A Statistical Model for Disocclusions in Depth-based Novel View Synthesis

Jayant Thatte, Bernd Girod
Department of Electrical Engineering, Stanford University, CA, USA
{jayantt, bgirod}@stanford.edu

***Abstract*—The occurrence of missing regions in output images is a critical issue when rendering a scene from novel vantage points using depth-based view synthesis. These regions typically have to be filled using inpainting algorithms, which are slow and might yield unconvincing results. Understanding the likelihood of the occurrence of these missing regions can help us design better, application-specific data representations and camera systems by knowing which vantage points should be captured and stored to minimize disocclusion holes in the synthesized novel views. In this paper, we propose a statistical model that predicts the likelihood of missing data in synthesized images as a function of the viewpoint translation. Scene-dependent model parameters are efficiently estimated using simple shift and scaling transformations on the source depth images without needing view synthesis.**

***Index Terms*—Novel view synthesis, depth-based rendering, virtual reality, omnistereo, statistical modeling**

I. Introduction

Depth-image-based rendering (DIBR) is one of the most common view synthesis techniques and is widely used in a variety of applications such as free viewpoint television systems, 3D displays, virtual reality (VR) with motion parallax, cinematic postproduction, light-field compression, parallax removal in video conferencing, and 3D cinema, among others.

When the vantage point from which a scene is rendered and that from which it is captured differ from each other, portions of the background that were occluded by foreground scene objects, as seen by the camera, may be visible from the novel vantage point. Since these regions were never observed by the camera, this will lead to missing areas, also known as "disocclusion holes", in the synthesized output. This is one of the major challenges in using DIBR [1][2] since the need for hole-filling degrades the rendering quality and speed.

In this paper, we propose a statistical model that can predict the occurrence of disocclusions as a function of output viewpoint translation using model parameters that implicitly capture the depth statistics of the scene. This allows us to estimate the disocclusions from any vantage point without performing view synthesis. This makes our method well-suited to quickly analyze various data representations and camera geometries without generating hundreds of novel views.

The main contributions of our work are as follows.

1) We propose a statistical model that predicts the occurrence of disocclusion holes in depth-based rendering
2) We demonstrate a method to efficiently estimate the model parameters using the source depth images
3) We extend the model to handle views synthesized from multiple source images
4) We provide practical guidelines for designing data representations as well as camera geometries for virtual reality to minimize disocclusions in synthesized views

II. Related Work

The authors in [1] summarize the typical processing pipeline for a depth-based rendering approach and recognize the need for hole-filling as one of the main challenges. Real-time inpainting methods [3] are rarely depth-aware and do not produce convincing results in DIBR. On the other hand, high-quality inpainting techniques [2][4] do not work in real-time. Additionally, none of these techniques ensure stereo consistency. Yet another approach is to stretch the background to eliminate the need for inpainting [5]. While such techniques could work in real-time, the results look unnatural. Thus, disocclusion holes are an unsolved challenge in DIBR and efforts to minimize their occurrence will prove crucial.

A vital step in estimating the occurrence of disocclusions is understanding the flow field induced in the scene as a result of ego-motion. Studies of such flow fields are presented in [6][7]. A statistical analysis of clutter in 3D scenes is proposed in [8].

III. Disocclusion Modeling

We wish to create a model that uses viewer translation and scene depth statistics to estimate the occurrence of disocclusion holes from the viewer's vantage point without performing view synthesis. Intuitively, we expect lateral, vertical, and looming motion to present differing behavior. For instance, a vertical pillar will cause holes due to lateral viewpoint motion, but not vertical. Moreover, for small translations, we expect the behavior to be linear – if a viewer moves twice as much, we expect the size of the holes to double. To verify our intuition, we rendered 150 texture-plus-depth images across 6 scenes based on 3D models of real-world environments. Using these as inputs, we synthesized novel views simulating the 3 different types of viewer translation. The results summarized in Fig. 1 are indeed consistent with what we expect.

A. The Model

Let $\overrightarrow{T}$ be the translation of a viewer from the source vantage point. Based on the viewing direction, we breakdown $\overrightarrow{T}$ into 3 orthogonal components – looming translation (x), lateral (y), and vertical (z). Based on Fig. 1, we expect the disocclusion probability to be piecewise linear along each axis. We therefore model the disocclusion probability P as

$$P(x, y, z) = ax + by + cz$$

where $a = \{a_+ \text{ if } x \geq 0;\, a_- \text{ else}\}$, $b = \{b_+ \text{ if } y \geq 0;\, b_- \text{ else}\}$, $c = \{c_+ \text{ if } z \geq 0;\, c_- \text{ else}\}$, and $\{a_+, a_-, b_+, b_-, c_+, c_-\}$ are constant model parameters that implicitly capture the geometry of the scene or scenes at hand. Note that by design $P(0,0,0)=0$ because if the viewer and the camera positions are identical, then by definition we cannot have disocclusions. The next task is to estimate $\{a_+, a_-, b_+, b_-, c_+, c_-\}$ so that we can use our model to predict disocclusions. This is addressed in the next two subsections.

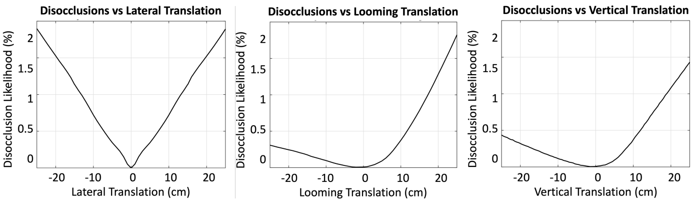

Fig. 1. The horizontal axis shows the lateral/looming/vertical view translation from the camera; The vertical axis shows the percentage of the synthesized viewport that is missing due to disocclusions, averaged across scenes

B. Motion-Driven Flow Fields

Translational ego-motion induces a flow field in any 3D scene. Any depth discontinuities not parallel to the local flow field result in occlusions or disocclusions. Estimating parameters $\{a_+, a_-, b_+, b_-, c_+, c_-\}$ therefore comes down to estimating the flow fields resulting from looming, lateral, and vertical components, respectively, of the viewer's translation.

In this subsection, we describe the flow fields induced by the 3 types of motion in panoramic as well as viewport source images. We describe the process in detail for lateral motion and present the outcomes for looming and vertical translations.

1) Panoramic Source: In panoramic images, viewer motion induces complex flow fields that look like the meridian lines on a globe, diverging in the direction of motion, and converging at a point diametrically opposite. However, if we focus on only a small section of the panorama that the viewer is directly looking at, then the flow can be modeled with relative ease under lateral, looming, or vertical translations.

Lateral Motion: Consider a viewer located at the center of a scene that is a perfect sphere. Looking in any direction, the viewer makes a small lateral step and we record a narrow section of the flow field that the viewer is looking at. We mosaic all these flows to give us a composite flow for lateral motion looking in any direction (Fig. 2). Notice that the flow lines are horizontal and parallel. The composite flow field can therefore be simulated using a cyclic shift of the panorama.

Vertical Motion: A similar analysis shows that the flow field comprises vectors along meridians joining the *zenith* and *nadir* points with the flow magnitude at any point proportional to the cosine of the elevation angle of that point. This can be approximated by splitting the panorama into top and bottom halves and vertically stretching one half while compressing the other to maintain the overall resolution.

Looming/Receding Motion: Scale the panorama up or down vertically, while keeping the horizontal resolution unchanged

2) Viewport Source: In viewports with limited fields-of-view, the flow fields resulting from left, right, up and down ego motion can be approximated simply shifting the source image in the opposite direction. Looming and receding motions are approximated by scaling the image up or down, respectively.

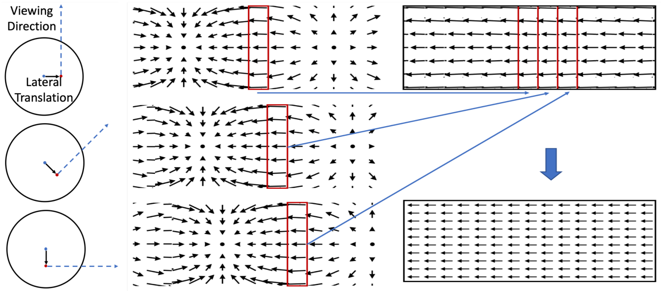

Fig. 2. The figure shows the top view of a viewer located at the center of a panoramic image. Looking in different directions (top to bottom), the viewer takes a step to the right. This generates a flow field (middle, equirectangular) with each red box showing a narrow slit that corresponds to the viewing direction. We obtain the composite flow by mosaicking these slits (top right). This flow can be approximated by parallel, horizontal lines (bottom right).

C. Estimating Model Parameters

The final step is to use the image transformations described previously to approximate the flow fields resulting from each of looming, lateral, and vertical ego-motion. Depth discontinuities that cut across the local flow field result in visibility changes and will be used to estimate model parameters $\{a_+, a_-\}$, $\{b_+, b_-\}$, and $\{c_+, c_-\}$, respectively, as follows. Steps 1 through 4 below are shown in Fig. 3.

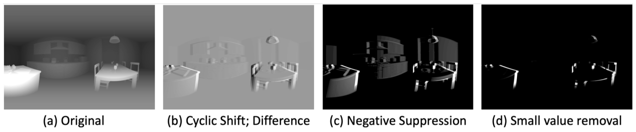

Fig. 3. Left to right: (a) Source depth map (in Diopters), (b) cyclically left-shifted depth map (simulates ego-motion to the right) subtracted from the original, (c) negative values suppressed, and (d) small value removal retains only those regions that will result in holes from a rightward shifted viewpoint.

Step 1: For each type of motion, transform the source depth map (in Diopters) following Sec. III-B. *E.g.* lateral ego-motion corresponds to a cyclic shift of the depth panorama.

Step 2: Subtract the transformed depth from the original. This pulls out the relevant depth edges, with the difference magnitude corresponding to the strength of the discontinuity.

Step 3: Negative difference corresponds to regions that are visible in the source and occluded in the novel view and do not lead to holes, whereas positive difference indicates disocclusions. We therefore suppress negative difference values.

Step 4: Small depth differences result from gradual changes in scene depth rather than from object boundaries. Suppressing small values helps retain only the strong depth edges.

Step 5: For each positive difference, we compute the effective magnitude of viewer translation that would locally result in the applied image transformation. *E.g.* for a cyclic shift s and a depth difference (Diopters) Δd, the effective baseline is given by $B = s/f\Delta d$, where f is the focal length.

Step 6: Quantize the computed baselines. For each bin, store a binary mask marking regions corresponding to that baseline.

Step 7: Vary the value of s in a range (*e.g.* range of negative to positive cyclic shifts for lateral ego-motion) and repeat steps 1-6. The min. and max. values of s are deduced from the range of viewer motion that is relevant to the application at hand.

Step 8: For each baseline bucket, take the union of the binary masks computed in step 6 resulting from all the values taken by s. This gives all the regions that will cause disocclusions when the viewer moves by that baseline value.

Step 9: Counting the areas of these regions for different baselines and for looming, lateral, and vertical motion allow us to estimate $\{a_+, a_-\}$, $\{b_+, b_-\}$, and $\{c_+, c_-\}$, respectively.

IV. Model Applications and Results

This section describes several scenarios in which the proposed model can be applied to gain insights into disocclusion behavior. Sec. V will discuss the practical implications of these observations for designing systems that minimize disocclusions in synthesized views. For our results, we use 6 synthetic scenes based on real-world environments – 4 indoor, 2 outdoor. *Convention*: translation in looming/right/up direction is positive whereas that in receding/left/down direction is negative.

A. Application: Singe Panorama as Source Image

In VR applications, DIBR systems typically use a panoramic input to synthesize novel views. In this subsection, we use as input a single omnistereo texture-plus-depth panorama [9][10] (left or right) with a radius of 15 cm and predict disocclusions from translated viewpoints. In each plot, the model prediction is shown as a solid line. The points show the disocclusion data generated for model verification using view synthesis.

Vertical Motion: Across all scenes, the model consistently predicts that disocclusions increase faster due to upward motion than downward, (Fig. 4, top left). This is because in natural environments, objects are located on a ground plane and moving up lets a viewer look past the foreground objects.

Looming Motion: The model shows that in scenes containing mostly tall objects (typical outdoor scenes), receding motion causes more holes than looming (Fig. 4, top right, red line). The effect is the opposite for indoor scene where most objects are located below the eye-level (black line).

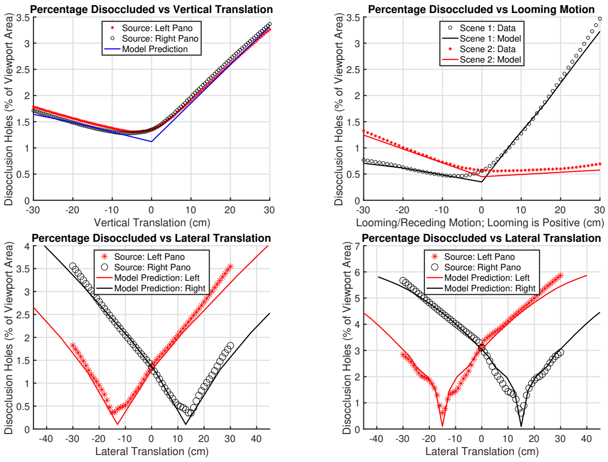

Fig. 4. The horizontal axis represents viewpoint translation and the vertical axis shows the disocclusion likelihood as percentage. Top left: vertical motion, top right: looming motion for a typical indoor and outdoor scene, bottom left: lateral motion, bottom right: lateral motion in a scene with thin objects.

Lateral Motion: In stereo panoramas, since the scene is imaged from points along the circumference, we expect the least disocclusions when the novel view lies on the periphery. This corresponds to a lateral translation of 15 cm to the left or right when the view is synthesized using the left or the right panorama respectively (Fig.4, bottom left).

Thin Objects: One of our scenes contain numerous thin, vertical poles. Thin scene objects tend to violate the linear relation between viewer translation and disocclusions. The method described in Sec. III-C is also correctly able to predict this nonlinear behavior (Fig.4, bottom right).

B. Model Improvement: Per-Pixel Viewpoint Translation

Stereo panoramas are multi-perspective images. That is, different pixels in the panorama are captured from different viewpoints on the viewing disk of the panorama [10]. This implies that when a novel view is synthesized using a panoramic source image, each output pixel, in effect, has a different viewpoint translation (Fig. 5, left). Taking into account the pixel-wise translation vectors makes the model predictions align more closely with experimental data (Fig. 5, right).

C. Model Extension: Multiple Sources

A common application of DIBR is synthesizing novel views that interpolate between two source views. Extending our model to a pair of source images requires the correlation between the disocclusions when a novel view is synthesized from each source individually (disocclusion indicator variables A_1, A_2). $E[A_{joint}] = E[A_1 \cap A_2] = E[A_1]E[A_2] + \sigma_1\sigma_2\rho$, where σ_1, σ_2 indicate standard deviations and ρ is the correlation.

To that end, we synthesized novel views using each individual image from stereo pairs with varying baselines. In each case, lateral motion is along the line joining the source vantage points and looming motion is orthogonal to it. We normalize all translations by the stereo baseline so that the source views are located at ±1. We found that upon such normalization, the correlation curves are almost identical across all different scenes and baselines, for both lateral and looming motion (Fig. 6). This is likely because the depth statistics of most natural scenes are self-similar over change of scale.

Using these correlation curves for lateral and looming motion, and using the model improvement described in Sec. IV-B, we are able to predict the disocclusion behavior for views synthesized by jointly using both the images of stereo panorama pairs for varying baselines (Fig. 7).

D. Application: Constructing Panoramas from Camera Rigs

Real-world environments can be viewed in stereo in VR when the raw images from camera rigs are used to construct stereo panoramas. Wide baseline, vertically separated panoramas can additionally respectively provide horizontal and vertical head-motion parallax in VR [11]. Our model can be used to analyze camera geometries.

As examples, we analyze 2 camera systems – Facebook Surround 360 (*https://github.com/facebook/Surround360*) (FB360) and a 2-tier system created by stacking 2 FB360 rigs one-above-the-other with a vertical separation of 20 cm. We use the model to predict the disocclusions in the panoramas constructed from these cameras as a function of the elevations and the stereo baselines of these panoramas.

Our model predicts a threshold panorama radius of about 13.7 cm, within which the constructed panoramas have low disocclusions and increase linearly outside (Fig. 8, left). This plot can be used to determine the largest baseline possible for

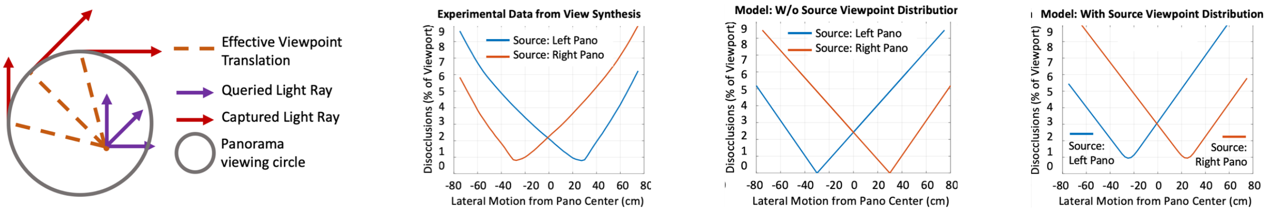

Fig. 5. Left to right: (1) Each light ray queried by the target viewport is captured from a distinct source viewpoint, (2) experimental disocclusion data generated by view synthesis, (3) model prediction without accounting for the source viewpoint spread, (4) improved model prediction.

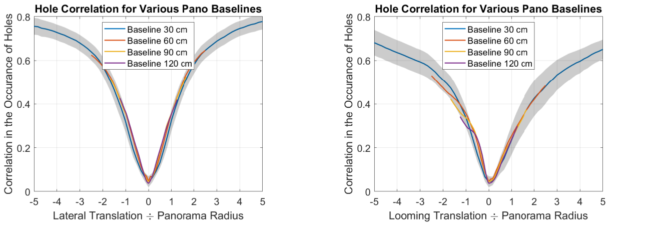

Fig. 6. The vertical axis shows the correlation in the occurrence of holes in novel views synthesized using each image from a stereo pair, one at a time. The horizontal axis shows viewpoint translation normalized by the source baseline; different lines show various stereo baselines. The gray regions show 1 standard deviation across scenes. Left: Lateral motion; Right: looming.

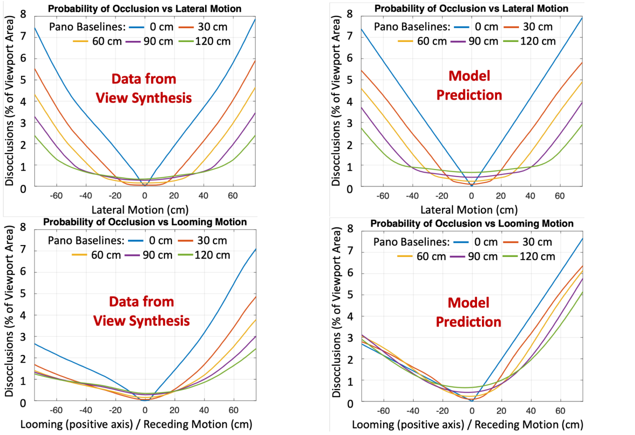

Fig. 7. Novel views from stereo panorama pairs: The vertical axis shows the disocclusion likelihood (as percentage), the horizontal axis shows translation (cm), and the different lines show various stereo baselines. Ground truth on the left, model predictions on the right. Top: Lateral motion; Bottom: Looming

the panoramas while keeping disocclusions under a predefined target. This in turn sets the range of a viewer's head-motion that can be supported with this camera geometry.

Next, we estimate disocclusions in panoramas as a function of their vertical displacement from the rig center. We use panoramas with a stereo baseline (diameter) of 13.7 cm$\times 2 =$ 27.4 cm. We observe that (1) 2-tier camera geometries are better at providing vertical parallax and (2) panoramas displaced upward have more holes than those displaced downward.

V. Conclusions

The occurrence of disocclusion holes in synthesized novel views can be predicted efficiently without performing view synthesis and instead analyzing the depth discontinuities in the source image in the context of the given viewpoint translation.

Using the model, we observe that, in general, more disocclusions are created when a viewer moves upward in a scene than downward. Therefore, scenes should be recorded and stored from vantage points towards the top end of the expected range of vertical ego-motion. Similarly, for indoor scenes, looming motion creates more holes than receding. The trend is reversed for outdoor scenes. This should be used to inform the placement of the viewer relative to the source vantage points and the objects of interest in the scene.

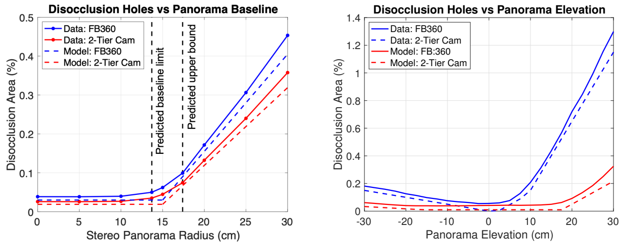

Fig. 8. The vertical axis shows disocclusions (%), solid lines show data generated from panorama synthesis, dashed lines show model predictions. Blue lines indicate analysis for FB360 and red lines for the 2-tier camera. Left: The horizontal axis show the radii of the constructed panoramas. The black lines indicate predictions for the cut-off radius beyond which disocclusions are expected to increase linearly. Right: The horizontal axis shows the vertical displacement of the panoramas from the center of each camera rig.

When using a pair of source viewport images, disocclusions are low when interpolating between the views and increase linearly outside. In general, for multiple sources, disocclusions are low within the convex hull of the source viewpoints. When using stereo panoramas, disocclusions are minimal within the viewing circle of the panoramas and increase linearly outside. The depth-augmented panoramas should hence be wide enough to accommodate the entire range of head-motion.

In the context of cinematic VR, given a camera rig geometry, we can use the model to predict a limit on the stereo baseline and vertical separation for panoramas constructed using the camera, such that the disocclusions in the panoramas remain under a predefined threshold. This will decide the range of viewer head-motion that can be supported with that camera.

References

[1] C. L. Zitnick et al., "High-quality video view interpolation using a layered representation," in *ACM SIGGRAPH*, 2004, pp. 600–608.

[2] P. Ndjiki-Nya et al., "Depth image-based rendering with advanced texture synthesis for 3-d video," *IEEE Trans. on Multimedia*, June 2011.

[3] J. Herling and W. Broll, "High-quality real-time video inpaintingwith pixmix," *IEEE Trans. on Visualization and Computer Graphics*, 2014.

[4] P. Buyssens et al., "Depth-aware patch-based image disocclusion for virtual view synthesis," in *ACM SIGGRAPH Asia Technical Briefs*, 2015.

[5] A. Smolic et al., "Disparity-aware stereo 3D production tools," in *2011 Conference for Visual Media Production*, Nov 2011, pp. 165–173.

[6] J. Gluckman and S. K. Nayar, "Ego-motion and omnidirectional cameras," in *Sixth International Conference on Computer Vision*, Jan 1998.

[7] J. E. Cutting, "Pictorial representations and their development in the work of james gibson," *Perception*, vol. 29, no. 6, pp. 635–648, 2000.

[8] M. S. Langer and F. Mannan, "Visibility in three-dimensional cluttered scenes," *J. Opt. Soc. Am. A*, vol. 29, no. 9, pp. 1794–1807, Sep 2012.

[9] S. Peleg et al., "Omnistereo: panoramic stereo imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, March 2001.

[10] J. Thatte et al., "Depth augmented stereo panorama for cinematic virtual reality with head-motion parallax," in *IEEE ICME*, July 2016, pp. 1–6.

[11] J. Thatte et al, "Stacked omnistereo for virtual reality with six degrees of freedom," in *IEEE VCIP*, Dec 2017, pp. 1–4.