that easily overshadows the halo of video recordings—first of all, the videos are two dimensional while the motion actually happens in 3D space. Thus, it is inevitable that we will lose some details of the actual motion. Secondly, when watching a video, it is hard to see the accurate path of the motion from the video as things happen too fast and we cannot easily memorize what happens.

In order to battle those difficulties that stop us from interpreting motion through video, mankind has developed many different approaches, including stroboscopic photography, shape-time photography and MoSculp—a newly developed method of interpreting motion in a video by turning a 2D video to a 3D motion sculpture. In this review, we will be focusing on the advantages and working process of this innovative method—MoSculp.

Before we delve into the discussion of MoSculp, it is helpful to briefly introduce the conventional way of interpreting motion in the video—stroboscopic photography (*fig 1*) and shape-time photography (*fig 2*). Both of the methods overlay the moving part of the video on the same picture to demonstrate the process of motion over time. Nevertheless, they still fail to give us a perspective of how motion happens in 3D space and may produce observable artifacts. Also, these methods may need the help of special video-capturing equipment and thus make the interpretation not as economical and approachable as conventional video. Thus, with the ability to overcome these weaknesses, MoSculp should be one of the most effective ways to interpret motions in videos.

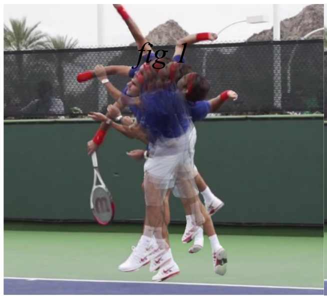

fig 1

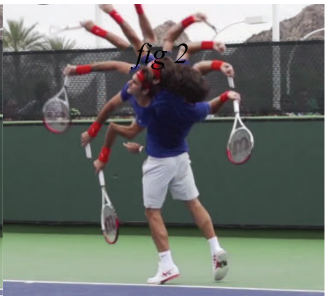

fig 2

The MoSculp method takes mainly four steps to give the ultimate motion sculpture—2D key points detection, reconstructing the 3D model over time, generating 3D sculpture, and rendering the sculpture back to the video and refinement. We will first discuss how MoSculp does key point detection and how key point detection helps MoSculp to reconstruct the 3D motion sculpture.

This part of the calculation is done automatically by a modified version of OpenPose, a standard method of locating joints of a person in a picture. OpenPose automatically records the joint position of the person inside a picture and record the location in 2D space (*fig 3*), which is referred to as the "pose" of the character inside a figure. However, as mentioned above, OpenPose needs to be modified since it is designed for static images and might produce errors when left and right sides are flipped during motion. Thus, the MoSculp team decide also to check the temporal coherency of the pose by penalizing the predictions that result in abrupt pose changes. After repeating the key point detection of each frame, we can obtain the poses of the figure inside each frame of the video and the changes of the pose over time. This information will be essential for the next step of MoSculp.

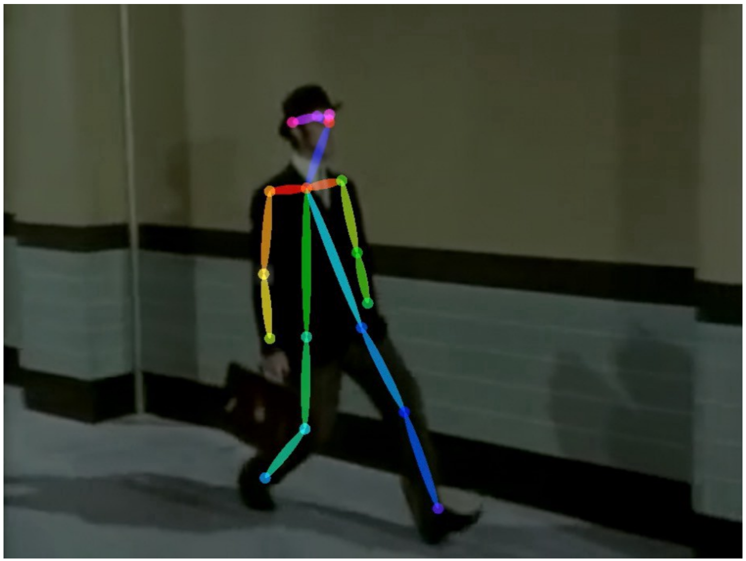

fig 3

After the pose of each frame has been recorded, MoSculp then reconstructs the 3D model of the person in each frame. This is done by a modified SMPL algorithm (a standard machine learning algorithm that reconstructs a 3D model of a person). The SMPL algorithm here will take three parameters—the transition of “pose” between each frame, the “pose” in each frame, and the 2D shape of the figure—and construct a 3D model of the figure in each frame. To ensure the motion in 3D is smooth and consistent with the video recording, the learning of SMPL is driven by a cost function that aims to minimize three terms simultaneously—projection error, the impossible pose, and the disruptive motion. The result after we have reconstructed the 3D model of the person in each frame is shown in *fig 4a.*

The third step of the MoSculp is to generate the 3D motion sculpture along with its depth information. There is nothing special about this step after we have calculated the 3D model frame by frame. The computer simply extracts the surface skeleton of all 3D model (*fig 4b*) we calculated in the last step and join them frame by frame (*fig 4c*).

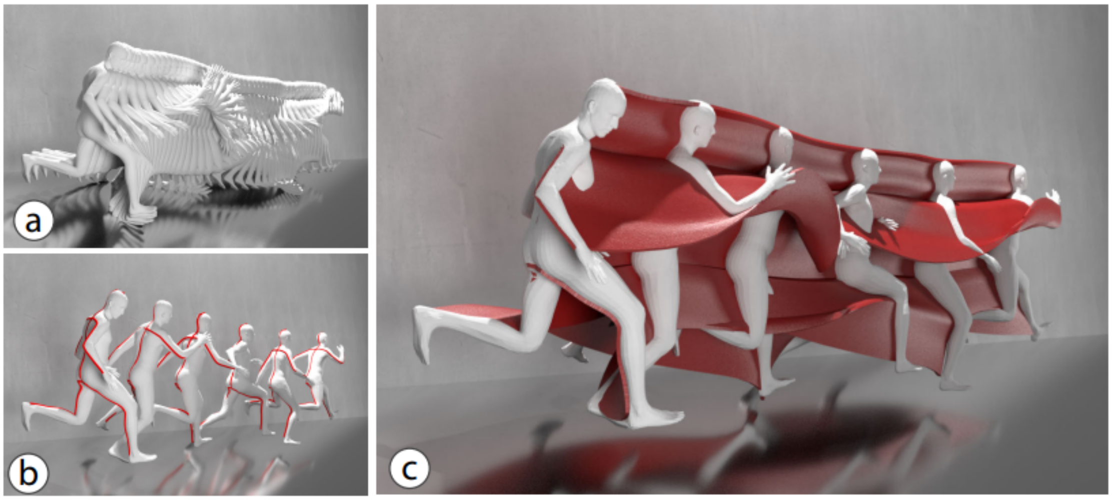

fig 4

Finally, the MoSculp will take an optional step to render the sculpture back to the video. In this step, the algorithm will estimate a depth map of the character with his/her clothes in each frame and overlay the motion sculpture back to the video according to the matches in the two depth maps. To fix the possible artifact due to inaccuracies in estimation, MoSculp will project the motion sculpture into the 2D plane of the video to adjust the x-y coordinate of the sculpture. To compensate for the distortion introduced by the x-y coordinate adjustment, each dimension will be smoothed by a Gaussian kernel. The result of this refinement is the final output of the MoSculp—a motion sculpture with 3D information of how the motion happens in real space-time without observable artifacts (*fig 5*).

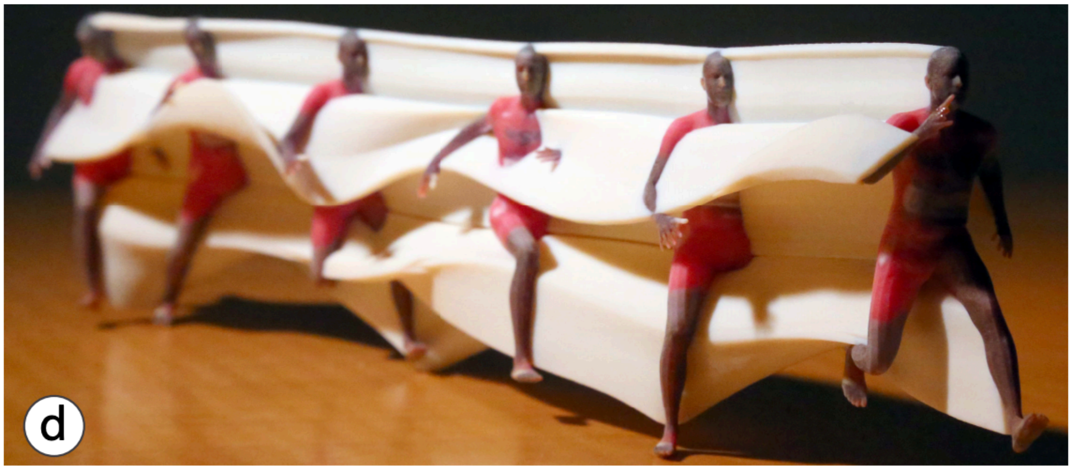

fig 5

The MoSculp is truly a revolutionary method to interpret motions from video recordings. Its ability to recover the 3D information from a 2D video using machine learning is a big leap in the field of motion modeling. Despite its limitation—can only handle video with one person in it, MoSculp has a lot of potential applications. Firstly, it can be used in fields that require intensive studies of motion such as dance and athletics; taking advantage of the 3D information recovered, it is much easier for the users to trace the details of the motion from the video. Secondly, it can be used in video artistic tools that allow users to trace the motion or render the process of the motion in a perceivable way. Finally, by output 3D information output of the video, it can be used to produce motion sculpture in other media such as 3D printed model. It is likely that this software would soon be applied in our lives in the future.

(1085 words)

Reference

Xiuming Zhang, Tali Dekel, Tianfan Xue, Andrew Owens, Qiurui He, Jiajun Wu, Stefanie Mueller, William T. Freeman, ACM Symposium on User Interface Software and Technology (UIST) 2018, MoSculp: Interactive Visualization of Shape and Time. http://mosculp.csail.mit.edu/assets/paper.pdf

Creating 3-D-printed "motion sculptures" from 2-D videos, MIT News, September 18, 2018. https://news.mit.edu/2018/creating-3-d-printed-motion-sculptures-from-2-d-videos-mit-csail-0919