DEPTH AUGMENTED STEREO PANORAMA FOR CINEMATIC VIRTUAL REALITY WITH HEAD-MOTION PARALLAX

Jayant Thatte, Jean-Baptiste Boin, Haricharan Lakshman, Bernd Girod

Department of Electrical Engineering, Stanford University
{jayantt, jbboin, harilaks, bgirod}@stanford.edu

ABSTRACT

Cinematic virtual reality (VR) aims to provide immersive visual experiences of real-world scenes on head-mounted displays. Current cinematic VR systems employ omnidirectional stereo videos from a fixed position, and therefore do not address head-motion parallax, which is an important cue for depth perception. We propose a new 3D video representation, referred to as depth augmented stereo panorama (DASP), to address this issue. DASP is developed considering data capture, postproduction, streaming, and rendering stages of the VR pipeline. The capabilities of this representation are evaluated by comparing the generated viewports with those from known 3D models. Results indicate that DASP can successfully create stereo and induce head-motion parallax in a predefined operating range.

Index Terms— Virtual reality, head-motion parallax

1. INTRODUCTION

Modern consumer head-mounted displays (HMDs) have the ability to display wide field-of-view content at high pixel densities to provide immersion. They also have the capability of tracking the viewer's head position and orientation with low latency. These features allow VR applications to create virtual environments with immersive 3D visual experience. Although traditionally VR content is generated using 3D computer models of virtual environments, recent developments in capture devices such as camera rigs and fisheye lenses, along with advances in computational methods for postproduction, have made it possible to acquire videos that can be stitched to produce wide field-of-view content. This facilitates the capture of live-action videos of real-world scenes to build VR experiences, making VR into a new "cinematic" medium.

To make VR attractive for cinematic applications and long-term usage, it is essential that the viewing experience feel comfortable and natural. For computer generated synthetic VR content, several depth cues of human visual perception such as binocular disparity, binocular occlusions, as well as head-motion parallax, are being accurately rendered using current HMDs. While binocular cues are addressed using separate frusta for the left and right eyes, head-motion parallax is supported by using positional tracking to render from the desired perspective. All of these cues are processed jointly by the human visual system to give us a sense of depth in the rendered scene.

However, supporting these depth cues, and particularly head-motion parallax, for real-world content is a rather challenging task for VR applications. Since constructing accurate 3D models of real-world scenes is difficult, image-based rendering techniques are typically used in cinematic VR. In order to update content according to head position, we need omnidirectional video data corresponding to each possible head position. A direct way would be to store omnidirectional videos at a discrete set of head positions and synthesize virtual views for intermediate positions using interpolation. However, this would require capture, storage and transmission of prohibitively large amounts of data.

Our contributions

1. In this paper, we propose a novel data representation for cinematic VR that supports stereo and head-motion parallax for all viewing directions, with only a small data overhead compared to existing representations. We demonstrate the accuracy of viewports synthesized using DASP by quantitatively evaluating them against known 3D models.
2. We perform a systematic study of how the representation radius affects the quality of the synthesized translated views. This analysis is helpful in deciding the optimal choice of capture radius in a practical system.
3. Complete coverage: Omnistereo [1] cannot represent scene points directly above or below the viewing circle and hence is not directly suitable for VR. We propose an extension to omnistereo to represent the entire scene including zenith and nadir regions, without artifacts.
4. We analyze the suitability of this format for different stages of the VR pipeline and discuss the pros and cons compared to existing approaches. We note that this representation can also be used to further optimize the visual experience, e.g., adapting inter-pupillary distance (IPD) according to user preference, disparity manipulations for comfort zone, etc.

Format	Geometry Modeling	Remarks	Data Capture	Post-production	Coding & Streaming	Head motion parallax
3D models	Explicit	Hard to model real-world scenes	*hard*	*hard*	*medium*	*simple*
Omnistereo [1]	No	Extrapolation required for head translation	*medium*	*medium*	*simple*	*hard*
Concentric mosaics [2]	Implicit*	Editing and streaming many rings	*medium*	*hard*	*medium*	*simple*
Mono panorama + depth	Implicit	Extrapolation required for head translation	*simple*	*simple*	*simple*	*hard*
Depth augmented stereo panorama	Implicit	View interpolation using warping	*medium*	*medium*	*simple*	*simple*

Table 1: A summary of 3D video formats for cinematic VR application. The last four columns provide estimates of relative complexity of different stages of the VR pipeline for each format. The asterisk in the 'geometry modeling' column of concentric mosaics denotes the usage of depth information to correct vertical distortions.

2. RELATED WORK ON 3D VIDEO FORMATS

A lot of research has been carried out on plenoptic modeling [3, 4] during the past decades resulting in different types of formats for different applications, such as 3D cinema, stereoscopic television with glasses, and autostereoscopic television. Cinematic VR has the potential to become a new medium in this landscape. To optimize the design of a cinematic VR system, we have to jointly take into account different stages of the pipeline: data capture, postproduction, coding, and display, due to the strong interdependencies. For instance, a scheme that supports random access to video data for display will affect the performance of a coding algorithm that is based on inter-frame prediction. The format determines the complexity of information extraction in postproduction and controls the capabilities of the rendering algorithm.

Table 1 provides a summary of the relative complexities of different stages of the VR pipeline for different 3D video formats. Although formats with explicit geometry offer immense possibilities during rendering, an underlying 3D model of the real-world scene has to be recovered. This might be suitable for simple scenes, but most real-world scenes are too complex and rich in details for this to be feasible. A setup consisting of over one hundred inward-looking cameras was used to construct a 3D model of the object of interest in [5]. However, for cinematic VR, we require an outward-looking configuration in order to experience an entire environment.

A panoramic format for stereo pairs, called omnistereo, can provide stereo in all directions, but with a fixed head position [1]. A closely related format, referred to as omnivergent stereo, has been shown to be useful for 3D reconstruction applications [6]. In [7], a camera rig is employed to capture scenes with moving objects and the videos from individual cameras are stitched together in the omnistereo format. Since omnistereo uses a fixed head position, it will be necessary to extrapolate if the user's head position changes. This would reduce the output quality due to inpainting artifacts.

In a technique referred to as concentric mosaics, a set of viewing circles with the same center but different radii are sampled and the resulting omnistereo panoramas are stored [2]. It was shown that this can support user head motion in a circular disk. Originally proposed for still imagery, capture, postproduction and streaming of concentric video mosaics are all unresolved challenges. To the best of our knowledge, no camera rigs that directly capture concentric video mosaics have been demonstrated. Furthermore, since ray space interpolation using images from different perspectives may lead to vertical distortions, these must be corrected with depth information [2]. Therefore, although intended to be purely image-based, depth information is additionally required in practical applications.

3. DEPTH AUGMENTED STEREO PANORAMA

In this paper, we propose depth augmented stereo panorama to support head-motion parallax in cinematic VR. This achieves a good trade-off between different stages of the cinematic VR pipeline. Our approach is inspired by video-plus-depth formats proposed by MPEG for autostereoscopic displays [8, 9]. If we use a monoscopic panorama (single center-of-projection) and a corresponding depth map, the production steps are relatively straightforward, but rendering a novel view for a translated head position would involve considerable extrapolation, which may result in visible artifacts. Hence, we propose to use left and right omnistereo texture panoramas and corresponding left and right omnistereo depth maps. This requires depth estimation during postproduction, which is a standard step in existing 3D video production workflows, possibly involving creative manipulations. Furthermore, in order to avoid extrapolation artifacts, the diameter of the camera circle during data capture and production stage is set such that it matches the head motion range to be supported (discussed in detail in Section 5). The transmission stage would involve coding and streaming two texture and two depth panoramas. During rendering, the desired viewports are computed using warping and displayed on the HMD.

Review of omnistereo. A representation to synthesize stereo output looking in all directions (with a zero elevation angle) was proposed by [1]. They use a rotating camera setup and form stereo panoramas by concatenating vertical slits captured at each position. The viewing circle is defined as the circle on which the camera (or the viewer's eyes) are rotating.

Extension to a sphere. The classical formulation of omnistereo has a limited vertical field of view. The points in the world that lie directly above or below the area of the viewing circle do not contribute to the panoramas and hence cannot be captured and represented [1]. Since VR applications require

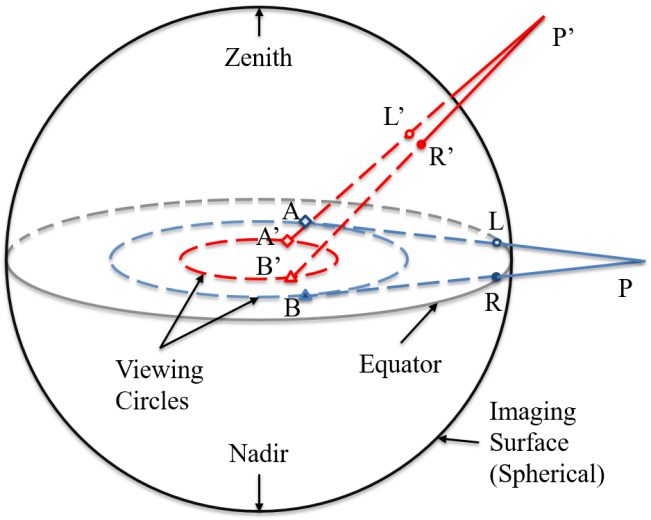

Fig. 1: Spherical omnistereo: A scene point P gets projected by drawing tangents to the viewing circle, meeting at points A and B respectively. The intersections of these tangential rays with the imaging surface determine the image points L and R on the omnistereo panoramas. The primed labels correspond to point P', not on the equatorial plane. Notice how the viewing circle gets smaller as the absolute elevation angle of the scene point increases. This allows us to map the entire scene, unlike [1].

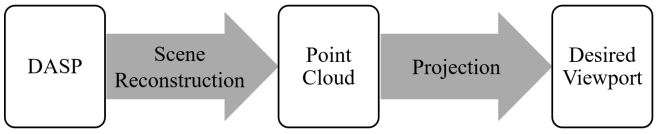

Fig. 2: Rendering using proposed format: (1) Scene reconstruction: Depth augmented stereo panoramas (DASP) are first mapped into point clouds. (2) Projection - The reconstructed points are then projected onto the desired viewport, given the translated viewpoint and the desired viewing direction.

a full 180° vertical and 360° horizontal view, it is important that any representation targeting VR applications be able to represent the whole space. With this in mind, we propose an extension to the classical omnistereo. We modify the imaging surface from cylindrical to spherical, and use a viewing "disk" rather than a viewing circle. For points on the equatorial plane, our representation is identical to the classical omnistereo. However, to map points at a non-zero elevation, we use a smaller viewing radius, scaled by a factor of cosine of the elevation angle. Thus, scene points at different elevations are mapped onto different viewing circles, which collectively form the viewing disk. This allows us to represent the entire 3D space, unlike classical omnistereo.

This means that considering the image information alone, the angular disparity for a given depth reduces and finally becomes zero as we look completely up or down towards the poles [10]. However, in our proposed formulation, we additionally have depth information and therefore can create stereo for all elevations from -90° to 90°. Furthermore, scene contents near the poles often consist of sky, ceiling, floor and so on with little detail, and thus limited disparity perception [11].

Omnistereo panorama and depth. Next, we tackle the problem of potential extrapolation required due to head motion. To this end, we set the diameter of a canonical viewing disk to match the head motion range to be supported. For instance, if we consider an IPD of 6 cm and a head motion range of 24 cm, we set the diameter of the canonical viewing disk to 30 cm. DASP panoramas are constructed as follows. Referring to Figure 1, for each scene point P, two eye positions (denoted as A and B) are chosen by drawing tangents to the appropriate viewing circle (depending on elevation of the scene point). The intersections of these tangents on the imaging surface determine the image points L and R, in the left and right texture panoramas, respectively. The texture panoramas are then augmented with the corresponding depth by storing lengths AP and BP in the left and right depth panoramas, respectively. This provides the ability to synthesize novel views within the predefined operating range. We recognize that postproduction of omnistereo videos with a large radius is a difficult task, but advocate this workflow in contrast to extrapolation during rendering, to minimize potential occlusion artifacts. Figure 2 shows the steps for rendering using the proposed format. These steps are described in the remainder of this section.

Scene reconstruction using proposed format. The goal is to reconstruct a point cloud by mapping every pixel (h, w) in each panorama in DASP to a scene point, represented by a 3D vector $\overrightarrow{r_P}$. Let O be the center of the coordinate system, which is also the center of the imaging sphere and the viewing disk. Let f and v be the radii of the imaging sphere and the viewing disk, respectively. For a pixel (h, w), let $L = (f, \theta_L, \phi_L)$ and $A = (v\cos\phi_L, \theta_A, 0)$ be the corresponding image point and eye position respectively (referring to Figure 1), where $[\theta_L, \phi_L] = \mathcal{L}(h, w)$ and $\mathcal{L}$ is the mapping from equirectangular pixel coordinates to polar image coordinates. The $\cos\phi_L$ term is to capture the shrinking viewing radius with elevation angle. By construction, $\overrightarrow{OA} \cdot \overrightarrow{AL} = 0$. This gives $\theta_A = \theta_L \pm \cos^{-1}(v/f)$, positive for the left eye. Additionally, A, L and P are collinear, and $|\overrightarrow{AP}| = d$, where d is the augmented depth value corresponding to pixel (h, w). Thus, $\overrightarrow{r_P} = \overrightarrow{OA} + \overrightarrow{AP} = \overrightarrow{OA} + (d/|\overrightarrow{AL}|)\overrightarrow{AL}$, which can be computed since we know d and coordinates of A and L. Thus, the coordinates of P can be recovered knowing (h, w, d).

Viewport synthesis

(1) *Viewpoints definition*: Given a translated eye-position and the viewing direction, a target image plane is chosen perpendicular to the viewing direction.

(2) *Depth warping*: Depth information from the point cloud previously computed is forward projected onto the target im-

age plane with the desired viewpoint as the center of projection, using bilinear interpolation at half-pixel accuracy. Disocclusion holes in the warped depth are left unaltered.
(3) *Texture warping*: Using the warped depth, texture on the target image plane is determined by a lookup into DASP texture panoramas. Note that depth warping is source to target, whereas image warping uses target to source mapping. Thus, image warping follows similar steps as depth warping, but in reverse order and uses a bicubic interpolation.
(4) *Hypothesis merging*: For each viewpoint, two warped texture hypotheses are generated: one from the left and one from the right panorama. These two hypotheses are then merged as follows. If a pixel is missing in one of the hypotheses due to occlusion, it is filled using the other hypothesis. If both hypotheses have valid textures, then the hypothesis with the smaller corresponding depth value takes precedence. Majority of the holes get filled in this merging step.
(5) *Hole-filling*: For each of the remaining holes, a vector pointing in the direction of the background is identified using depth information surrounding the hole. The background pixel values are then propagated into the hole area. Since the holes left after the merging step are usually small in their spatial extent, the results produced by this simple pixel propagation scheme appear plausible.

4. EXPERIMENTAL SETUP

The proposed approach is evaluated with both synthetic and real-world data. The experiments with synthetic data are used to validate the functionality of the proposed approach by comparing it with viewports created using known 3D geometry.

Synthetic scene. We use *Blender*[1], an open-source 3D computer graphics software, which contains a realistic physics-based ray-tracer to render a synthetic scene. We adapt Blender to render DASP panoramas, whose construction is explained in Section 3. The known 3D geometry is used to generate DASP panoramas as well as ground truth views at translated head positions with an IPD of 6 cm. We use a viewing disk of radius 15 cm and an imaging sphere of radius 1.9 m, which is a typical distance to the virtual screen when using head-mounted displays. The effectiveness of this format is evaluated in Section 5 by generating views for the same set of translated head positions and comparing the synthesized viewports with the ground truth.

Real-world capture setup. For real-world scenes, most existing approaches for omnidirectional stereo capture use rotating cameras, which are not suitable for dynamic environments. Ideally, a large camera rig would capture rays in all directions such that it supports the desired viewing circle radius. After the capture, several research and engineering challenges have to be addressed to minimize artifacts in stereo panorama stitching [7]. Since the main focus of this paper is the representation of 3D video and not the production itself, we build a simple proof-of-concept prototype to show the capabilities of the DASP format. We use a setup with two commercial off-the-shelf point 360° cameras (Ricoh Theta m15) placed one above the other, i.e., the baseline is vertical. Each of these cameras consists of two fisheye lenses to capture spherical video from a single center-of-projection. Currently, the resolution of this camera is 3584×1792 for image capture and 1920×960 for video capture. It is important to note that no fixed arrangement of two point 360° cameras can capture stereo in all directions because there will be no parallax along the extended baseline. Since stereo in the vicinity of the equator is more important, a vertical baseline is preferred to a horizontal one.

Depth estimation and panorama creation. Capturing spherical images from two vertically displaced viewpoints allows estimating scene depth. Care is taken that the centers of the two spherical cameras align vertically during data capture. Nevertheless, due to limited mechanical precision, slight horizontal offsets can be observed in the captured images. This is addressed by cyclically shifting one image about the vertical axis so that the misalignment is minimized. For small offsets, this step can be considered as an approximation to spherical rectification. Following this step, the coordinate poles can be regarded as the epipoles and epipolar lines become longitudes or columns in equirectangular representation. Stereo matching is thus simplified to a simple 1D search along the column in the equirectangular image, and depth can then be computed using triangulation [12]. Next, omnistereo panoramas and depth maps are created according to Figure 1. Finally, errors in the panoramas and depth maps can be corrected during postproduction, together with creative manipulations to create a more compelling visual experience.

5. RESULTS AND ANALYSIS

Synthetic Scene Setup. An artificial environment is setup consisting of background and foreground objects with known 3D geometry. We use a combination of diffuse (for most objects) and glossy materials (for the silver monkey head). Distances of the objects from the center of the scene vary from roughly 1 m for the closest objects, to up to 5 m for the outer set of spheres. These distances are chosen to be in a range where motion parallax is an important cue. It is also a typical range where important objects are likely to be placed in real scenes. The background is a hemispherical sky texture at infinity, while the floor is textured with a checkerboard pattern where each square has a side of 1 m. The scene is lit by a combination of a directional light, giving realistic shadows, as well as ambient lighting from the environment map (sky). An example of omnistereo panorama in equirectangular format is shown in Figure 3.

[1] https://www.blender.org

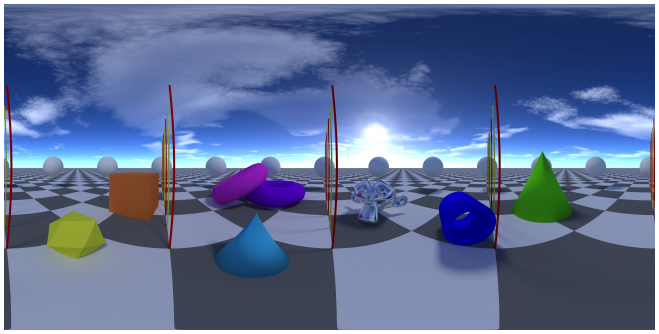

Fig. 3: Example omnistereo panorama (left eye) for an exaggerated IPD of 30 cm. The vertical poles appear bent since the panorama is being depicted on a spherical imaging surface with viewing radius changing with elevation angle.

DASP Viewing Radius	15 cm	12 cm	9 cm	6 cm	3 cm	0 cm
Avg. PSNR (dB)	37.4	37.9	37.9	36.9	33.8	30.9
Std. Dev. (dB)	2.3	2.8	3.2	3.9	5.9	5.8

Table 2: Mean and standard deviation of PSNR, across different head translations and viewing directions, as a function of the representation radius.

Quantitative Evaluation. For quantitative evaluation of the proposed method, viewports with a field-of-view of 90° and a resolution of 1024×1024 were synthesized using the proposed representation. These viewports were then compared against the ground truth in terms of PSNR. The viewing radius for the representation was varied from 15 cm to 0 cm (the latter being equivalent to a central projection). A total of 110 viewports, spanning different eye positions (translation ranging from -15 cm to 15 cm) and viewing directions were synthesized with each of those representations. The results are shown in Figure 4. It can be seen that the PSNR is high as long as the translated viewpoint is within the representation radius, and drops rapidly as the translation exceeds the radius. This is expected since interpolating between views is easier than extrapolating. This motivates the use of a representation radius that is close to the desired range of head-motion.

Table 2 summarizes PSNR values across different representation radii. The mean PSNR is high until a certain critical viewing radius and thereafter drops sharply. This is because input representations with small radii perform very poorly for large values of head translation. Additionally, it can be observed that the standard deviation is the smallest for the largest representation radius and increases progressively as the representation radius decreases. This means that representations with large viewing radii perform well consistently across different head translations.

Subjective Evaluation. An example of a generated viewport is shown in Figure 5. This viewport is generated from a representation with a viewing radius of 15 cm. The viewport corresponds to a translation of 9 cm from the center. We do not show the corresponding ground truth because it is almost

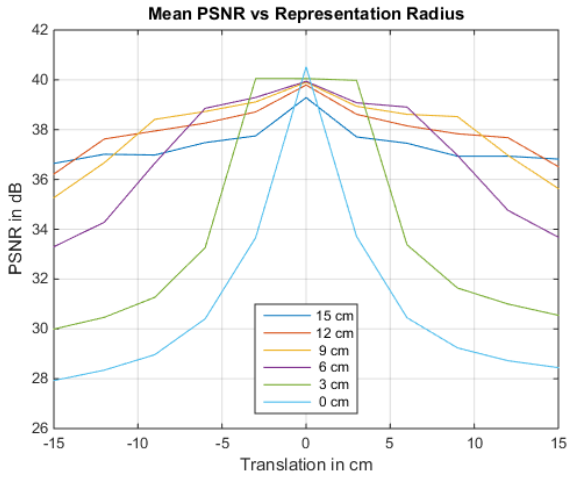

Fig. 4: PSNR averaged across viewing directions as a function of head translation relative to the center. The representation radius is varied from 15 cm to 0 cm and the viewport quality is recorded.

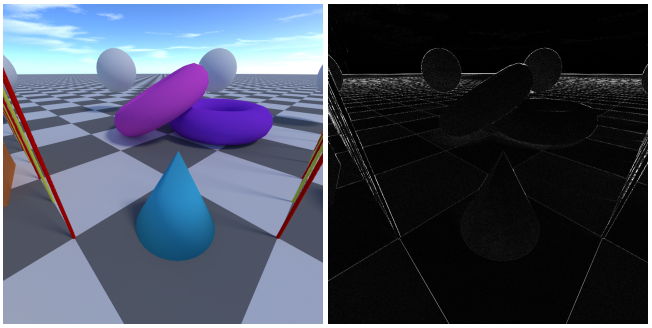

Fig. 5: Left: image synthesized by the proposed representation; Right: pixel-wise absolute difference compared to ground truth, amplified by a factor of 8. This specific example has a PSNR of 36.2 dB with respect to the ground truth.

indistinguishable from our result. Instead, we show the amplified difference between these images. Predictably, most of the errors are small and situated at edges and depth discontinuities. These inconsistencies are due to two factors: (a) the depth maps we employed were not anti-aliased, and (b) depth warping was performed with half-pixel accuracy. These artifacts could be mitigated by performing depth warping with a finer resolution.

Disocclusion Handling. The benefit of using a large representation radius is also clear when it comes to disoccluded regions. Figure 6 shows such an example. If the radius is zero (central projection), all the background pixels that were hidden behind the object become visible when the viewpoint is shifted. Since it is not possible to recover the values of these pixels because they were not included in the original representation, resorting to hole filling is necessary. On the other hand, since DASP maps the scene using stereo panoramas, most of disoccluded regions get filled during the hypothesis merging step.

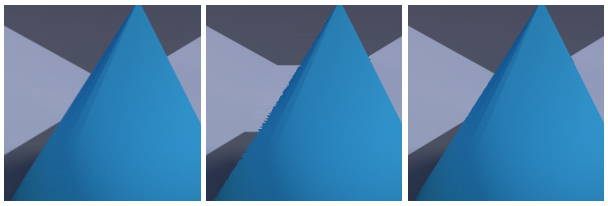

Fig. 6: Disocclusion handling. Left: ground truth, middle: central projection, right: proposed. The central projection produces disocclusion artifacts while the proposed method produces results close to the ground truth.

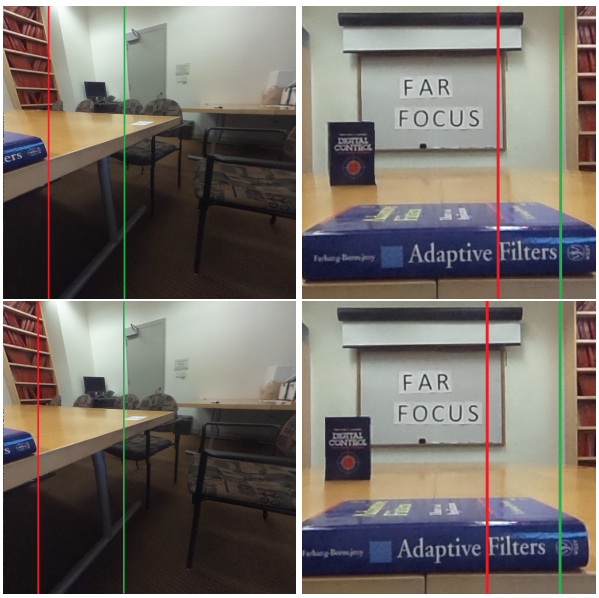

Fig. 7: Outputs for left-right head translation of 8 cm (left column) and 4 cm (right column). Notice how the red line (which tracks a near object) moves much more than the green line (which tracks a far object). The parallax appears plausible, with no evident artifacts.

Real World Examples. Figure 7 shows examples of how the content in viewport changes when there is a left-right head translation. The results are organized in two columns for two different viewing directions. Within each column, the images show viewports for two different head positions. For visualizing the changes between the two rows, a red line is marked in each image at corresponding positions on an object close to the viewpoint and a green line is marked on a far object. From the relative positions of the marked lines, it is evident that the near object shows a large shift while the far object shows almost no shift. This demonstrates that the proposed representation can produce visually plausible parallax.

Similar to other depth-based rendering approaches, a limitation of our method is that errors in depth estimation can lead to artifacts in synthesized views, especially at object boundaries. However, depth estimation is a rapidly evolving area in computer vision. Better stereo matching algorithms, together with active sensing and depth editing can be used to overcome the inaccuracies in depth estimation.

6. CONCLUSIONS

The emergence of VR as a new cinematic medium critically depends on an immersive viewing experience for users. We proposed depth augmented stereo panorama as a novel solution to support head-motion parallax in cinematic VR. The proposed representation supports head-motion parallax for translations along arbitrary directions. The representation radius can be adjusted to increase or decrease the extent of translation supported. We believe this representation to be pivotal in establishing a practical link between emerging omnidirectional capture and display systems.

7. REFERENCES

[1] Shmuel Peleg, Moshe Ben-Ezra, and Yael Pritch, "Omnistereo: Panoramic stereo imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 279–290, 2001.

[2] Heung-Yeung Shum and Li-Wei He, "Rendering with concentric mosaics," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1999, pp. 299–306.

[3] Leonard McMillan and Gary Bishop, "Plenoptic modeling: An image-based rendering system," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. ACM, 1995, pp. 39–46.

[4] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski, "Layered depth images," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. ACM, 1998, pp. 231–242.

[5] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan, "High-quality streamable free-viewpoint video," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 69:1–69:13, July 2015.

[6] Steven M Seitz, Adam Kalai, and Heung-Yeung Shum, "Omnivergent stereo," *International Journal of Computer Vision*, vol. 48, no. 3, pp. 159–172, 2002.

[7] Christian Richardt, Yael Pritch, Henrik Zimmer, and Alexander Sorkine-Hornung, "Megastereo: Constructing high-resolution stereo panoramas," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013*. IEEE, 2013, pp. 1256–1263.

[8] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," *Signal Processing: Image Communication*, vol. 22, no. 2, pp. 217 – 234, 2007, Special issue on three-dimensional video and television.

[9] Karsten Müller, Philipp Merkle, and Thomas Wiegand, "3-D video representation using depth maps," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 643–656, 2011.

[10] Murray Eisenberg and Robert Guy, "A proof of the hairy ball theorem," *American Mathematical Monthly*, pp. 571–574, 1979.

[11] Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, Hans-Peter Seidel, and Wojciech Matusik, "A luminance-contrast-aware disparity model and applications," *ACM Transactions on Graphics, Proceedings SIGGRAPH Asia, Singapore*, vol. 31, no. 6, 2012.

[12] Hansung Kim and Adrian Hilton, "3D scene reconstruction from multiple spherical stereo pairs," *International journal of computer vision*, vol. 104, no. 1, pp. 94–116, 2013.