Towards Perceptual Evaluation of Six Degrees of Freedom Virtual Reality Rendering from Stacked OmniStereo Representation

Jayant Thatte, Stanford University, Stanford, CA, Unites States
Bernd Girod, Stanford University, Stanford, CA, United States

Abstract

Allowing viewers to explore virtual reality in a head-mounted display with six degrees of freedom (6-DoF) greatly enhances the associated immersion and comfort. It makes the experience more compelling compared to a fixed-viewpoint 2-DoF rendering produced by conventional algorithms using data from a stationary camera rig. In this work, we use subjective testing to study the relative importance of, and the interaction between, motion parallax and binocular disparity as depth cues that shape the perception of 3D environments by human viewers. Additionally, we use the recorded head trajectories to estimate the distribution of the head movements of a sedentary viewer exploring a virtual environment with 6-DoF. Finally, we demonstrate a real-time virtual reality rendering system that uses a Stacked OmniStereo intermediary representation to provide a 6-DoF viewing experience by utilizing data from a stationary camera rig. We outline the challenges involved in developing such a system and discuss the limitations of our approach.

Introduction

Cinematic virtual reality is a subfield of virtual reality (VR) that deals with live-action or natural environments captured using a camera system, in contrast to computer generated scenes rendered from synthetic 3D models. With the advent of modern camera rigs, ever-faster compute capability, and a new generation of head-mounted displays (HMDs), cinematic VR is well-poised to enter the mainstream market. However, the lack of an underlying 3D scene model makes it significantly more challenging to render accurate motion parallax in natural VR scenes. As a result, all the live-action VR content available today is rendered from a fixed vantage point disregarding any positional information from the HMD. The resulting mismatch in the perceived motion between the visual and the vestibular systems gives rise to significant discomfort including nausea, headache, and disorientation [1] [2]. Additionally, motion parallax is an important depth cue [3] and rendering VR content without motion parallax also makes the experience less immersive. Furthermore, since the axis of head rotation does not pass through the eyes, head rotation even from a fixed position leads to a small translation of the eyes and therefore cannot be accurately modelled using pure rotation.
The following are the key contributions of our work.

1. We present a subjective study aimed at understanding the contributions of motion parallax and binocular stereopsis to perceptual quality of experience in VR
2. We use the recorded head trajectories of the study participants to estimate the distribution of the head movements of a sedentary viewer immersed in a 6-DoF virtual environment
3. We demonstrate a real-time VR rendering system that provides a 6-DoF viewing experience

The rest of the paper is organized as follows. The following section gives an overview of the related work. The next three sections detail the three contributions of our work: the results of the subjective tests, the estimated head movement distribution, and the proposed real-time rendering system. The last two sections outline the future work and the conclusions respectively.

Related Work

While several studies have demonstrated the importance of motion parallax as a prominent depth cue [4], [3], few have done so in the context of immersive media. The authors in [5] used a fish tank VR setup and showed that rendering motion parallax significantly enhances the viewers' subjective feeling of presence and reduces visual fatigue. In this work, we focus on understanding the perceptual significance of head-motion parallax and binocular vision for content rendered in modern HMDs.

Since constructing and storing a 3D scene mesh per frame is challenging for natural scenes shot using a camera system, several intermediary representations have been proposed to render cinematic VR content with motion parallax. Concentric Mosaics [6] do not require depth reconstruction, but can only provide parallax for horizontal motion, have a large data footprint, are challenging to capture for dynamic scenes, and introduce perspective distortions that cannot be corrected without first estimating the scene depth. Depth Augmented Stereo Panoramas [7] were proposed as a depth-based solution that overcomes many of the drawbacks of Concentric Mosaics. Stacked OmniStereo (SOS) [8] extended this idea to also support vertical motion parallax in addition to horizontal. SOS can synthesize high-quality novel views from vantage points within a predefined 3D volume.

The authors in [9] presented an approach for real-time 6-DoF rendering from natural videos captured using a single moving 360 camera. They use camera motion to infer the scene depth resulting in incorrect rendering for moving objects. In this work, we focus on content that is captured from a single, stationary camera rig and use SOS intermediary representation to provide real-time 6-DoF rendering within a limited viewing volume.

Subjective Testing

This section summarizes the setup, methodology, and the findings of the subjective study that we conducted with the aim of understanding the importance of motion parallax and binocular stereopsis in shaping the perceptual quality of experience in VR.

Test Setup

The subjective study involved showing the participants various modes of VR rendering and asking them to rate the relative quality of experience. We tested 6 different modes – {no motion parallax (3-DoF), only horizontal parallax (5-DoF: supports

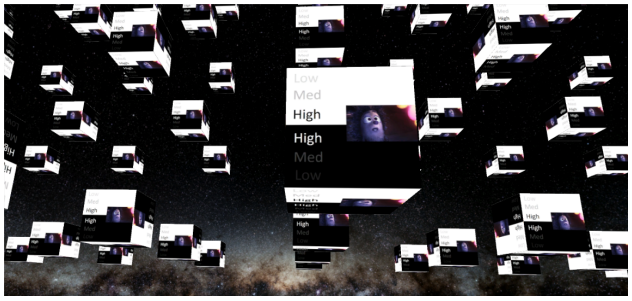

Figure 1. *The test scene used for the subjective study*

lateral and looming/receding motion; but not up-down movement) and full motion parallax (6-DoF)} $\times$ {monoscopic, stereoscopic}. We used Unity[1] to construct and render the test scene and Oculus Rift headset to display the rendering. The test scene comprised many equal sized cubes floating in space against a starry background at infinity (Fig. 1). The cubes were placed between 1 and 4 meters from the viewer. The participants were seated in a swivel chair and asked to explore the scene freely without any constraints. Additionally, during the study, 6-DoF head trajectories (translations along 3 axes and yaw, pitch, and roll) were recorded for each participant. After testing for normal stereo vision, 17 test subjects were allowed to participate in the study.

Testing Methodology

For subjective evaluation, we used the Comparison Category Rating (CCR) method – also known as Double Stimulus Comparison Scale (DSCS) or Pairwise Comparison (PC) method – as described in [10]. In this method, each participant is shown a pair of stimuli (each stimulus being one of the six VR modes listed in the previous subsection) and asked to rate the overall perceived quality of the second stimulus relative to the first on a 7 point scale (+3: much better, ..., 0: the same, ..., -3: much worse). We prefer this rating method over Absolute Category Rating since CCR ratings are minimally influenced by factors such as the subject's opinion of the scene content, display quality, the rendering engine used etc. (section 7.1.3.1 in [10]). The 6 VR settings yield 30 randomized pairs of stimuli covering all possible permutations. Thus, per subject, each stimulus pair gets tested twice – once with each ordering – minimizing the influence of ordering bias on the final rating statistics. Each stimulus is shown to the subjects for a fixed duration of 10 seconds with a resting period of 1.5 seconds, followed by a response screen to record the rating for that pair.

Results

The main findings of the subjective study are summarized in Fig. 2. Looking at the chart, we can make a few key observations.

First, motion parallax is more important than stereo vision. Enabling stereoscopic rendering in a fixed-viewpoint renderer does not seem to enhance the perceptual quality. However, enabling motion parallax leads to a significant increase in the mean opinion score, even if the rendering is still monoscopic – adding horizontal only parallax yields an improvement of ~ 1.3 and additionally enabling vertical motion leads to a further gain of ~ 0.5. A possible explanation for this could be that the effective baseline of lateral head motion is much larger than the inter-pupillary

[1] https://unity3d.com

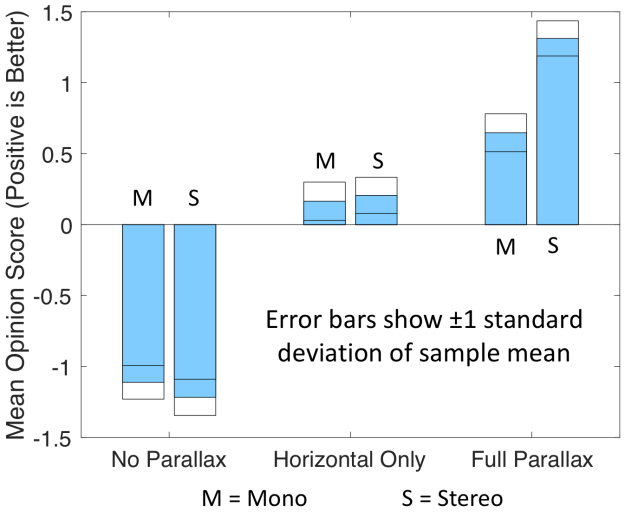

Figure 2. *In this graph, the Y axis shows the mean opinion score (positive is better) and X axis shows different VR modes that were tested. The two bars on the left represent no parallax, the two in the middle are for only horizontal parallax (only lateral and looming/receding motion), and the bars on the right indicate scores for rendering with motion parallax along all 3 dimensions. In each pair, the left bar is for monoscopic rendering and the right is for stereo. The boxes at the end of each bar indicate 1 standard deviation of the mean.*

distance (IPD), and therefore has a stronger impact on depth perception. This trend is consistent with the findings of [3].

Second, stereo rendering causes a significant gain in the perceived quality only in the presence of motion parallax along all 3 dimensions, but not when the parallax is either limited or absent.

Third, in the absence of parallax, the subjects had a small preference for monoscopic rendering over stereoscopic. A possible explanation could be that monoscopic vision is more consistent with the lack of motion parallax, corresponding to a world where everything is far away. However, the difference between the associated mean opinion scores is within 1 standard error and the result must be confirmed with a larger number of test subjects.

The table in Fig. 3 shows the mean comparative scores of each of the 6 rendering modes, with every other mode. The largest absolute values in the table correspond to comparisons of render modes that provide full motion parallax with those that support no parallax. This again indicates that motion parallax is more important than binocular disparity for perceptual quality of experience.

It is important to note that the findings of this study are a function of the test scene composition, specifically the distances of various scene objects from the viewer. Further work is needed to understand how these results vary with changes to the test scene. For instance, in a virtual environment where all objects are far away from the viewer, it could be expected that all the six modes tested here would perform similarly.

Head Movement Distribution

Having an estimate of the typical range of head translation for a sedentary viewer is essential for designing camera systems and algorithms that aim to support motion parallax in immersive video. Despite this need, there is a lack of studies that attempt to systematically estimate such head motion statistics. To bridge this gap, we recorded 6-DoF head trajectories of all the test partic-

M	S	M	S	M	S
0	-0.0882	1.2353	1.2647	1.1471	2.0000
0.0882	0	1.3824	1.1765	1.4118	2.0294
-1.2353	-1.3824	0	0.2059	0.7059	0.8824
-1.2647	-1.1765	-0.2059	0	0.5294	1.0882
-1.1471	-1.4118	-0.7059	-0.5294	0	0.5588
-2.0000	-2.0294	-0.8824	-1.0882	-0.5588	0
No Parallax		Horizontal Onlv		Full Parallax	

Figure 3. *The figure shows a 6×6 matrix of the comparative opinion scores of each of the 6 rendering modes with every other mode, averaged across all the test subjects. An entry in row i, column j gives the mean score (rating scale: -3 to +3, positive is better) of mode j relative to mode i. The 6 render modes in order are: (1) mono, no parallax, (2) stereo, no parallax, (3) mono, horizontal parallax, (4) stereo, horizontal parallax, (5) mono, full parallax, and (6) stereo, full parallax. The diagonal entries are zero by definition since they correspond to the comparsion of one of the 6 settings with itself.*

ipants during our subjective study using Oculus Rift's positional tracking sensor that tracks a constellation of infrared LEDs on the headset and estimates its position and orientation at 90Hz.

For each test subject, a separate head trajectory was recorded per stimulus. Since each stimulus lasts for 10 seconds, each trajectory has 900 data points recorded at 90Hz over the 10 seconds. At the beginning of each new stimulus, the viewer's position was reset to origin thereby recentering the viewer in the virtual environment and the orientation was left unaltered. Since each viewer was shown 30 pairs of stimuli, 60 ten second head trajectories were recorded per test participant. Finally, all of the recorded trajectories were used to estimate a distribution of the horizontal and vertical head translation for a typical sedentary viewer immersed in a 6-DoF virtual environment.

The distributions of head translation in the horizontal and the vertical direction are shown in Fig. 4. We found that over 90% of the horizontal head motion falls within a circle of radius 30 cm and over 90% of the vertical head motion is contained within $\pm$8.5 cm. As expected, a sedentary viewer has a much larger range of horizontal (lateral and looming/receding) translation and a relatively smaller extent of vertical head movement.

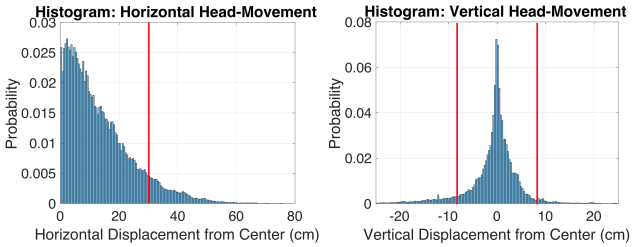

Figure 4. *Histograms of the distribution of head movements of a sedentary viewer exploring a virtual environment with 6-DoF: In each graph, the Y axis represents the bin probability. Left: X axis shows the magnitude of horizontal displacement of the viewer's head position from the origin. Right: X axis shows vertical head translation (positive is up). The red lines indicate the boundaries within which 90% of the head translation is contained.*

Real-time 6-DoF Rendering

This section describes in detail the proposed real-time 6-DoF rendering system. We summarize the design and the implementation of our system, show a few sample rendered viewports, outline the performance of the renderer, and finally discuss the challenges involved and the limitations of our approach.

Stacked OmniStereo Representation

Stacked OmniStereo(SOS) is a depth-based intermediary VR representation that can provide stereo rendering with head-motion parallax along all 3 dimensions from viewpoints lying within a predefined 3D volume. As shown in Fig. 5, SOS comprises 2 vertically separated planes with a pair of texture-plus-depth omnistereo-style panoramas recorded from each plane. SOS panoramas are constructed with a large enough stereo baseline and vertical separation so that the vantage points resulting from a viewer's head motion fall within the cylindrical viewing volume. For our rendering system, we set the vertical separation to 20 cm and the SOS diameter to 40 cm.

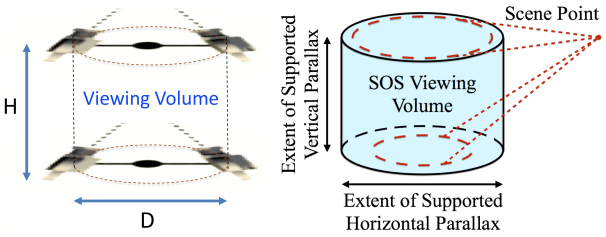

Figure 5. *Left: A visualization of how SOS could be constructed from 4 cameras on a turn table. In practise, however, the representation would be typically constructed from a camera rig system. Right [8]: Each point in the scene gets recorded from 4 distinct viewpoints, as shown.*

Construction of the SOS from a stationary camera rig is described in the original paper that introduced the representation [8]. In this work, we assume that SOS has already been constructed offline and focus on real-time motion parallax rendering using the representation and the 6-DoF HMD pose as inputs.

Rendering from Stacked OmniStereo

The rendering algorithm broadly consists of two parts: (1) hypothesis generation, which runs once per SOS panorama (or 4 times per frame for each eye), and (2) hypothesis merging, which is computed once per frame for each eye.

(1) Hypothesis Generation: This part of the algorithm accepts a single texture-plus-depth SOS panorama and the HMD position and orientation as inputs and generates a hypothesis of what the novel view would look like from the target pose. We obtain one such hypothesis per SOS panorama. Hypothesis generation involves the following steps, as shown in Fig. 6: (a) forward depth warp, (b) depth interpolation, and (c) backward texture lookup.

(1a) Forward depth warp: Each pixel in an SOS panorama corresponds to a specific light ray. Knowing how the SOS panoramas are constructed [8], the direction[2] $\hat{r}$ and the point of incidence A of each such light ray can be uniquely evaluated. We have $\hat{r} = [1, \theta = 2\pi j/w, \phi = \pi i/h]$ and $A = \{\rho\cos\phi, \theta \pm \pi/2, \pm\lambda\}$, where (h, w) is the resolution of the panorama, (i, j) are coordinates of the pixel, ρ and 2λ are SOS radius and height respectively. The first $\pm$ is positive for right panorama and negative for

[2]*Notation*: Parentheses, braces, and square brackets denote Cartesian: (x, y, z), polar: $\{$radius, azimuth, $z\}$, and spherical coordinates: [radius, azimuth, elevation] respectively.

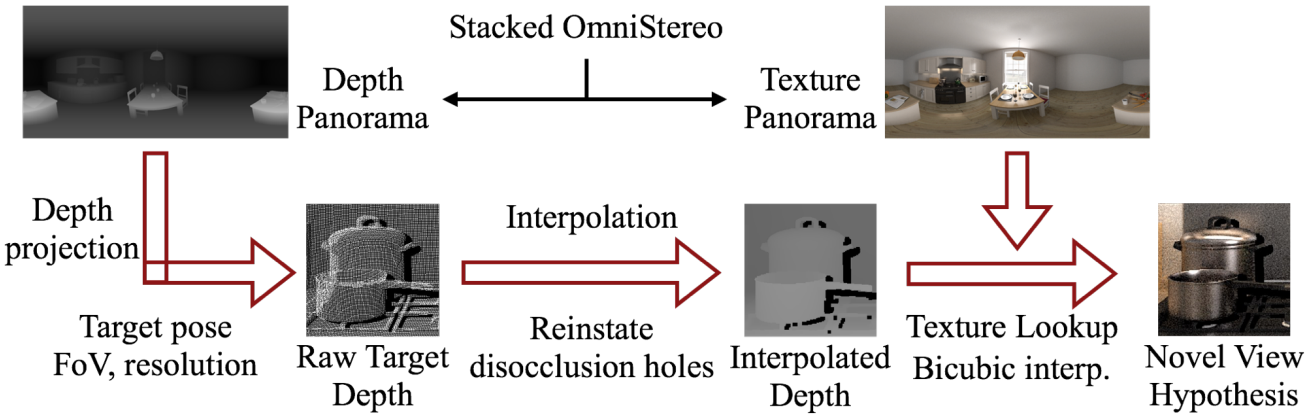

Figure 6. *Hypothesis generation: Each SOS texture-plus-depth panorama is used to generate one hypothesis for target viewport. This is done using 3 steps as shown: forward depth projection, depth interpolation, and backward texture lookup. The 4 SOS panoramas generate 4 such hypotheses.*

left. The second $\pm$ is positive for the top plane and negative for the bottom plane. Knowing the $\hat{r}$, A, and the depth value for each pixel, all the pixels of the SOS depth panorama are warped onto the target viewport to yield a raw target depth map (Fig. 6 (a)).

1(b) Depth Interpolation: The depth values warped using the previous step in general land at non-integer pixel coordinates in the target viewport and need to be interpolated to generate a smooth target depth map. This is often implemented using low pass filtering or by minimizing the total depth error over all projected values under smoothness constraints [9]. However, any filtering or optimization formulation that uses all of the projected depth points identically will yield an incorrect target depth map. This is because the projected depths being points cannot occlude each other and any region of the scene where the foreground and the background overlap will have contributions from both. Naïvely interpolating across such regions results in the foreground depth blending with the background rather than occluding it. To overcome this, we first round the projected pixel coordinates to the nearest half-integer (half-pixel binning) and then interpolate using morphological closing to ensure that foreground occludes background. This is shown in Fig. 7.

(1c) Backward Texture Lookup: At this stage we have the interpolated target viewport depth map. For each pixel in the viewport, we can now invert the calculations in (1a) and compute the coordinates in the SOS source panorama that map to that viewport pixel. The source texture at the computed coordinates is copied over using bicubic interpolation and assigned to the target pixel.

(2) Hypothesis Merging: Repeating the above steps for each SOS panorama produces 4 hypotheses for the target viewport. The final task is to merge these into a single output view (Fig. 8). The hypotheses are identical to each other everywhere except (i) the locations of the disocclusion holes, and (ii) the colors of non-Lambertian surfaces. For pixels where the 4 depth values diverge, smaller depth takes precedence (foreground occludes background). When all the hypotheses have the same depth (within a tolerance), the texture for that pixel is computed as a weighted summation of all the hypotheses. The weights are produced by taking the softmax of the physical distances between the output viewpoint and the vantage points from which that region of the scene was recorded in the 4 SOS panoramas. This allows for non-Lambertian reconstruction to some degree and yields view-dependent specular highlights. An example is shown in Fig. 9.

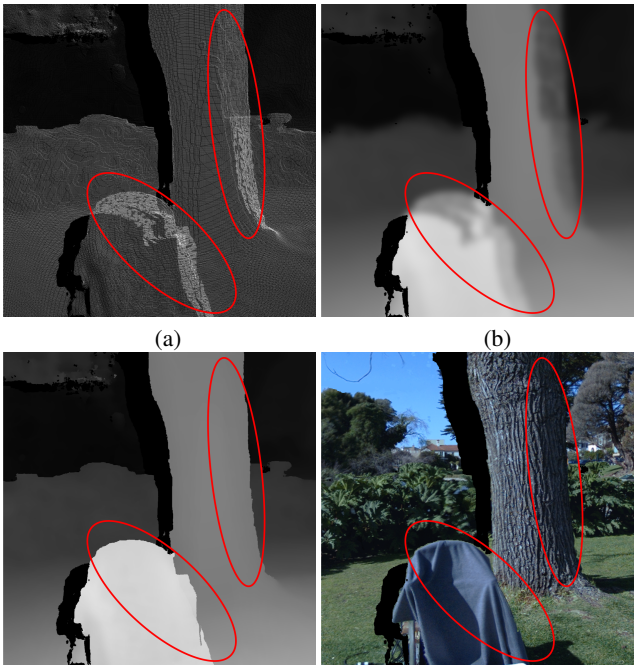

Figure 7. *Depth interpolation: (a) Projected depth from SOS depth panorama. Notice that in the marked areas, background depth projections are visible through the foreground. (b) Results of Gaussian interpolation leads to incorrect depth due to the mixing of the foreground and the background depths. Such incorrect depth would cause distortions in the synthesized view. (c) Depth interpolated using half-pixel binning followed by morphological closing operation (ours). Object boundaries are sharp and foreground occludes background. (d) The corresponding warped image.*

Implementation Details

The real-time renderer was implemented on a 6 GB Nvidia GeForce GTX 980 Ti and Oculus Rift HMD was used to display the results. The depth warping and depth interpolation stages were implemented using CUDA and the texture lookup was achieved using OpenGL pixel shaders. Finally, each frame was postprocessed using OpenGL's frame buffer objects. The final views are predistorted to undo the lens distortion in the HMD. This was implemented using Oculus Rift SDK.

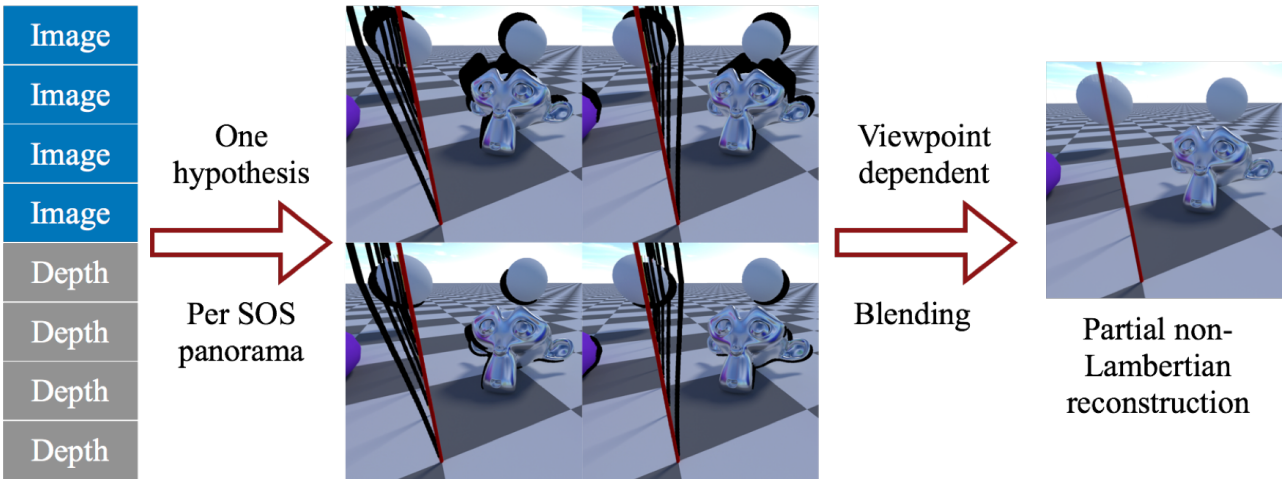

Figure 8. *Hypothesis Merging: 4 SOS panoramas give rise to 4 hypotheses for the target viewport. The disocclusion holes appear in different directions around a foreground object in the different hypotheses, as shown. Thus, the holes get filled in the final view usually without inpainting.*

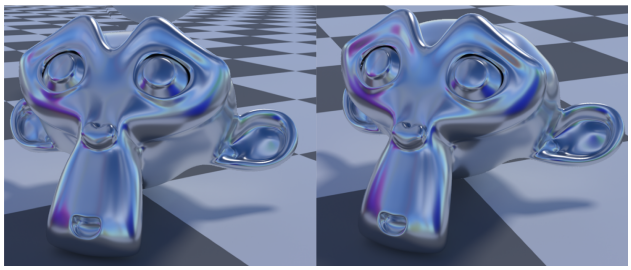

Figure 9. *A shiny object rendered from two distinct vantage points using Stacked OmniStereo: Notice the changes in the specular highlights.*

Results

We are currently able to render stereo views at 45 frames per second with a resolution of 1344×1600 per eye. Morphological depth interpolation is slower than filtering, but produces more accurate target depth maps and hence better novel views. The timing performance can be improved by using – (1) downsampled versions of SOS depth panoramas, and (2) integer-pixel binning rather than half-integer rounding during the depth interpolation stage. The run times for depth warping and depth interpolation stages are quadratic in those parameters respectively. A few examples of synthesized novel views are shown in Fig. 10.

Limitations

Reflections and Refractions: Reflective or transparent surfaces are rendered incorrectly. This is a fundamental limitation since the SOS format assigns a single depth value per pixel. This means that multi-path effects such as reflection and refraction cannot be modeled using SOS. Decomposing the scene into two additive components – Lambertian and reflective/transmittive – with their own depth maps could be a possible solution [11].

Hole-filling: It is possible that a part of the scene that is occluded from all of the four source viewpoints becomes visible from a certain output vantage point. In such cases, the corresponding pixels in the output viewport must be filled using inpainting since the data required to synthesize them is not available in any of the SOS panoramas. In practice, so long as the target viewpoint is within the SOS viewing volume, these holes are quite rare. Additionally, they have a small spatial extent and hence are easy to inpaint. Averaged across 1200 viewports with random positions and orientations, across 6 photorealistic indoor and outdoor scenes, we found that less than 0.1% of the pixels needed hole-filling.

Future Work

This paper presents preliminary work aimed towards understanding the importance of motion parallax in VR and developing a renderer that can support it. The subjective study conducted in this work uses a single test scene. It would be useful to understand how the results of these tests change when different scenes are used. Additionally, the head translation distribution that we estimate is valid for a sedentary viewer. It might be of interest for certain VR applications to estimate a similar distribution for a typical non-sedentary user. Similarly, measuring 6-DoF head trajectories using realistic video content and over a longer duration could provide a better estimate of this distribution. With regards to the rendering system, further work will focus on improving the speed and reducing the aliasing artifacts at object boundaries. Another interesting direction would be to apply the approach in [11] to better handle the reflections and the refractions in the scene.

Conclusions

The subjective study in this work shows that in order to achieve a higher overall quality of experience in virtual reality, rendering accurate motion parallax is in fact more important than providing viewers with stereoscopic vision. Stereopsis seems to significantly boost the perceptual quality only when motion parallax is also provided for lateral, looming, as well as vertical movements. Interestingly, switching to stereo rendering does not seem to produce a significant benefit either when parallax is absent or supported only for the horizontal component of the viewer's head-translation. Furthermore, we show that it is possible to implement a system that can utilize a Stacked OmniStereo intermediary representation to render novel views using a single GPU, and can respond in real-time to the viewer's six degrees of freedom head movements within a predefined viewing volume. Our work shows the importance and the feasibility of developing future virtual reality systems that are capable of rendering immersive video content with accurate, real-time motion parallax.

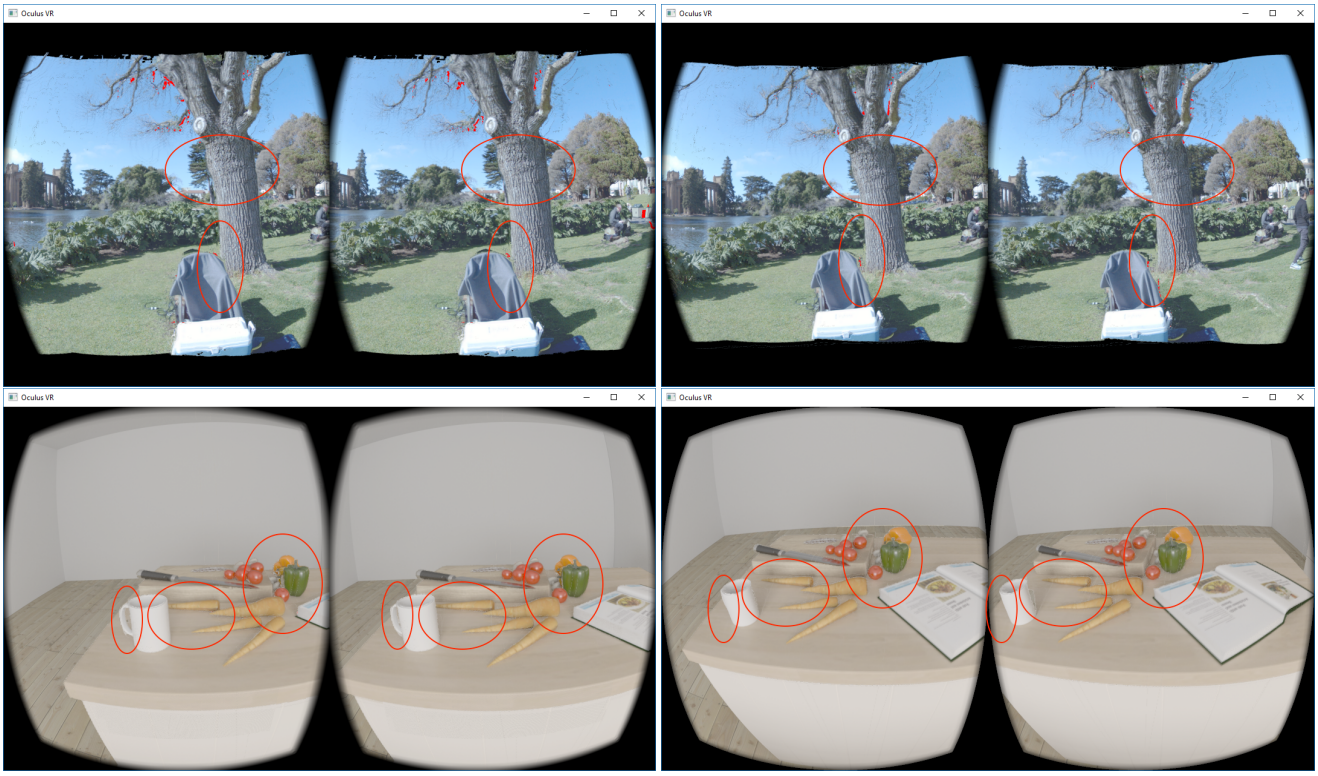

Figure 10. *Motion parallax using the proposed system, rendered on a head-mounted display: Each row shows novel stereo views from two distinct viewpoints within the SOS viewing volume. The ellipses show the regions where the parallax is most noticeable. Top row: A natural scene captured using Facebook Surround 360. Bottom row: A photorealistic synthetic scene.*

References

[1] E. M. Kolasinski, *Simulator Sickness in Virtual Environments*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, 1995.

[2] J. J. LaViola, Jr., "A discussion of cybersickness in virtual environments," *SIGCHI*, vol. 32, no. 1, pp. 47–56, 2000.

[3] J. Cutting and P. Vishton, *Perceiving layout and knowing distances: the interaction, relative potency, and contextual use of different information about depth*. Cambridge, MA, USA: Academic Press, 1995.

[4] B. Rogers and M. Graham, "Motion parallax as an independent cue for depth perception," *Perception*, vol. 8, no. 2, pp. 125–134, 1979.

[5] S. Kongsilp and M. N. Dailey, "Motion parallax from head movement enhances stereoscopic displays by improving presence and decreasing visual fatigue," *Displays*, vol. 49, no. Supplement C, pp. 72 – 79, 2017.

[6] H.-Y. Shum and L.-W. He, "Rendering with concentric mosaics," in *Proc. of the 26th Annual Conf. on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '99, 1999, pp. 299–306.

[7] J. Thatte, J.-B. Boin, H. Lakshman, and B. Girod, "Depth augmented stereo panorama for cinematic virtual reality with head-motion parallax," in *IEEE International Conference on Multimedia & Expo (ICME)*, 2016, pp. 1–6.

[8] J. Thatte, T. Lian, B. Wandell, and B. Girod, "Stacked omnistereo for cinematic virtual reality with six degrees of freedom," in *IEEE International Conference on Visual Communications and Image Processing*, 2017.

[9] J. Huang, Z. Chen, D. Ceylan, and H. Jin, "6-DOF VR videos with a single 360-camera," in *2017 IEEE Virtual Reality (VR)*, March 2017, pp. 37–44.

[10] *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment*, International Telecommunication Union (ITU) Recommendation P.913, March 2016.

[11] S. Sinha, J. Kopf, M. Goesele, D. Scharstein, and R. Szeliski, "Image-based rendering for scenes with reflections," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, 2012.

Author Biography

Jayant Thatte received his B.Tech and M.Tech in electrical engineering from Indian Institute of Technology Madras (2014) along with Philips India Award for the best academic record. He is currently a Ph.D. candidate in electrical engineering at Stanford University. His work is focused on the development of image processing algorithms and systems that provide a more natural and comfortable cinematic virtual reality experience.

Bernd Girod is the Robert L. and Audrey S. Hancock Professor of Electrical Engineering at Stanford University, California. He is a Fellow of IEEE, a EURASIP Fellow, a member of the US National Academy of Engineering, member of the German National Academy of Sciences, as well as a recipient of both the Technical Achievement Award of the IEEE Signal Processing Society and the EURASIP Technical Achievement Award.