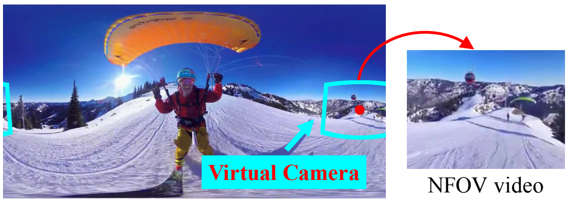
Eurographics Workshop on Intelligent Cinematography and Editing (2017)
W. Bares, V. Gandhi, Q. Galvane, and R. Ronfard (Editors)

Pano2Vid: Automatic Cinematography for Watching 360° Videos

Yu-Chuan Su, Dinesh Jayaraman, and Kristen Grauman

The University of Texas at Austin



360° video (input)   NFOV video (output)

Figure 1: *Input and output of the Pano2Vid problem. The task is to control a virtual camera in the 360° camera axis in order to capture a normal video that looks as if captured by human videographer.*

A 360° camera captures the entire visual world from its optical center, which provides exciting new ways to record and experience visual content by relieving restrictions on the field-of-view (FOV). Videographers no longer have to determine what to capture in the scene, and human viewers can freely explore the visual content. On the other hand, it also introduces new challenges for the video viewer. The most common interface for watching 360° videos is to display a small portion of the video as a normal field-of-view (NFOV) video captured by a virtual camera. The video viewer has to decide "where and what" to look at by controlling the direction of the virtual camera throughout the full duration of the video. For example, on YouTube, the human viewer has to control the virtual camera direction by dragging the video or clicking on the control panel. Because the viewer has no information about the content beyond the current FOV, it may be difficult to find interesting content and determine where to look in real time.

To address this difficulty, we define "Pano2Vid", a new computer vision problem. The task is to design an algorithm to automatically control the pose and motion of a virtual NFOV camera within an input 360° video. The output is the NFOV video captured by this virtual camera. See Fig. 2. Camera control must be optimized to produce video that could conceivably have been captured by a human observer equipped with a *real* NFOV camera. A successful Pano2Vid solution would therefore take the burden of choosing "where to look" off both the videographer and the end viewer: the videographer could enjoy the moment without consciously directing her camera, while the end viewer could watch intelligently-chosen portions of the video in the familiar NFOV format.

We propose the AUTOCAM algorithm to solve the Pano2Vid problem in a data driven approach that does not require human labor [SJG16, SG17]. The algorithm first learns a discriminative model of human-captured NFOV web videos. The NFOV videos are crawled automatically from the web and do not need human annotations. It then uses this model to identify candidate viewpoints and events of interest to capture in the 360° video, before finally stitching them together through optimal camera motions using a dynamic programming formulation for presentation to human viewers. Unlike prior attempts at automatic cinematography, which typically focus on virtual 3D worlds and employ heuristics to encode popular idioms from cinematography, AUTOCAM 1) tackles real video from dynamic cameras, and 2) considers *learning* cinematographic tendencies from data. See some example output videos[†].

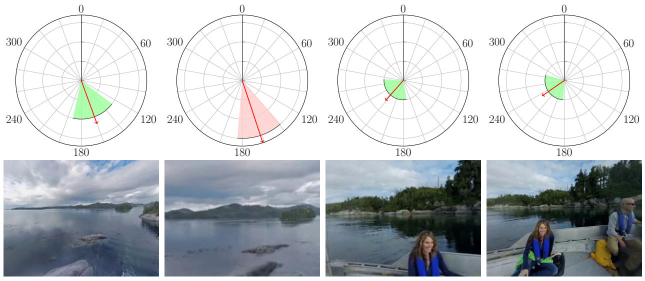


Figure 2: *Example frames of* AUTOCAM *outputs and the corresponding camera poses. We show the camera FOV and azimuthal angle by the circular sector and polar angle by the color. Red/green indicates the angle is greater/smaller than 0, and more saturated color indicates larger value. The camera first captures distant scene with long focal length and close objects with short focal length.*

For evaluation, we compile a dataset of 360° videos downloaded from the web, together with human-selected NFOV camera trajectories. The dataset consists of 86 360° videos with a combined length of 7.3 hours and 3.4 hours of human-selected trajectories. We also define multiple evaluation metrics that measure how close a Pano2Vid algorithm's output videos are to human-generated NFOV videos. The metrics fall into two groups: *HumanCam-based metrics* measure how much the algorithm-generated video content looks like human-generated videos, and *HumanEdit-based metrics* measure the similarity between algorithm-generated and human-selected trajectories. All the metrics can be reproduced easily, and data and evaluation code are available[‡]. The AUTOCAM algorithm outperforms the baselines in all metrics, and indicates the promise of learning "where to look" for 360° video.

References

[SG17] SU Y.-C., GRAUMAN K.: Making 360° video watchable in 2d: Learning videography for click free viewing. In *CVPR* (2017). 1

[SJG16] SU Y.-C., JAYARAMAN D., GRAUMAN K.: Pano2vid: Automatic cinematography for watching 360° videos. In *ACCV* (2016). 1

[†] http://vision.cs.utexas.edu/projects/Pano2Vid/output_examples.mp4

[‡] http://vision.cs.utexas.edu/projects/Pano2Vid

DOI: 10.2312/wiced.20171071