A Depth-Guided Inpainting Scheme Based on Foreground Depth-Layer Removal for High Quality 2D to 3D Video Conversion

Jangwon CHOI\(^{[a]}\), Nonmember, Yoonsik CHOE\(^{[b]}\), Member, and Yong-Goo KIM\(^{[c]}\), Nonmember

SUMMARY This letter proposes a novel depth-guided inpainting scheme for the high quality hole-filling in 2D-to-3D video conversion. The proposed scheme detects and removes foreground depth layers in an image patch, enabling appropriate patch formation using only disoccluded background information. This background only patch formation helps to avoid the propagation of wrong depths over hole area, and thus improve the overall quality of converted 3D video experience. Experimental results demonstrate the proposed scheme provides visually much more pleasing inpainting results with better preserved object edges compared to the state-of-the-art depth-guided inpainting schemes.

key words: depth-guided inpainting, foreground depth layer, hole filling, depth image-based rendering, 2D to 3D conversion

1. Introduction

Although the creation of high quality 3D video is technically feasible nowadays, 2D to 3D conversion has gained more interests because the production cost of 3D video is still high and a substantial amount of valuable films was already produced as 2D video. Typically, 3D conversion of 2D video starts with the creation of depth map to assign each pixel the distance from a viewpoint. Then, DIBR (Depth Image-Based Rendering) technique [1] is applied to generating a new viewpoint image via viewpoint projection based on the reference image and its associated depth map. This viewpoint projection brings about a horizontal disparity for each pixel, resulting in pitted area to appear in the vicinity of object boundaries. Since the information of this pitted area (i.e. disoccluded background) affects the depth perception of human binocular vision, the quality of hole-filling is crucial to the visual comfort as well as the overall quality of converted 3D video experience.

In order to fill such holes in new viewpoint image, there have been various approaches in the literature, but most of them focused on the viewpoint interpolation in Free-viewpoint TV environment. In this environment, the generated viewpoint is in between two neighboring reference viewpoints, and thus the holes from DIBR using one reference image can be reliably filled in by using the other reference viewpoint image [2], [3]. In contrast, the hole-filling for 2D to 3D conversion employs only one reference viewpoint, bringing about inherent limits on the access to disoccluded background information. Because of these inherent limits, most of the previous hole-filling schemes resort to an intra-image-based approach such as pixel-by-pixel image interpolation and extrapolation, or patch-by-patch image inpainting, etc. [4], [5]. These concealment style approaches, however, relied only on the intra-image candidates without exploiting depth information, leaving room to improve the quality of converted 3D video.

As for the exploitation of depth information for hole-filling, recently, a few of researches have been reported. In [6], Daribo et al. proposed a depth-guided image inpainting, where depth information is used as regularization terms for patch priority and patch-matching error calculations. This scheme was further enhanced by Gautier et al. using 3D tensor based on depth and color combination such that the object structure near hole-area can be diffused along depth as well as color information [7]. By considering depth structure and combining depth with color for hole-filling, these schemes exhibit a significant quality improvement over the previous intra-image-based approaches. However, such combination of depth and color yields, sometimes, the mixture of wrong depth layers (i.e. when color activity surpasses the effect of depth on inpainting). This mixture triggers the propagation of wrong depths via repeated patch concealment, resulting in a limited performance, especially, for high resolution and/or active depth videos producing large holes.

In order to solve this problem, this letter proposes a novel depth-guided inpainting scheme, where foreground (object) layers near hole area are further detected and removed from inpainting process. This foreground layer removal helps to fill the patch with proper background information, and thus to avoid the propagation of wrong depths. Moreover, by defining the priority of a patch based only on depth information and in more controllable form, the proposed scheme efficiently trades off the confidence term of a patch against its data term, leading to a more adaptive algorithm.

2. Proposed Algorithm

In order to explain the proposed algorithm, let us assume that we have a reference viewpoint image of \( \Omega = (\Omega_R, \Omega_c, \Omega_B, \Omega_G) \), where the image contains 5 color compo-
nents $\Omega_x, \Omega_y, \Omega_z$ and a depth component $\Omega_z$. After DIBR, this reference image is converted to a new viewpoint image $\Omega^D = (\Omega^D_x, \Omega^D_y, \Omega^D_z, \Omega^D_0)$, all the components of which have the same pitted area caused by horizontal disparities. Then, if we denote the pixels of this pitted area (holes) and the pixels of its boundary by $H^D$ and $B^D$, respectively, the proposed hole filling scheme can be described as follows.

2.1 Foreground Depth-Layer Detection

In conventional image inpainting schemes, a block of pixels consisting of known as well as unknown (hole) areas is concealed recursively by finding a number of well-matched blocks. Here, the matching is performed using the known area pixels and the concealment is performed using the unknown area pixels of well-matched blocks. However, when we consider the rationale behind the horizontal disparity of viewpoint projection, foreground (object) pixels having relatively high depth values shall not be used as the known area pixels for patch matching because foreground object produces disoccluded background which is discontinuous with it. One attempt to accommodate this is to define a set of partial boundary pixels $\Omega_B$ such that

$$\partial B^D = \{p \mid p \in B^D \text{ and } p + u \in H^D\}, \quad (1)$$

where $u$ is the horizontal unit vector $(-1, 0)$ under the assumption that the new viewpoint image is for the right-eye view of stereo vision. This set of partial boundary pixels basically excludes such foreground objects from forming the known area pixels. However, some of the right-hand side boundary pixels may belong to another foreground object which should also be removed from inpainting process.

In order to figure out this situation in more detail, let us consider an example shown in Fig. 1, where Fig. 1 (a) shows a reference viewpoint image and Fig. 1 (b) depicts the result of DIBR with hole area across the depth layer 1 and others. If we assume that we are dealing with the block of pixels centered at $p$ (the black rectangle in Fig. 1 (c)), we can easily identify that the depth layer 2 shall be excluded from the known area because it is another foreground object which occlude the background of depth layer 3. Thus, as shown in Fig. 1 (c), this foreground depth layer shall be detected and removed from the known area pixels of the block to produce the desired inpainting result shown in Fig. 1 (d).

In order to detect such foreground layer, we focus on the differences of neighboring depths in reference and projected viewpoint images (see Fig. 2). To be more specific, let us assume that a patch $\Psi_p \subset \Omega_z$ centered at $p \in \partial B^D$ is defined by

$$\Psi_p = \{q \mid |p_x - q_x| \leq W_x \text{ and } |p_y - q_y| \leq W_y\}, \quad (2)$$

where $W_x$ and $W_y$ represent the horizontal and the vertical size of the patch. And then, let $q^N$ denote the neighboring pixel of $q$, which is across the hole area $H^D$, (i.e., $q^N \in B^D - \partial B^D$ and $q^N = q_0$). Here $q = (q_x, q_y)$ is an element of $\Psi_p \cap \partial B^D$ and $q^N = (q^N_x, q^N_y)$. Further, let us denote by $q^{NC} \in \Omega_z$, the corresponding pixel of $q^N \in \Omega^D_z$ (i.e. inverse DIBR transforms $q^N$ to $q^{NC}$). Then, as shown in Fig. 2, the difference between $q$ and $q^N$ would be quite different from the disparity between $q^{NC}$ and $q^{NC} + (1, 0)$ if $q$ is on a foreground object such as the depth layer 2 of our example. Hence, for each pixel $q$, we define two differences $d^B_p$ and $d^{NC}$ such as

$$d^B_p = |q - q^N| \text{ and } d^{NC} = |q^{NC} - (q^{NC} + (1, 0))| \quad (3)$$

Then, based on the threshold $K_{pb}$, the partial boundary set $\partial B^{DR}_p$ is defined by

$$\partial B^{DR}_p = \{q \mid q \in \Psi_p \cap \partial B^D \text{ and } |d^B_q - d^{NC} | \leq K_{pb}\} \quad (4)$$

Using this set, we further define the local background depth layers $\Psi^B_p$ based on the threshold $K_{bd}$. That is, for each $k \in \partial B^{DR}_p$,

$$\Psi^B_p = \{q \mid q \in \Psi_p \text{ and } |\Omega^D_0(k) - \Omega^D_z(q)| \leq K_{bd}\} \quad (5)$$

Here, one exception is the case where $\partial B^{DR}_p = \emptyset$. In this case, $\Psi^B_p$ is also defined by an empty set such that this type
of patch will never be selected until the neighboring hole area of the foreground object is filled by some background layer data.

### 2.2 Patch Priority

After foreground depth layers are detected and removed from the patch $\Psi_p$, the priority $P(p)$ is calculated for the patch $\Psi_p^B$ by

$$P(p) = S(p) \times \left( \frac{Z_{\text{max}} + 1 - \Omega_D^p(p)}{Z_{\text{max}} + 1} \right),$$  

where $Z_{\text{max}}$ is the maximum depth value of $\Omega_D^p$ and the sparsity $S(p)$ is defined by

$$S(p) = C(p) + \lambda \times D(p)$$  

In (7), $C(p)$ and $D(p)$ are the confidence and the data terms of the patch $\Psi_p^B$, which are defined, respectively, by

$$C(p) = \frac{N(\Psi_p^B)}{N(\Psi_p)},$$

$$D(p) = \frac{1}{N(\Psi_p)} \sum_{(x,y) \in \Psi_p^B} \left( \frac{\Omega_D^p(x,y) - \Omega_D^p(x,y-1)}{\|p - (x,y)\|_1} \right)^2$$

where $N(X)$ means the cardinality of the set $X$, and $\|x\|_1$ denotes the L1 norm of $x$. Note that the data term in (9) is defined by using only the depth value $\Omega_D^p$ and the sparsity in (7) by the weighted sum of $C(p)$ and $D(p)$, instead of the color and depth combination (i.e. $\Omega_D^p$) and the multiplication of $C(p)$ and $D(p)$ as in the previous works [6], [7]. With these definitions, the relative importance of the horizontal edge strength in $D(p)$ can be more adaptively controlled by appropriately choosing the parameter $\lambda$.

### 2.3 Patch Selection and Hole Filling

After completing the assignment of $P(p)$ to each $p \in \partial B^D$, the highest priority patch $\Psi_p^B$ is selected.

$$p^* = \arg \max_{p \in \partial B^D} P(p)$$

Then we find the most similar nearby candidate patches $\Psi_q^B$ by calculating

$$q^* = \arg \min_{q \in L_{p^*}} D(\Psi_p^B, \Psi_q^B),$$

where $L_{p^*}$ is the local search range centered at $p^*$ and the distortion measure $d(\cdot, \cdot)$ is given by

$$d(\Psi_p^B, \Psi_q^B) = \sum_{s \in \Psi_p^B, X \in \{R,G,B,Z\}} \left( \Omega_X^p(s) - \Omega_X^p(s + q) \right)^2$$

for

Note that foreground depth layers were already removed from the highest priority patch, and thus they are not involved in the above patch matching.

In order to get more robust hole filling results, we used 5 most similar candidate patches, combined using the weights depending on their exponential distortions as in the previous work of [7]. With this method, the unknown hole area (i.e. $\Psi_{p^*} \cap H^p$) is inpainted by the corresponding color and depth pixels of the combined most similar patch blocks. Finally, the local search range $L_{p^*}$ in (11) was set as the 20 times larger size of a patch (i.e. using 20 · $W_x$ and 20 · $W_y$ for the definition of a patch in (2)) and these above patch-based hole filling procedures are iterated until no more hole area pixels remain.

### 3. Experimental Results

To verify the performance of the proposed algorithm, we selected 15 high resolution stereoscopic images from the datasets of Middlebury website [8]. For each test image, we generated the right-eye view image and its associated depth map and then filled the produced hole area using the proposed scheme. As a comparative study, we also implemented two recent depth-aided hole filling algorithms, proposed by Daribo et al. in [6] and by Gautier et al. in [7], to perform the hole filling for the same hole area. In the implementation of the proposed scheme, we used $W_x = W_y = 6$, $K_{pb} = 5$, $K_{bd} = 10$, and $\lambda = 1.0$, and the test results are summarized in Table 1.

As shown in Table 1, the proposed algorithm outperforms the tested two previous works by 1.19dB and 0.49dB, respectively, in terms of average PSNR (Peak Signal to Noise Ratio). Here, the PSNRs were calculated only on the hole area of the generated color image (not on the depth map). Although the average PSNR of the proposed scheme is higher than those of the other methods, we can find in the table that not all the test images provide positive gains such as the case of the test image ‘baby1’. However, if we look into the bottom part of Fig. 3 and the mid row of Fig. 4, we can easily identify that the depth quality and the subjective

| Test image | Daribo et al. (dB) | Gautier et al. (dB) | Proposed (dB) |
|------------|------------------|------------------|--------------|
| bowling1 | 27.46 | 28.93 | 29.14 |
| baby1 | 27.58 | 28.22 | 27.96 |
| baby2 | 23.63 | 28.98 | 29.08 |
| baby3 | 26.44 | 26.01 | 27.24 |
| cloth2 | 25.97 | 26.52 | 26.89 |
| flowerpots | 22.29 | 23.20 | 23.57 |
| lampshade1 | 27.18 | 27.61 | 29.72 |
| midd2 | 23.62 | 23.41 | 24.30 |
| monopoli | 25.27 | 25.78 | 26.39 |
| rocks1 | 24.60 | 26.63 | 26.73 |
| wood1 | 29.08 | 29.08 | 28.72 |
| wood2 | 30.30 | 31.30 | 32.21 |
| dolls | 25.37 | 25.23 | 25.28 |
| books | 25.60 | 22.14 | 23.18 |
| noobias | 27.61 | 27.46 | 27.44 |

Average 26.00 26.70 27.19
color image quality are much better in the proposed scheme than those of the previous works. The reason of such low objective quality (PSNR) in the test image ‘baby1’ can be explained by the mismatched dark yellow and the dark green parts in the reconstructed hole area (see the rightmost image of the mid row of Fig. 4). This type of subjective superiority has been identified for other lower objective quality cases, and is quite helpful in improving the quality of stereo image viewing experience.

Figure 3 depicts the reconstructed depth maps from different hole filling methods. The top row is the results for the test image ‘bowling1’ and the bottom row for ‘baby1’. If we compare the results shown in Fig. 3(b) with the results of the proposed algorithm (Fig. 3(c)), we can verify that the hole area was appropriately filled in the proposed scheme by using only background depth layers information. However, we can see the mixture of depth layers in the reconstructed images by the two previous works.

Finally, Fig. 4 shows the reconstructed color images by different hole filling methods. As was explained before, the hole filling results by the previous works exhibit some mixtures of foreground and background texture information, while the proposed scheme provides edge-preserving high quality reconstruction. Especially, the result from the test image ‘baby1’ shows the best subjective quality although its objective quality is worse than those of the other hole filling algorithms.

4. Conclusions

In this letter, we proposed a novel depth-guided inpainting scheme for high quality hole-filling of 2D-to-3D video conversion. The proposed algorithm detects foreground depth layers in each patch to remove them from patch priority and patch matching processes. By efficiently removing such foreground depth layers, the proposed scheme successfully inplants hole area using only appropriate background depth layers information, and thus prevents the propagation of wrong depths into hole area. Experimental results demonstrated better objective as well as subjective inpainting quality, which is quite helpful in improving the overall quality of converted 3D video experience.

Acknowledgements

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the “ITRC (Information Technology Research Center)” support program supervised by the NIPA (National IT Industry Promotion Agency)” (NIPA-2013-H0301-13-1008)

References

[1] C. Fehn, “Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV,” Proc. SPIE, vol.5291, pp.93–104, 2004.
[2] K. Oh, S. Yea, and Y. Ho, “Hole-filling method using depth based inpainting for view synthesis in free viewpoint television (FTV) and 3D video,” Proc. 27th Conference on Picture Coding Symposium, 2009.
[3] L. Do, S. Zinger, and P. With, “Quality improving techniques for free-viewpoint DIBR,” Proc. 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009.
[4] C. Vazquez, W.J. Tam, and F. Smeranza, “Stereoscopic imaging: Filling disoccluded areas in depth image-based rendering,” Proc. SPIE, vol.6392, 63920D, 2006.
[5] L. Po, S. Zhang, X. Xu, and Y. Zhu, “A new multidirectional extrapolation hole-filling method for depth-image-based rendering,” Proc. 18th ICIP, pp.2589–2592, 2011.
[6] I. Daribo and B. Pesquet-Popescu, “Depth-aided image inpainting for novel view synthesis,” Proc. MMSP 2010.
[7] J. Gautier, O. Meur, and C. Guillemot, “Depth-based image completion for view synthesis,” Proc. 3DTV Conf. 2011.
[8] http://vision.middlebury.edu/stereo/data/