360DIV: 360° Video Plus Depth for Fully Immersive VR Experiences

Julie Artois, Glenn Van Wallendael and Peter Lambert
IDLab, Ghent University - imec, Ghent, Belgium
julie.artois@ugent.be

***Abstract*—360° cameras can easily capture events and let viewers across the world re-live them in Virtual Reality (VR). However, the viewer can only rotate their head and there is no correct sense of depth, breaking the immersion. This work describes a framework that reconstructs the captured scene in 3D from a 360° image/video. To do this, it requires depth information, either from depth estimation or depth sensors. The reconstruction is rendered in VR in real-time. Areas of the scene that were not captured, are inpainted in a fast and non-intrusive way. A demo is made available online and will be used for user experiments to assess the perceived quality of experience.**

***Index Terms*—VR, 360° video, 6DoF, depth-image-based-rendering**

I. Introduction

The multimedia industry is pushing towards more and more immersive experiences. Virtual Reality (VR) headsets or holographic displays present the viewer with a more realistic sense of depth compared to regular flat screens. VR especially has the potential to fully immerse the user in the virtual world [1] [2].

For all these applications, hand-crafting photorealistic scenarios in 3D modeling software is a laborious task. Moreover, VR requires considerable processing capacity to render two high-resolution images (one per eye) at least 90 times per second to avoid motion-sickness.

For these reasons, the VR industry is shifting towards showing 360° videos instead [3]. These can easily be acquired and displayed in VR. However, 360° videos do not provide motion parallax, as illustrated in Fig. 1. When the viewer moves, the virtual scene follows, which significantly lowers the quality of experience.

In this work, we process 360° videos to deliver VR experiences with realistic depth perception. In this way, we try to achieve the photorealism and flexibility of 360° cameras combined with the motion parallax of hand-crafted 3D scenes.

We present two contributions. Firstly, we implemented a framework to automatically convert 360° images/videos plus depth information into a 3D scene reconstruction that can be viewed in real-time in VR. Secondly, we made a tool to render the reconstruction in real-time, with support for multiple VR platforms. The tool has several features that are useful in many multimedia applications, such as the ability to incorporate 2D or 3D overlays or objects, to add avatars, and to track where the viewer is looking.

We encourage the reader to try out our demo of the tool[1]. The demo shows an example environment in VR, created by the proposed framework from a 360° video and static depth map. The demo can also display the original monoscopic 360° video, as well as the stereoscopic variant. By switching between these three options, the viewer can compare the level of realism. We intend to use the proposed framework and tool for user experiments to query about the realism of the experience.

This work was funded in part by the Research Foundation – Flanders (FWO) under Grant 1SD8221N, in part by IDLab (Ghent University – imec), Flanders Innovation and Entrepreneurship (VLAIO), and the European Union.

[1]Demo available at https://github.com/IDLabMedia/360DIV

Fig. 1. Left: (Half of) a 360° image is shown to a viewer wearing a VR headset. In essence, the image is projected on a sphere which remains stationary with respect to the viewer's head. There is no motion parallax. Right: Depth information about the original scene is added. The geometry does not move together with the viewer, resulting in the correct depth perception.

II. Related work

To deliver VR applications from 360° videos with motion parallax, the geometry of the captured scene needs to be known. In this work, the distance of each pixel in the 360° video to the camera is stored in a depth map. Most commonly, the depth is estimated from multiple images using Structure from Motion (SfM) [4] and Multi-View Stereo (MVS) [5]. After some refinement steps, the depth maps are often of decent quality but this approach requires a lot of processing power [6].

Deep learning has also proven to be a beneficial approach for depth estimation [7] [8]. The accuracy of the depth is often poor, though, leading to these methods only being used in tandem with rendering techniques that can handle the badly reconstructed geometry.

To reduce the processing load, depth sensors like in the Matterport, Microsoft Kinect or smartphone cameras can be used [9]. For example, the Matterport3D dataset [10] and the dataset by Armeni et al. [11] contain a good collection of 360° images with depth maps which can be used as input to the proposed framework. The depth sensors are not infallible, though, and struggle with reflective or far away surfaces. Additionally, the Matterport camera only works in static environments.

III. Proposed technology

A. Textured triangle mesh

The proposed framework takes two equirectangular 360° videos as input, one for the color and one for the *depth map*. The depth is leveraged to calculate the original position

Fig. 2. (a) The triangle mesh before the deletion of the elongated triangles. (b) The elongated triangles were deleted, leaving behind the dark grey pixels. (c) and (d) are the same views as (b) and (a) respectively, but now with inpainting.

of each pixel in 3D space. In essence, the original scene is reconstructed as a large 3D point cloud of $w*h$ colored points, where $w \times h$ is the resolution of the input video.

A colored point cloud can be viewed in VR from all angles. However, if the viewer moves closer to the geometry, they will notice the gaps in between the scattered points, breaking the immersion. Therefore, it is opted to use textured triangle meshes instead. The 3D points are replaced by triangles that are connected to their neighbors. The level of mesh detail is customizable, but it is best to start with two triangles per pixel.

For performance reasons, the number of triangles in the mesh is decimated [12]. Afterwards, the texture coordinates that define how the 360° video is stretched over the 3D mesh are re-calculated. For many environments, we found that reducing the number of triangles to about one million does not lower the quality in a noticeable way.

B. Inpainting

At this point, all triangles are still connected, as shown in Fig. 2a. This results in elongated triangles connecting foreground to background objects. This stretched-out geometry is not present in the original scene and needs to be removed to maintain a sense of immersion for the viewer. The algorithm deletes triangles based on their circumference and requires manual tweaking for the best result.

A spectator viewing the resulting mesh will notice the holes that the deleted triangles leave behind, for example in Fig 2b. Typically, inpainting is used to color the holes in a plausible way. Deep-learning-based inpainting methods deliver the most convincing results, but the implementations are too slow to be incorporated into VR applications [13]. Other techniques simply interpolate colors from the pixels at the edges of the holes.

We propose a novel implementation of such straightforward inpainting methods that is computationally very fast. The implementation re-uses the previously deleted triangles as a separate mesh, that is rendered behind all other geometry. The texture is changed so that only colors of the background are used. Lastly, the texture is blurred to reduce the spatial frequency and catch less of the viewer's attention. The result is temporally stable and does not distract the viewer, as long as the viewer's head stays close to the original 360° camera. The result is demonstrated in Fig. 2c and 2d.

IV. Conclusion

360° cameras allow scenes to easily be captured and re-played in VR. However, a viewer that is moving around will quickly notice the absence of motion parallax, which breaks the immersion. We present a framework and tool to build realistic VR experiences where the viewer can move around freely, although restricted in volume for optimal performance. The framework takes a 360° image/video and a depth map as input to create a 3D reconstruction of the scene. We encourage the reader to try out the VR demo of the tool. We achieve promising results in terms of photorealism and quality of experience. In the future, the tool will be used to perform user experiments to assess the perceived level of realism.

References

[1] D. De Luca, F. Lodo, and F. M. Ugliotti, "Evaluation of the effects of anxiety on behaviour and physiological parameters in vr fire simulations," in *2021 IEEE International Conference on Consumer Electronics (ICCE)*, 2021, pp. 1–6.

[2] D. Nguyen, H. T. T. Tran, and T. C. Thang, "A telepresence-based remote learning system," in *2022 IEEE International Conference on Consumer Electronics (ICCE)*, 2022, pp. 1–2.

[3] L. Bassbouss, S. Steglich, and I. Fritzsch, "Interactive 360° video and storytelling tool," in *2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT)*, 2019, pp. 113–117.

[4] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4104–4113.

[5] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016, pp. 501–518.

[6] G. Riegler and V. Koltun, "Free view synthesis," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 623–640.

[7] K. Takeda, Y. Iwamoto, and Y.-W. Chen, "Color guided depth map super-resolution based on a deep self-learning approach," in *2020 IEEE International Conference on Consumer Electronics (ICCE)*, 2020, pp. 1–4.

[8] 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 Trans. Graph.*, vol. 39, no. 4, jul 2020.

[9] J. Lei, L. Li, H. Yue, F. Wu, N. Ling, and C. Hou, "Depth map super-resolution considering view synthesis quality," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1732–1745, April 2017.

[10] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.

[11] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017.

[12] M. Garland and P. S. Heckbert, "Surface simplification using quadric error metrics," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '97. USA: ACM Press/Addison-Wesley Publishing Co., 1997, p. 209–216.

[13] M.-L. Shih, S.-Y. Su, J. Kopf, and J.-B. Huang, "3d photography using context-aware layered depth inpainting," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 8025–8035.