CasualVRVideos: VR videos from casual stationary videos

Stefanie Zollmann
stefanie.zollmann@otago.ac.nz
University of Otago
Dunedin, NZ

Anthony Dickson
dican732@student.otago.ac.nz
University of Otago
Dunedin, NZ

Jonathan Ventura
jventu09@calpoly.edu
Cal Poly
San Luis Obispo, US

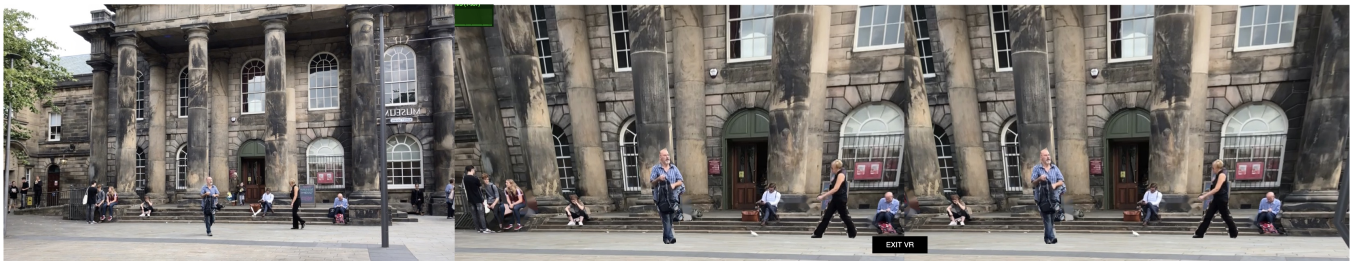

Figure 1: CasualVRVideos. Left) Input video frame captured by mobile phone camera (Museum sequence). Middle and Right) CasualVRVideos displayed for VR usage (displays image for left and right eye).

ABSTRACT

Thanks to the ubiquity of devices capable of recording and playing back video, the amount of video files is growing at a rapid rate. Most of us have now video recordings of major events in our lives. However, until today, these videos are captured mainly in 2D and are mostly used for screen-based video replay. Currently there is no way for watching them in more immersive environments such as on a VR headset. They are simply not optimized for playback in stereoscopic displays or even tracked Virtual Reality devices.

In this work, we present CasualVRVideos, a first approach that works towards solving these issues by extracting spatial information from video footage recorded in 2D, so that it can later be played back in VR displays to increase the immersion. We focus in particular on the challenging scenario when the camera itself is not moving.

CCS CONCEPTS

• **Hardware → Emerging technologies**; • **Computing methodologies → Virtual reality**;

KEYWORDS

virtual reality, single view geometry, content creation

ACM Reference format:
Stefanie Zollmann, Anthony Dickson, and Jonathan Ventura. 2020. CasualVRVideos: VR videos from casual stationary videos. In *Proceedings of 26th ACM Symposium on Virtual Reality Software and Technology, Virtual Event, Canada, November 1–4, 2020 (VRST '20),* 3 pages.
https://doi.org/10.1145/3385956.3422119

1 INTRODUCTION

Mobile phones and other video-capable devices such as action cameras led to a drastic increase of public and private video material. Thanks to improved video quality, we increasingly record videos instead of photos for major events in our lives. At the same time, we see more displays in the consumer market that are capable of playing back more immersive video footage such as stereo projectors, 3D TVs, but also differently priced VR devices ranging from mobile phone driven VR headsets to fully tracked desktop PC powered VR headsets. Common to all of these devices is that they are capable of playing back videos in a more immersive way but require full 3D or some kind of depth information. However, for creating such 3D videos, additional equipment [5] or multiple-cameras setups [6] are required. Setups for capturing more sophisticated light fields are even more complex [2]. These setups are often not accessible to casual users. Computing 3D with depth from mono approaches [4] creates its own problems for VR consumption such as the egomotion of the camera that is likely to create motion sickness. In order to address those challenges, we propose CasualVRVideos a novel way to create VR experience from casual videos.

2 CASUALVRVIDEO COMPUTATION

The main goal of our approach is to extract 3D data from stationary videos for VR consumption. We define stationary videos as videos that are captured by a person in one location only changing their viewing direction, a movement pattern that is often observed when capturing outdoor events. A challenging scenario as we can not rely on standard structure-from-motion (SfM) approaches. Our approach aims to address rotating as well as completely fixed captures.

To achieve our goal we need to extract 3D data for each frame of the 2D video sequence. In our approach, we differentiate between static scene elements (e.g. building walls) and dynamic scene elements (e.g. people) in the scene to maintain consistency of static parts of the scene and reduce the amount of data that is stored.

VRST '20, November 1–4, 2020, Virtual Event, Canada
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7619-8/20/11.
https://doi.org/10.1145/3385956.3422119

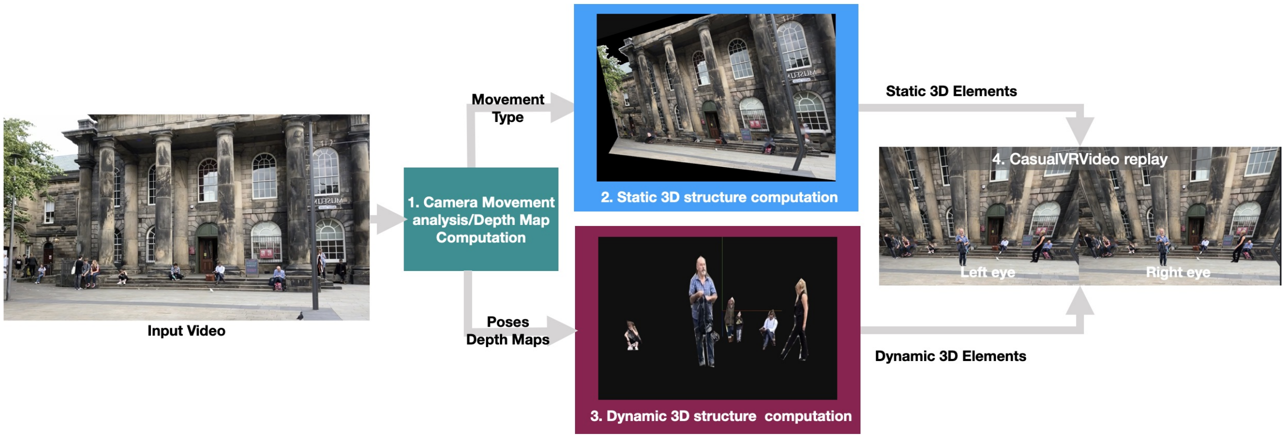
Input Video
1. Camera Movement analysis/Depth Map Computation
Movement Type
Poses Depth Maps
2. Static 3D structure computation
3. Dynamic 3D structure computation
Static 3D Elements
Dynamic 3D Elements
4. CasualVRVideo replay
Left eye
Right eye

Figure 2: Overview CasualVRVideo Computation. 1) The input video is analysed for camera movements and is used to extract depth maps. 2) Based on the type of camera movement, we use for different approaches to extract 3D data representing the static background. 3) Computing 3D meshes from the dynamic foreground elements by combining instance segmentation and single view geometry. 4) Rendering in a VR headset.

Our method is based on four steps (Figure 2): We start with analyzing the input video for camera movements using spherical SfM [1] and compute depth maps. The next step is the extraction of static scene elements. Based on the type of camera movement, we use different approaches to extract 3D data representing the static scene elements. For sequences with only small rotations we use 3D structure analysis on the starting frame to extract a photopopup like 3D structure [3]. For larger rotational movements, we use spherical SFM for creating a stereo panorama from the static background [1]. We use the pose data that results from the camera movement analysis step in combination with instance segmentation and single view geometry to compute billboard like 3D meshes from the dynamic foreground elements. Finally we render the computed 3D representations for the static scene elements represented by a photopopup model or a stereo panorama as well as the dynamic billboards in a VR headset.

3 CASUALVRVIDEO REPLAYS

Once we have the 3D representations for static scene elements and the dynamic parts for each frame of the video computed, we render them in a VR headset. Our CasualVRVideo replay application is based on WebXR[1] in order to provide platform flexibility. The results of the *CasualVRVideo computation* step are untextured 3D meshes. In order to give those meshes the correct appearance, we use projective texture mapping to re-project the background image, as well as the video frames onto the computed 3D meshes. For both texture mappings, we setup a projection matrix that is given by the intrinsics and extrinsic of the camera for each frame of the video sequence. The texture for the static elements in a static camera sequence is fixed and a single image. In contrast for the dynamic elements in all types of sequences, we use video texturing. Here, we make sure that we use the right video frame as well as the correct transformation matrix for the corresponding dynamic mesh by using frame indexing. For this purpose, we store each dynamic mesh with a frame index describing the video frame it was extracted from.

4 RESULTS

We tested our method with a set of different video sequences. As there is currently no standard dataset available for creating casual VR videos, we created our own dataset to cover different egomotion patterns such as fixed camera (Figure 1), very small camera movements (rotations), as well as lager camera movements (rotation).

We tested the results of the CasualVRVideo for suitability of rendering them in a VR headset. All CasualVRVideo results for our test sequences were possible to be display as a complete sequences in our VR headset (Oculus Quest, 60FPS). This is in contrast to our results with regards to using 3D models created from depth from mono approaches.

5 CONCLUSIONS AND FUTURE WORK

We proposed CasualVRVideos, an approach for creating immersive experiences from casually captured videos by stationary users, posing a challenging input to standard 3D reconstruction methods. By combining image processing, single view geometry and deep learning we are able to extract simplified 3D information for each video frame. We show that this 3D representation creates compact data representations and is suitable for display in a VR headset.

The results that we achieved so far allow users to explore all our test video sequences within immersive way. We tested them in a VR headset (Oculus Quest) that allows for 6DOF motion.

Currently our approach purely focuses on stationary video sequences. In the future we plan to expand this by integrating traditional SfM into our pipeline.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the New Zealand Marsden Council through Grant UOO1724.

[1] https://www.w3.org/TR/webxr/

REFERENCES

[1] L. Baker, S. Mills, S. Zollmann, and J. Ventura. 2020. CasualStereo: Casual Capture of Stereo Panoramas with Spherical Structure-from-Motion. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 782–790.

[2] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew DuVall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. 2020. Immersive Light Field Video with a Layered Mesh Representation. 39, 4 (2020), 86:1–86:15.

[3] Derek Hoiem, Alexei A. Efros, and Martial Hebert. 2005. Automatic photo pop-up. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 577. http://portal.acm.org/citation.cfm?id=1073204.1073232

[4] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2019. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. (2019). arXiv:cs.CV/1907.01341

[5] Christian Richardt, Carsten Stoll, Neil A. Dodgson, Hans Peter Seidel, and Christian Theobalt. 2012. Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. In *Computer Graphics Forum*, Vol. 31. 247–256. https://doi.org/10.1111/j.1467-8659.2012.03003.x

[6] Enliang Zheng, Dinghuang Ji, Enrique Dunn, and Jan-Michael Frahm. 2015. Sparse Dynamic 3D Reconstruction from Unsynchronized Videos. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 4435–4443. https://doi.org/10.1109/ICCV.2015.504