Towards Casually Captured 6DoF VR Videos

1st Haoxi Sun
Department of Computer Science
University of Otago
Dunedin, New Zealand
sunao770@student.otago.ac.nz

2nd Stefanie Zollmann
Department of Computer Science
University of Otago
Dunedin, New Zealand
stefanie.zollmann@otago.ac.nz

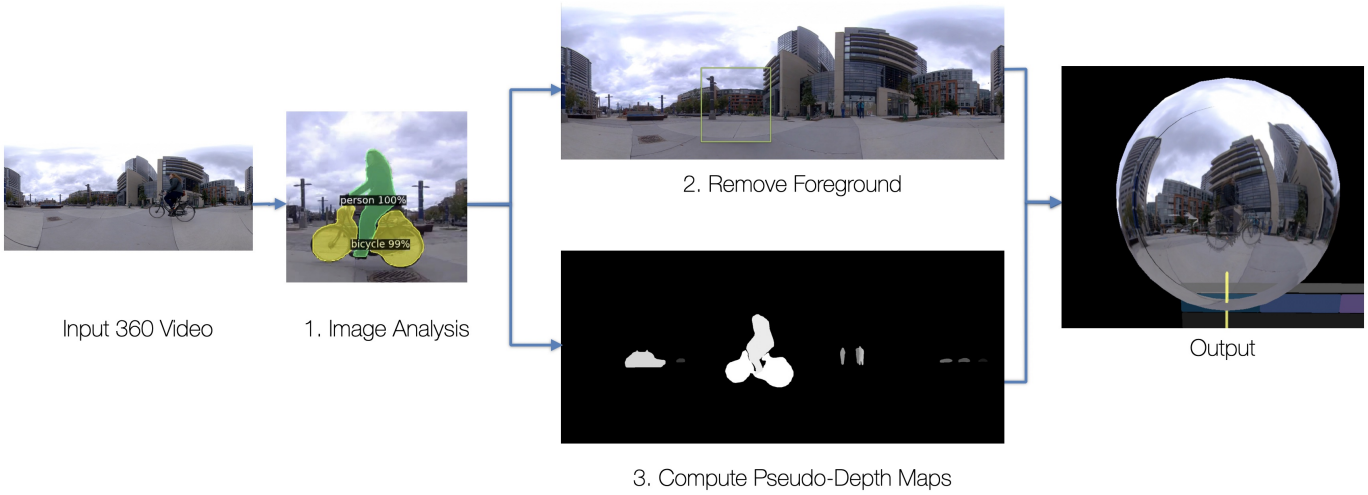

Fig. 1: Overview of 6DoF VR Video approach. 1) We start with a single 360 video as input and analyze each frame for foreground objects which are assumed to be dynamic. 2) We use this information to remove foreground objects from the video sequence and to create a static representation of the scene background. 3) We then compute pseudo-depth maps that create a layered representation of the 360 video. Finally, we render the result in VR using a custom WebXR VR video player.

***Abstract*—In contrast to traditional media, content production tools and capturing for immersive experiences such as 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 videos can already be viewed on virtual reality devices, they do not provide a sense of depth to the user. Traditional methods for reconstructing depth or stereo information for 360 footage 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. We also provide some first results on the results achieved.

***Index Terms*—Virtual Reality, Capturing, Content Creation, Casual, 6DoF**

I. 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. In addition, recent developments now allow for fully integrated tracking, controllers, and computation in one device and as such these devices can be used by a wider audience. Often these devices are used for gaming, but also for 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 (degree of freedom) videos [9] or 3DoF+ [3]) are more immersive as they provide depth, but they are still expensive to produce with only limited products available to the consumer market. For instance, 6DoF videos can be created using a depth-from-stereo approach from stereo videos [5] or more sophisticated multi-camera setups [2]. The main goal of our work is to simplify content creation for virtual reality viewing to make it accessible to casual users.

II. Related Work

Recent work by Broxton et al. [2] presents an approach that synthesizes 6DoF videos from over 40 single cameras placed

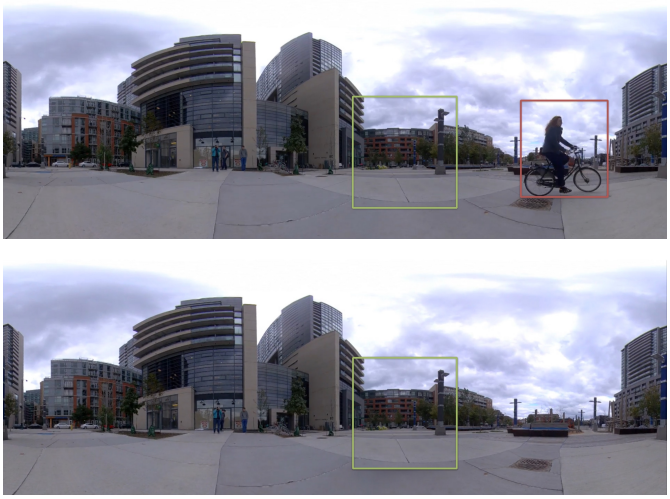

Fig. 2: Identifying a dynamic object in the scene to replace in another sequence frame. Top) A dynamic object (red frame) is identified to no longer occupy the background (green frame) in this video frame. Bottom) Areas in the start frame of the sequence that contained dynamic objects are replaced with content from the initial frame (top frame).

on a sphere. Other methods synthesize high-quality 6DoF video in real-time from 360 omnidirectional stereo (ODS) footage [1]. Recent depth estimation methods allow for an estimated depth not only for monocular images [8] but also for 360 images [12]. While a lot of depth estimation models are trained on indoor data, more recent methods also focused on 360 depth estimation in the wild [4]. While these depth estimation methods often focus on single image depth estimation, Serrano et al. proposed a method for computing motion parallax for 360 videos [10].

However, existing methods still often require expensive hardware to be used to support the image processing, long processing times or require special input for the processing, such as ODS footage or footage from complex camera rigs. Also when monocular depth estimation methods are applied to videos, they often create the problem of temporal inconsistencies between the frames. In our work, we look into options for creating more immersive 360 videos with more affordable approaches that are available to a wide range of users. We call these casual 6DoF VR Videos. The main idea is that we separate the scene's background from the dynamic foreground and computing depth information for each dynamic element.

In this work, we describe our approach for computing casually captured 6DoF videos and provide some first results on the results achieved. We thereby build on the assumption that the 360 camera that captures the scene is static and make use of the assumption that the background of the scene is mostly static. We consider this assumption to be a valid one as 360 footage with a lot of ego motion is often causing motion sickness [7]. Thus, often content is captured with a static camera on a tripod. It is important to mention that the output of our method only provide a limited amount of 6DoF movement around the original capture position.

III. 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 that is captured in a video often happens in the foreground but only covers a small number 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 2). 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. While a similar outcome could be achieved by using image inpaiting [13], our method uses actual video pixels instead of synthesizing information.

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 field representation in previous works [2] as well as for panoramic images [11]. Based on the results, we compose a stacked video that contains the RGB information

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

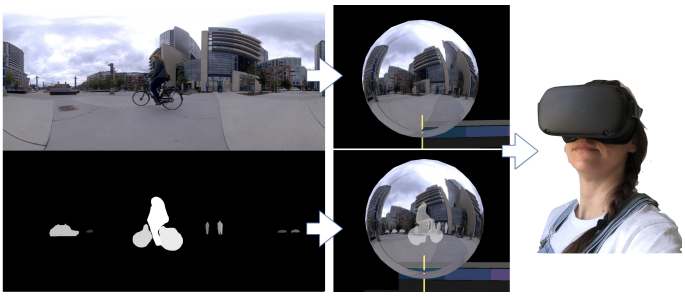

Fig. 3: The output of our method is used for rendering in VR. Top) A presentation of the static background is rendered onto a sphere. Bottom) The dynamic foreground layers are also rendered on a sphere. Thereby the pseudo-depth values are used as displacement values to render the objects closer or further away from the camera.

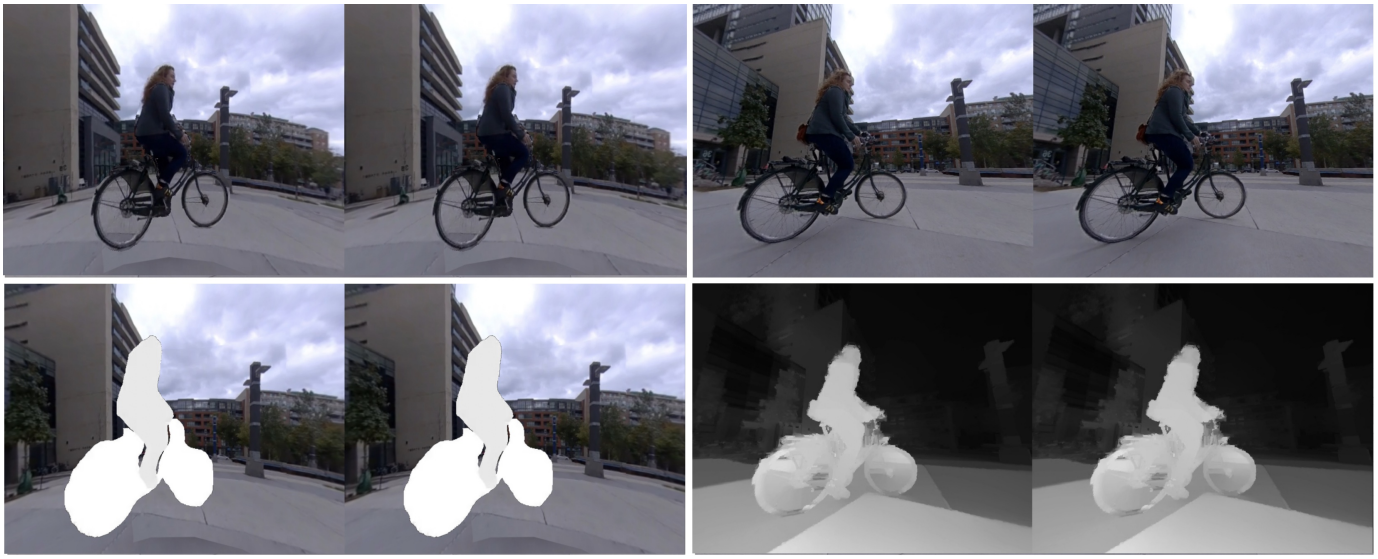

Fig. 4: Results: Virtual Reality stereo views of 6DoF VR videos. Left: Casual 6DoF VR videos. Right: Captured with a stereo camera and processed with Stereo2Depth ([5]). Using video material publicly available (http://pseudoscience.pictures). Bottom: Depth maps for (Left) Casual 6DoF VR videos (0verlay onto panoramic background image, (Right) depth maps for Stereo2Depth.

on the top and the depth information at the bottom (Figure 3 Left).

IV. Rendering

Once, we computed both the static background and the depth of dynamic foreground objects (Figure 1), we use this information for rendering. We use a web-based video editor based on WebXR[2] for the replay of the Casual VR video [6]. 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. For each frame rendered during replay, we only update the foreground layer including the RGB data and the pseudo-depth. Using WebXR for replay allows for platform independence. We tested the replays on the Oculus Quest[3] and the WebXR simulator [4]. The user then experiences an immersive presentation of the VR video using a VR headset (Figure 3).

V. Results

As this work is still in an early stage, we focused on exploring the visual quality of our proposed approach and existing methods. As our approach focuses on making the method affordable and available for casual users, a certain loss in quality could be expected and acceptable.

We compared our results (Figure 4, Left) with the quality that can be achieved when capturing the video material with stereo cameras (Figure 4, Right). We experienced a loss in details (e.g. fine depth variations, missing shadows) and in quality.

Our preliminary tests in a VR headset[5] were promising as a sense of depth was provided to the viewer and visual artifacts were not disturbing the experience (Figure 4, 5). However, these preliminary tests need to be confirmed with a user study.

VI. Conclusion and Future Work

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 2, Right) and Casual 6DoF VR videos (Figure 2, 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. Another issue is that our method might recognize static objects as foreground objects (for example as visible in Figure 1 a bench is recognized as a foreground object). In our user study, we will also look into such artifacts and their impact on user experience. In addition, we are interested to compare the results to other existing work such as the work by Serrano et al. [10].

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

[3]https://www.oculus.com/quest-2/

[4]https://github.com/MozillaReality/WebXR-emulator-extension

[5]https://youtu.be/0BeygmNxatU

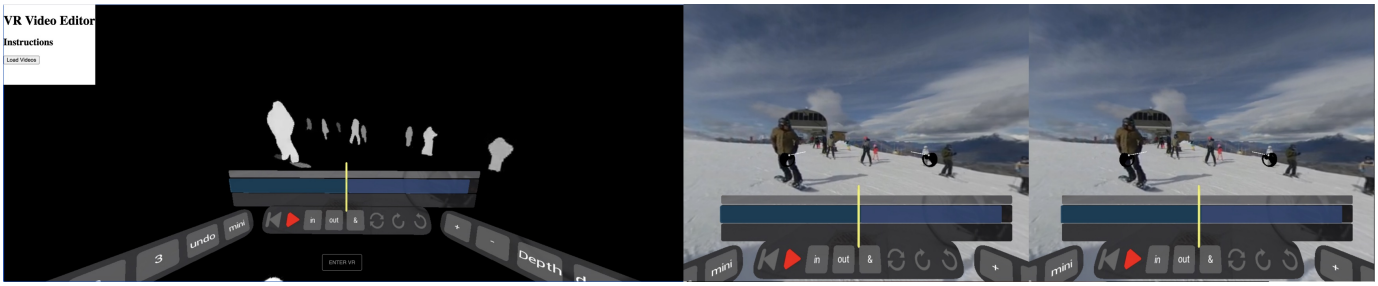

Fig. 5: Results Snowboarding sequence in WebXR renderer. Left: Pseudo Depth Maps for Snowboarding Sequence. Virtual Reality stereo views of 6DoF VR videos. Right: Stereo Rendering Casual 6DoF VR videos (showing Left and Right eye).

ACKNOWLEDGMENTS

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

REFERENCES

[1] B. Attal, S. Ling, A. Gokaslan, C. Richardt, and J. Tompkin. MatryODShka: Real-time 6dof video view synthesis using multi-sphere images. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, vol. 2020, pp. 441–459. Springer, Nov. 2020. European Conference on Computer Vision 2020, ECCV ; Conference date: 24-08-2020 Through 28-08-2020. doi: 10.1007/978-3-030-58452-8_26

[2] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. DuVall, J. Dourgarian, J. Busch, M. Whalen, and P. Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 39(4):86:1–86:15, 2020. doi: 10.1145/3388536.3407878

[3] T. L. T. da Silveira and C. R. Jung. Dense 3d scene reconstruction from multiple spherical images for 3-dof+ vr applications. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 9–18, 2019. doi: 10.1109/VR.2019.8798281

[4] Q. Feng, H. P. H. Shum, and S. Morishima. 360 depth estimation in the wild - the depth360 dataset and the segfuse network. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 664–673, 2022. doi: 10.1109/VR51125.2022.00087

[5] J. Gladstone and T. Samartzidis. Pseudoscience stereo2depth. https://github.com/n1ckfg/pseudoscience_stereo2depth, 2019. Accessed on 05.05.2021.

[6] R. Griffin, T. Langlotz, and S. Zollmann. 6dive: 6 degrees-of-freedom immersive video editor. *Frontiers in Virtual Reality*, 2, 2021. doi: 10.3389/frvir.2021.676895

[7] P. Hu, Q. Sun, P. Didyk, L.-Y. Wei, and A. E. Kaufman. Reducing simulator sickness with perceptual camera control. *ACM Trans. Graph.*, 38(6), nov 2019. doi: 10.1145/3355089.3356490

[8] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[9] C. Richardt, P. Hedman, R. S. Overbeck, B. Cabral, R. Konrad, and S. Sullivan. Capture4VR: From VR Photography to VR Video. In *ACM SIGGRAPH 2019 Courses*, SIGGRAPH '19, pp. 1–319. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3305366.3328028

[10] A. Serrano, I. Kim, Z. Chen, S. DiVerdi, D. Gutierrez, A. Hertzmann, and B. Masiá. Motion parallax for 360 RGBD video. *IEEE Transactions on Visualization and Computer Graphics*, 25:1817–1827, 2019. doi: 10.1109/TVCG.2019.2898757

[11] J. Waidhofer, R. Gadgil, A. Dickson, S. Zollmann, and J. Ventura. PanoSynthVR: Toward Light-weight 360-Degree View Synthesis from a Single Panoramic Input. In *2022 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2022.

[12] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 459–468, 2020. doi: 10.1109/CVPR42600.2020.00054

[13] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5840–5848, 2019.