SPATIALLY CONSISTENT VIEW SYNTHESIS WITH COORDINATE ALIGNMENT

Lam Tran†  Ramsin Khoshabeh†  Ankit Jain†  Christopher Pal⋆  Truong Nguyen†  
† ECE Department, University of California San Diego, La Jolla CA 92093  
⋆École Polytechnique de Montréal, Montréal, QC, Canada.

ABSTRACT

In this paper, we propose a novel method that uses coordinate alignment and background pixel extraction to synthesize highly accurate and spatially consistent intermediate views from a pair of stereo images and disparity maps. In contrast to the traditional depth image-based rendering (DIBR) method, where useful background pixels are discarded in the warping process, the proposed method extracts these background pixels and uses them as candidates for an exemplar-based image in-painting technique (EBIIT) to synthesize realistic content in disocclusion regions. Our second contribution is a coordinate alignment algorithm that aligns disocclusion regions in each view together and simultaneously synthesizes disocclusion regions to enhance spatial consistency across all virtual views. The proposed method compares favorably in quantitative measures to those obtained by existing techniques and has superior potential for stereo-to-multiview conversion.

Index Terms— autostereoscopic display, disparity map, stereo images, and view synthesis.

1. INTRODUCTION

Depth image-based rendering (DIBR) is commonly used to generate new virtual viewpoints for autostereoscopic displays. The three main steps of the DIBR framework are: preprocessing of disparity maps, image warping, and hole filling. The challenge of this method is the hole filling step in which one must restore the occluded pixels in the new virtual view. Disocclusion refers to the process of recovering scene information obstructed by visible points and we refer to any occluded pixels that have been restored as disoccluded pixels.

In [1], a Gaussian filter is used to smooth the disparity map in the preprocessing step to eliminate disoccluded pixels. This method is easy to implement and computationally efficient; however, the synthesized images are unrealistic due to geometry distortion especially when the disoccluded region is large.

A better approach [2] combines depth-based-hole-filling and in-painting to restore the disoccluded pixels more accurately compared to prior in-painting methods without using depth information. While in-painting is a powerful tool to restore small disoccluded regions, it produces a notable blur and can become computationally inefficient when the disoccluded region is large in the virtual view. Both of these methods described above produce visual artifacts shown in [3] and degrade the 3D effect when the synthesized images are interlaced into a multiview image.

The work in [4] oversegments the virtual view, and pixels in segments connected to each disoccluded region are used with edge information to fill in the regions. The filling process is done by merging each disoccluded pixel to an attached segment and selecting a pixel in a neighboring segment to fill in that disoccluded pixel. This method fails when there is no edge and complicate textures are present, making it difficult to merge disoccluded pixels to the correct segment. The method in [5] suggests generating two virtual views at the extreme left and right and uses view interpolation to generate the intermediate views. To enhance spatio-temporal consistency, the authors add disparity to calculate priority that determines the filling order of occluded pixels; however, this method requires additional hole filling to each intermediate view.

In this work, we extend our work [4] and propose a new method to synthesize consistent intermediate stereo images from a pair of stereo images that achieves high accuracy in quantitative metrics. We propose to warp the pixels and their
2. PIXEL CLASSIFICATION

Each pixel in the virtual view is classified as stable, unstable, or disoccluded. Stable pixels have only one pixel candidate and remain constant throughout the inference process. Unstable pixels have multiple pixel candidates and the candidate is selected that matches best with its neighboring pixels. Finally, disocclusion pixels have no pixel candidate and are occluded in both reference images. The candidates are obtained with pixel extraction and the disoccluded pixels are aligned and filled in with EBIIT.

3. SYNTHESIZE INITIAL VIRTUAL VIEW

In the initial step, color segments in the left $I_L$ and right $I_R$ reference images are extracted by [7]. In each segment, pixels with disparity that exceed $\pm 20$ from the mode are labelled as occluded, and the occluded pixels are then filled with the disparity of the nearest non-occluded neighboring pixels in the segment to generate left $D_l$ and right $D_r$ refined disparity maps. After the initial disparity refinement step, a set of $n$ predefined virtual camera positions $\theta_i$ are defined, and each $\theta_i \in \{0, 1\}$ is used to compute two disparity maps for virtual view $i$ as

$$D_{L,i} = \theta_i D_l \text{ and } D_{R,i} = (1 - \theta_i) D_r.$$  \hspace{2cm} (1)

These disparity maps are used to generate placement matrices to warp pixels in reference images to the virtual view. To evaluate the proposed method with the Middlebury’s [8, 9] data set the three virtual cameras are positioned at $\theta = \{1/4, 1/2, 3/4\}$ and the reference left and right camera are positioned at 0 and 1 respectively.

The disparity maps for each virtual view are used to compute placement matrices that warp pixels from the reference view to the virtual view as described in our previous work [4] for the stable pixels $P_{L}^1$ and $P_{R}^1$ and the unstable pixels $P_{L}^2$ and $P_{R}^2$. The advantage of warping pixels using placement matrices is that the refinement of small cracks and round off errors are done on the coordinate of the image to preserve texture of the virtual view.

The remainder of this paper is organized as follows. Section 2 describes the type of pixels in the virtual view. Section 3 describes how to generate the initial virtual view with placement matrices. Section 4 describes coordinate alignment and background pixel extraction. Section 5 describes the hole filling process. Sections 6 shows the simulation and experiment results, and section 7 concludes the paper.

4. SPATIAL CONSISTENCY WITH COORDINATE ALIGNMENT AND PIXEL EXTRACTION

The three initial virtual views $I_1$, $I_2$, and $I_3$ are synthesized using the placement matrices. Each occluded region appears in all three views but at different coordinate locations shown in the first three images of row one in Fig. 2. To enforce consistency of synthesized regions across all views, the proposed method aligns the coordinate of the disoccluded region onto the destination virtual view and backtracks these coordinates to the reference views to extract background neighboring pixels used to fill in the disoccluded pixels. Both of these methods will be explained in the next two subsections.

4.1. Coordinate Alignment to a Target View

To enforce consistency of synthesized pixels in the same occluded region across all views, each virtual view is partitioned into $n$ segments. When a disoccluded region appears in segment $k$, virtual view $k$ is selected as the destination view. The
neighboring pixel coordinates in \( C_{s,k} \) are extracted for the selected disoccluded region to be synthesized. The coordinates from other virtual views are then aligned with the neighboring coordinates in destination virtual view.

The alignment matrix indicates whether pixel \( \beta \) in virtual view \( k \) is aligned to a pixel in virtual view \( l \) for an arbitrary row \( i \) is computed as

\[
A_{k,l}(i,\beta) = \begin{cases} 
1 & \text{if } \Phi(\beta) = 0 \\ 
0 & \text{otherwise} 
\end{cases}
\]

such that

\[
\Phi(\beta) = F_k(\beta) - F_l(\alpha^*) + C_k(\beta) - C_l(\alpha^*),
\]

where

\[
\alpha^* = \arg \min_{\alpha} (F_k(\beta) - F_l(\alpha)) + (C_k(\beta) - C_l(\alpha)).
\]

The process is repeated for all of the boundary coordinates of the disoccluded region in view \( k \) with each of the virtual view to be synthesized. After all the neighboring pixels are aligned, we perform background pixel extractions to recover all of the background pixels as described in the next section.

### 4.2. Pixel Extraction From Reference Images

The most common method for stereo-to-multiview conversion is to first warp the pixels to the virtual view and perform hole filling after all the pixels are warped. This method removes useful information from the reference images to aid the hole filling process. To recover information from the reference images, the proposed method extracts pixels in the reference images that are neighboring pixels to the disoccluded pixels in the virtual view.

However the neighboring pixels to be extracted might be obstructed in the virtual view. To extract the neighboring pixels, a reference indicator map is used to track which reference image was used to warp pixels to the virtual view. The background pixel extraction process for view \( k \) starts on the boundary of the disoccluded region and is extracted from the reference view as follows:

\[
E_k(x, y) = \begin{cases} 
0 & \text{if } F_k(x, y) = 0 \\ 
I_L(x, y - D_{L,k}(x, y)) & \text{if } F_k(x, y) = 1 \\ 
I_R(x, y + D_{R,k}(x, y)) & \text{if } F_k(x, y) = 2 
\end{cases}
\]

and stored the extracted pixels in \( E_k \). This process continues until the disparity levels between neighboring pixels disagree. (d), (e) and (f) of Fig. 2 show the background pixel extraction of a disoccluded region across 3 virtual views.

### 4.3. Fusion of Extracted Pixels and Alignment Matrices

After the alignment and pixel extraction processes are completed, each extracted region is aligned and fused together to view \( k \) as follows

\[
I_F = \sum_{i \neq k} E_i A_{i,k} + E_k.
\]

Fig 2 (g) shows examples of fused background pixel extraction. The proposed method compared to DBIR not only provides more information but removes irrelevant objects to simplify the hole filling process of the disoccluded region. The alignment step only requires each occluded region to be synthesized once independent of how many intermediate views one wishes to synthesize and also enhances spatial consistency across all intermediate views. In contrast to the conventional DBIR method where each view is synthesized independently as shown in (a), (b), and (c) of Fig. 2, the proposed method only synthesizes one view, as shown in the last column of each row in of Fig. 2.

### 5. HOLE FILLING WITH PATCH MATCHING

After the disoccluded region is processed by the proposed method, the image is partitioned into patches of 5×5 and filled with the exemplar-based image in-painting technique [6]. We simplified the filling priority calculation of each patch center at \( x \) and \( y \) to

\[
P(x, y) = \frac{|G_\parallel(x, y)| + |G_\parallel(x, y)|}{|G_\parallel(x, y)|G_\parallel(x, y)| + \epsilon},
\]

where \( G_\parallel \) and \( G_\parallel \) are the horizontal and vertical gradients of the image and \( \epsilon \) is a constant that is not zero. After the patches are filled, the disoccluded region is shifted back to each of the virtual views shown in row 2 of Fig. 1.

### 6. EXPERIMENT AND RESULTS

In our experiment, the images from the Middlebury data set [8, 9] are used to evaluate the proposed method. This data set consists of two ground truth disparity maps and 7 stereo images for each data set. Views 1 and 5 are used with the disparity maps to synthesize virtual view 3. These synthesized images are compared with the ground truth to quantify the algorithm performance based on PSNR and SSIM[10] shown in Table 1. The SSIM index value 1 is only reachable when two images are identical and a higher PSNR normally indicates that the synthesized image is of higher quality. Fig. 3 shows the hole filling of the disocclusion regions of the proposed method compared with [4] and the ground truth.

The experimental results show that the proposed method achieved on average over 34 dB in PSNR and 0.95 index value in SSIM on the Middlebury stereo data sets. The results reported by [2, 3] are from different stereo data sets and are not compared with our experiment. We compare to the result in [4], which currently achieves the highest PSNR
and SSIM value, with our framework in Table 2. Although the quantitative measure only improved by a small factor, the proposed method synthesizes the intermediate view more accurately with textures and edges preserved shown in the above figure. Due to space limitation, only portion of the experimental results are shown in this paper. The full size images and full experiment can be viewed or downloaded on our website ([http://videoprocessing.ucsd.edu/~lamtran/ICASSP2011.html](http://videoprocessing.ucsd.edu/~lamtran/ICASSP2011.html)).

### 7. CONCLUSION

In this paper, we have proposed a framework to synthesize spatially consistent intermediate views. The proposed method recovers background pixels discarded by the warping processes to provide these pixels for the hole filling process to synthesize the disoccluded regions more accurately. The alignment enhances spatial consistency between immediate views and reduces the filling complexity by requiring each disoccluded region to be filled only once independent of how many immediate views one wishes to synthesize. The proposed method yields better objective results as well as producing synthesized views with better subjective quality.

### 8. REFERENCES


### Table 1: Experiment results for Middlebury Stereo Database.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Proposed PSNR</th>
<th>Proposed SSIM</th>
<th>Tran et al. [4] PSNR</th>
<th>Tran et al. [4] SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art</td>
<td>32.82</td>
<td>0.95</td>
<td>32.66</td>
<td>0.95</td>
</tr>
<tr>
<td>Books</td>
<td>31.06</td>
<td>0.95</td>
<td>30.92</td>
<td>0.93</td>
</tr>
<tr>
<td>Cloth 1</td>
<td>35.99</td>
<td>0.97</td>
<td>35.99</td>
<td>0.97</td>
</tr>
<tr>
<td>Dolls</td>
<td>33.22</td>
<td>0.95</td>
<td>33.05</td>
<td>0.95</td>
</tr>
<tr>
<td>Laundry</td>
<td>32.16</td>
<td>0.95</td>
<td>32.13</td>
<td>0.95</td>
</tr>
<tr>
<td>Moebius</td>
<td>34.58</td>
<td>0.96</td>
<td>34.30</td>
<td>0.94</td>
</tr>
<tr>
<td>Monopoly</td>
<td>32.27</td>
<td>0.95</td>
<td>32.19</td>
<td>0.95</td>
</tr>
<tr>
<td>Reindeer</td>
<td>34.10</td>
<td>0.95</td>
<td>33.70</td>
<td>0.94</td>
</tr>
<tr>
<td>Plastic</td>
<td>37.90</td>
<td>0.98</td>
<td>37.77</td>
<td>0.98</td>
</tr>
<tr>
<td>Wood</td>
<td>37.50</td>
<td>0.94</td>
<td>37.47</td>
<td>0.94</td>
</tr>
</tbody>
</table>

### Table 2: Experiment results comparison.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR (dB)</th>
<th>Tran et al. [4] PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>34.16</td>
<td>34.02</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.955</td>
<td>0.946</td>
</tr>
</tbody>
</table>