Casual 6DOF Virtual Reality Video Creation

HAOXI SUN* and STEFANIE ZOLLMANN*, University of Otago, New Zealand

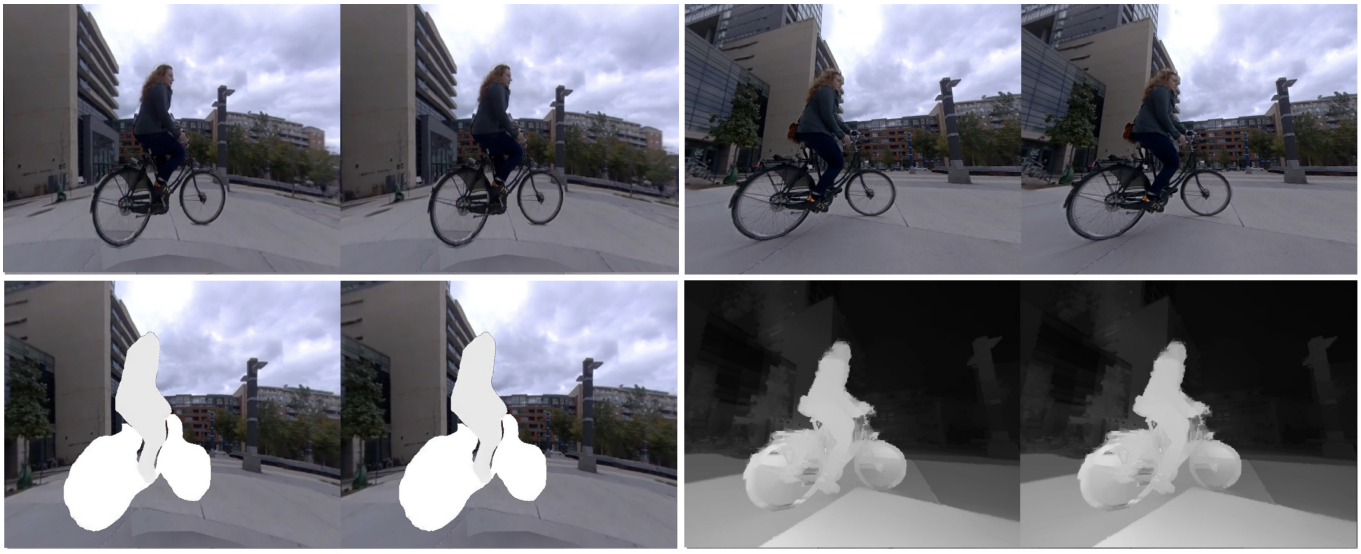

Fig. 1. Virtual Reality stereo views of 6DoF VR videos. Left: Casual 6DoF VR videos. Right: Captured with stereo camera and processed with Stereo2Depth ([3]). Using video material publicly available (http://pseudoscience.pictures). Bottom: Depth maps for (Left) Casual 6DoF VR videos (0verlay onto panoramic background image, (Right) depth maps for Stereo2Depth.

In contrast to traditional media, content production tools and capturing for immersive experiences such as for Virtual Reality (VR) headsets are still not widely accessible to casual users. Often expensive hardware or sophisticated software tools are involved in the capturing and content production pipeline. The main goal of our work is to address this gap and simplify content creation for virtual reality viewing to make it accessible to casual users.

While traditional 360 panorama photographs and 360-degree cameras can already be viewed on virtual reality devices, they do not provide a sense of depth to the user. Traditional methods for reconstructing stereo panoramas on the other hand involve specialized hardware. In this work, we investigate methods that allow computing 6-degree-of-freedom (6DOF) videos from standard 360-degree cameras. Thereby, we extract multiple layers of interest representing the background and the foreground.

CCS Concepts: • **Computing methodologies** → **Virtual reality**; *Image-based rendering*.

Additional Key Words and Phrases: Virtual Reality, Capturing, 6DOF, casual

ACM Reference Format:
Haoxi Sun and Stefanie Zollmann. 2022. Casual 6DOF Virtual Reality Video Creation. In *Workshop on Advanced Visual Interfaces for Augmented Video.* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this research.

AVIxAV2022, June 07, 2022, Rome, Italy, June 2022

1 INTRODUCTION

Virtual reality applications are getting more and more popular. In particular, with the recent availability of affordable VR headsets, VR applications and games are getting more attention in the consumer market. One factor here is affordability but also major developments that now allow for fully integrated tracking, controllers, and computation in one device allow that these devices can be used by a wider audience. Often these devices are used for gaming, but also media consumption, such as watching immersive videos (360-degree videos or 360 + Depth videos) is becoming more popular. However, while 360 videos are easy to produce by capturing a scene of interest with a panorama camera, these videos are less immersive as they do not provide any depth information about the scene surrounding the user. In contrast, 360 + Depth videos (also called 6DOF videos [5]) are more immersive as they provide depth, but they are still expensive to produce with only limited products available to the consumer market. 6DoF videos can also be created using a depth from stereo approach [3]. Recent work by Broxton et al. [2] presents an approach that synthesizes 6DoF videos from over 40 single cameras placed on a sphere. Other methods synthesize high-quality 6DoF video in real-time from 360 ODS footage [1]. Recent depth estimation methods allow for an estimated depth not only for single images but also for 360 images.

However, they still often require expensive hardware to use. In our work, we look into options for creating more immersive 360 videos with more affordable approaches. We call these casual 6DOF VR Videos. The main idea is that we are separating the scene's

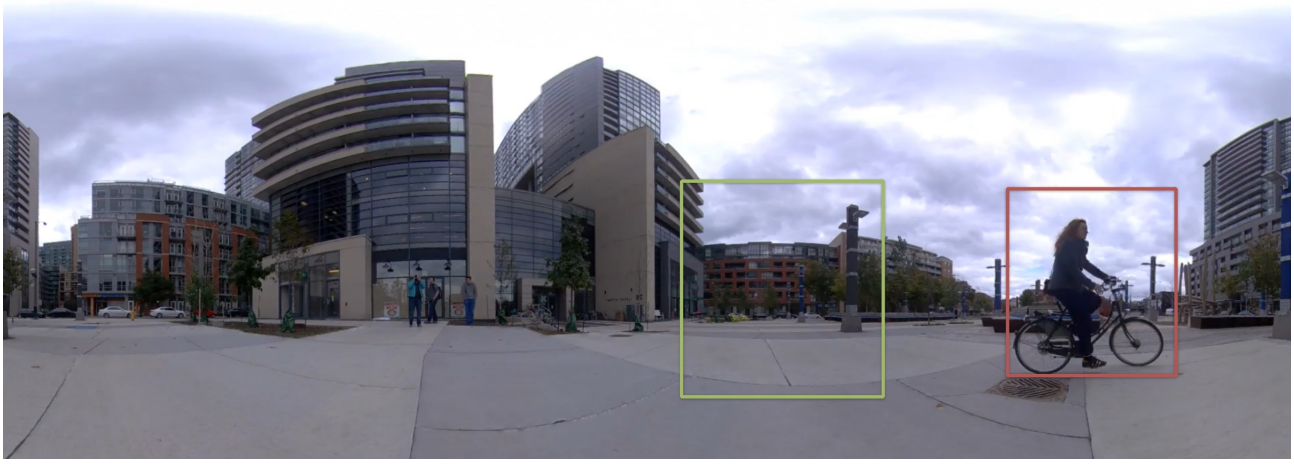

Fig. 2. Identifying a dynamic object in the scene to replace in other sequence frame. Dynamic object (red frame) is identified to not longer occupy background (green frame) in this video frame.

background from the dynamic foreground and compute depth information for each dynamic element.

2 COMPUTING CASUAL 6DOF VIDEOS

Our method is based on the assumption that often large parts of a scene will remain static, such as a street scene that consists of buildings and street furniture. The important action often happens in the foreground but only covers a small amount of pixels in the overall images. Based on this assumption, we separate the static background from the dynamic foreground first. For this purpose, we use instance segmentation[1] to extract dynamic foreground objects. This step gives all regions in each frame of the 360 video that are covered by dynamic objects such as persons or cars (Figure 3). We then use a patch-based search and iterate over all video frames to identify the best image patches to replace dynamic content in the first frame of the sequence. As dynamic content is moving there is a high likelihood of finding pixels that were not occupied with pixels of dynamic objects at some stage of the sequence. This step gives us a panoramic representation of the static background with a low amount of dynamic objects occluding the scene.

Once, we have extracted the static background of the scene as a panoramic image, in the next step we have to assign depth values to each dynamic foreground object. For this purpose, we use the extracted dynamic objects and compute the size of each object in pixels. Objects that are further away are assumed to be smaller in image space and objects that are closer to the camera are assumed to be larger. While this is just a very coarse approximation, it allows us to assign different distances to objects in the scene and allows us to compute a layered representation of the scene. Layered mesh representations have been used for light representation in previous works [2]

[1]https://detectron2.readthedocs.io

3 RENDERING

Once, we computed both the static background and the depth of dynamic foreground objects (Figure 1), we can use this information for rendering. We use a web-based video editor based on WebXR[2] for replay of our Casual VR video [4]. We adapted the video editor to display a 360 panoramic image as background using a spherical geometry. The dynamic foreground layers are also rendered on a spherical geometry, but the depth values are used as displacement values to render the objects closer or further away from the camera. Using WebXR for replay allows for platform independence. We tested the replays on the Oculus Quest [3] and the WebXR simulator [4].

4 CONCLUSION

In this work, we highlight the gap in tools for creating immersive experiences from casually captured 360 videos and we explore methods for creating immersive experiences such as footage without the need for expensive hardware. Our aim is thereby to make content creation for VR headsets and immersive experiences more accessible to users that might not have access to expensive hardware and sophisticated software tools. For our work, we plan to run user studies to investigate the difference in quality between content that is captured for instance with stereo cameras (Figure 3, Right) and Casual 6DoF VR videos (Figure 3, Left). As there is a loss in quality as visible in these sample images and image loss (such as shadows), we are interested in how much users will notice this and how important it is for their experience.

ACKNOWLEDGMENTS

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

[2]https://www.w3.org/TR/webxr/
[3]https://www.oculus.com/quest-2/
[4]https://github.com/MozillaReality/WebXR-emulator-extension

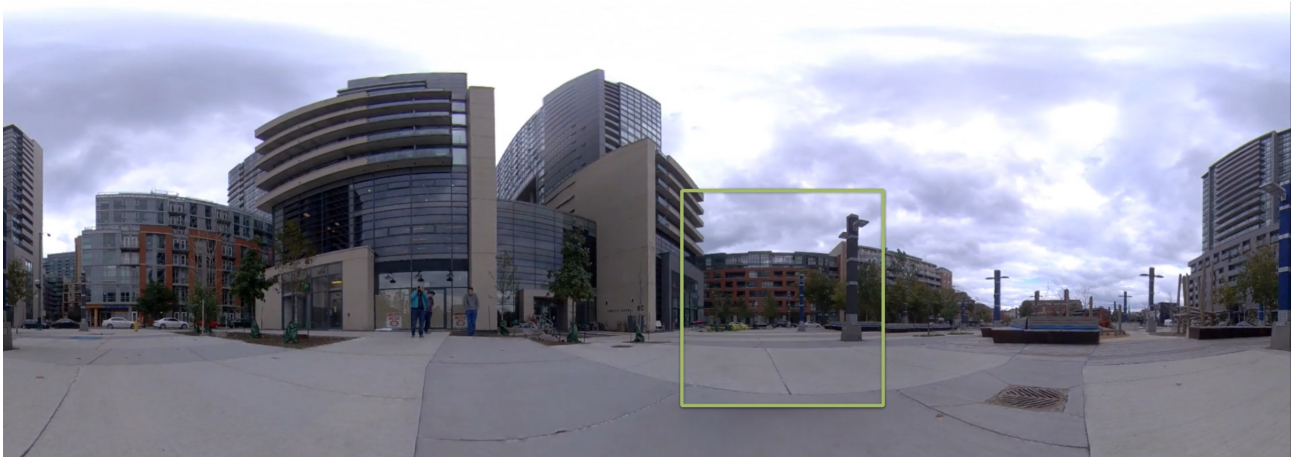

Fig. 3. Identifying a dynamic object in the scene to replace in other sequence frame. Areas in the start frame of the sequence that contained dynamic object are replaced with content from (top frame).

REFERENCES

[1] Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. 2020. MatryODShka: Real-time 6DoF Video View Synthesis using Multi-Sphere Images. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, Vol. 2020. Springer, 441–459. https://doi.org/10.1007/978-3-030-58452-8_26 European Conference on Computer Vision 2020, ECCV ; Conference date: 24-08-2020 Through 28-08-2020.

[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. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 39, 4 (2020), 86:1–86:15. https://doi.org/10.1145/3388536.3407878

[3] Josh Gladstone and Timotheos Samartzidis. 2019. Pseudoscience stereo2depth. https://github.com/n1ckfg/pseudoscience_stereo2depth. Accessed on 05.05.2021.

[4] Ruairi Griffin, Tobias Langlotz, and Stefanie Zollmann. 2021. 6DIVE: 6 Degrees-of-Freedom Immersive Video Editor. *Frontiers in Virtual Reality* 2 (2021). https://doi.org/10.3389/frvir.2021.676895

[5] Christian Richardt, Peter Hedman, Ryan S. Overbeck, Brian Cabral, Robert Konrad, and Steve Sullivan. 2019. Capture4VR: From VR Photography to VR Video. In *ACM SIGGRAPH 2019 Courses* (Los Angeles, California) *(SIGGRAPH '19)*. Association for Computing Machinery, New York, NY, USA, Article 4, 319 pages. https://doi.org/10.1145/3305366.3328028