Disocclusion Hole-Filling in DIBR-Synthesized Images using Multi-Scale Template Matching

Smarti Reel[1], K. C. P. Wong[1], Gene Cheung[2], Laurence S. Dooley[1]

[1]*Department of Computing and Communications, The Open University, Milton Keynes, United Kingdom*
[2]*National Institute of Informatics, Tokyo, Japan*

Email: [1]{smarti.reel, k.c.p.wong, laurence.dooley}@open.ac.uk, [2]cheung@nii.ac.jp

***Abstract*—Transmitting texture and depth images of captured camera view(s) of a 3D scene enables a receiver to synthesize novel virtual viewpoint images via Depth-Image-Based Rendering (DIBR). However, a DIBR-synthesized image often contains disocclusion holes, which are spatial regions in the virtual view image that were occluded by foreground objects in the captured camera view(s). In this paper, we propose to complete these disocclusion holes by exploiting the self-similarity characteristic of natural images via nonlocal template-matching (TM). Specifically, we first define self-similarity as nonlocal recurrences of pixel patches within the same image across different scales–one characterization of self-similarity in a given image is the scale range in which these patch recurrences take place. Then, at encoder we segment an image into multiple depth layers using available per-pixel depth values, and characterize self-similarity in each layer with a scale range; scale ranges for all layers are transmitted as side information to the decoder. At decoder, disocclusion holes are completed via TM on a per-layer basis by searching for similar patches within the designated scale range. Experimental results show that our method improves the quality of rendered images over previous disocclusion hole-filling algorithms by up to *3.9dB* in PSNR.**

***Index Terms*—Free viewpoint video, depth-image-based rendering, image inpainting**

I. Introduction

By transmitting both texture maps (color images) and depth maps (per-pixel distance between objects in the 3D scene and the capturing camera) captured from one or more camera view(s), *free viewpoint video* [1] enables a user the ability to synthesize novel virtual view images via depth-image-based rendering (DIBR) [2]. In a nutshell, DIBR copies color pixels in the camera-captured view(s) to corresponding pixel locations in the virtual view image, given 3D geometric information provided by the depth map(s). A DIBR-synthesized image often contains *disocclusion holes*, however, which are spatial regions in the virtual view that were occluded by *foreground* (FG) objects in the camera-captured view(s) but became visible after the view-switch. Satisfactory filling of disocclusion holes is essential to the free viewpoint visual experience. This paper addresses the disocclusion hole-filling problem.

978-1-4799-6139-9/14/$31.00

Coincidentally, the computer vision community has studied a related *image inpainting* problem over the last decade [3]–[5]: completion of missing pixel regions in a natural image. Broadly speaking, there are two categories of approaches. In the first category are schemes that locally extrapolate signals based on partial differential equations [4], Fourier analysis [5] etc. While intuitive, these local schemes do not perform well when the missing pixel regions are large. In the second category are nonlocal schemes such as Criminisi's *template-matching* (TM) [3] algorithm, that fill in missing pixels in a target region by identifying similar pixel patterns in faraway known region, assuming the well recognized self-similarity characteristic commonly observed in natural images. Recently, TM has been adopted for disocclusion hole-filling in DIBR-synthesized images as well [6], [7]. There remain two problems. First, there may not always exist self-similar patches of the same scale in a given image for TM to properly fill in missing pixels in disocclusion holes. Second, nonlocal TM schemes tend to be computationally complex due to the exhaustive search employed in the large known pixel region.

In this paper we propose a new sender-guided disocclusion hole-filling scheme that addresses the two aforementioned problems in previous nonlocal TM schemes. Specifically, we first redefine self-similarity in a *multi-scale* manner for natural images—a characterization of self-similarity for a given natural image is then how well target pixel patches will match with nonlocal patches of the same image resized by a specified range of scaling factors. Next, we design a sender-guided disocclusion hole-filling algorithm, where *i*) at encoder we divide a camera-captured texture image into multiple depth layers, characterize self-similarity in each layer and transmit the characterization parameters to decoder as side information (SI); *ii*) at decoder we perform TM for disocclusion hole-filling only within suitable depth layers but across multiple scales as specified by the transmitted self-similarity parameters. Performing TM within a subset of layers means search complexity for matching patches in known region is drastically reduced. Experimental results show that our disocclusion hole-filling algorithm outperforms previous schemes by up to *3.9dB* in PSNR at comparable or smaller computation complexity.

The outline of the paper is as follows. In Section II we for-

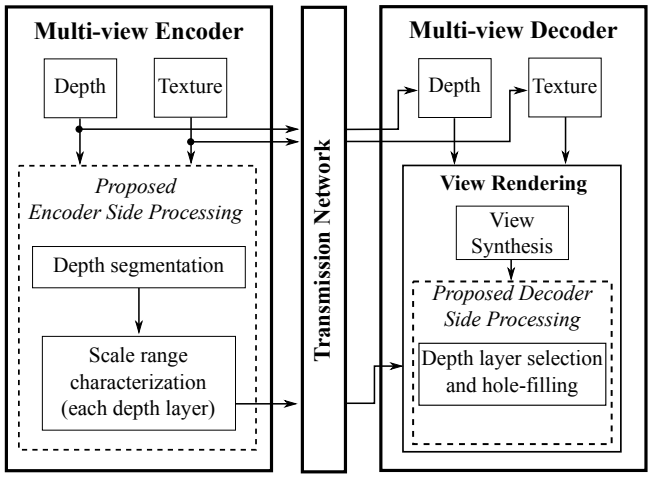


Fig. 1. Proposed processing blocks in multi-view context

mally define our proposed notion of self-similarity in a multi-scale manner. Using this definition, we describe our sender-guided disocclusion hole-filling algorithm based on multi-scale self-similarity in Section III. We present experimental results and conclusion in Section IV and V, respectively.

II. Characterization of Self-Similarity

It is observed that natural images are self-similar in general; *i.e.*, a given pixel patch is likely to recur one or more times in faraway (nonlocal) spatial regions in the same image. While existing works on inpainting using TM [3] assume the recurrence take place in the same scale, in this paper we generalize the notion to assume that the recurrence of a pixel patch can take place across multiple scales. More precisely, we characterize self-similarity in natural images as the *scale range* (SR) (parameterized by upper and lower bounds) over which a given pixel patch is likely to recur within the same image. This *multi-scale self-similarity* is an intuitive generalization; for example, repeating textural patterns like wallpaper vary in size as the distance to the capturing camera changes.

In practice, we compute the upper and lower bound that characterize multi-scale self-similarity as follows. A reference texture patch of size $w \times w$ pixels is first selected in a color image. Then each sliding window of size $(w+n) \times (w+n)$ pixels is rescaled to $w \times w$ pixels where n denotes the scaling factor within a given candidate range. Using *mean squared error* (MSE) as the distortion metric, for each n we identify the number of best-matched patches at this scaling factor and compare against a threshold T. The range of n values for which the number of best matched patches is higher than T defines the upper and lower bounds that characterize multi-scale self-similarity in this image.

III. Sender-Guided Hole-Filling Algorithm

Having defined our notion of multi-scale self-similarity in natural images, we now describe our disocclusion hole-filling algorithm that exploits this multi-scale self-similarity via TM. We first describe operations at the encoder to characterize self-similarity of camera-captured color images;characterization parameters are then transmitted to the decoder as SI. We then describe the operations at the decoder to efficiently perform TM guided by the received SI.

A. Encoder Side Processing

At the encoder (see Fig.1), the objective is twofold: *i*) segment the camera-captured texture image into depth layers−contiguous spatial areas with similar depth values, and *ii*) define and transmit SR for each depth layer to the decoder for sender-guided disocclusion hole-filling. We discuss them in order next.

1) Depth Layer Segmentation: The goal of depth layer segmentation is to divide a camera-captured texture image into contiguous spatial areas that roughly correspond to physical objects in the 3D scene. This is done so that multi-scale TM performed at the decoder can be done per layer instead of per image, reducing complexity. This is reasonable, since repeated textural patterns likely recur within the same physical object, contained in a depth layer. Let I_T and I_D be the texture and depth maps of a camera-captured view, respectively. We first divide depth map I_D into k layers by detecting peaks and valleys in a constructed histogram of depth values (see Fig 2(a) for an example) [8]. The depth cut-off values (based on valleys) $D = \{d_i\}$ correspond to the segmented depth layers $Y = \{y_i\}$ where $i = 1, 2, ..., k$ and $d_k = 255$. The same segmentation is then applied to the corresponding color image I_T as well.

2) Scale Range Characterization: We now characterize multi-scale self-similarity for each computed texture (color) layer. For target in TM, we consider only patches near the boundary of given layers. The reason is that disocclusion holes tend to appear near FG object boundaries [2]. In our experiment, we consider scale value n within range $[-3, 3]$. As illustration, Fig 2(b) shows the number of best matches for various scale values within the range $[-3, 3]$ for a depth layer in Middlebury's `Aloe` image. Here, it is observed that the number of best matches exceeds T[1] at $n = 0$ and $n = 1$, hence the SR= $[0, 1]$. The chosen SR for each layer is transmitted as SI to the decoder for sender-guided disocclusion hole-filling. Note that the SI transmission accounts for only a very small signaling overhead (0.01%) compared to the size of the camera-captured texture and depth maps.

B. Decoder Side Processing

The decoder receives a pair of texture and depth maps I_T, I_D and SR per depth layer, as shown in Fig. 1. For disocclusion hole-filling, a recent *Joint Texture and Depth Inpainting* (JTDI) algorithm [7] fills texture and depth hole pixels alternately: use available depth information to fill in textural pixel holes, then use inpainted textural information to fill corresponding depth pixel holes. However, JTDI still employs full-image TM, which is computation-expensive. In contrast, we perform joint texture and depth pixel filling of disocclusion holes as done in [7], but employ multi-scale

[1]The value of T is chosen empirically as 75. The experimental results are not particularly sensitive to this chosen value.

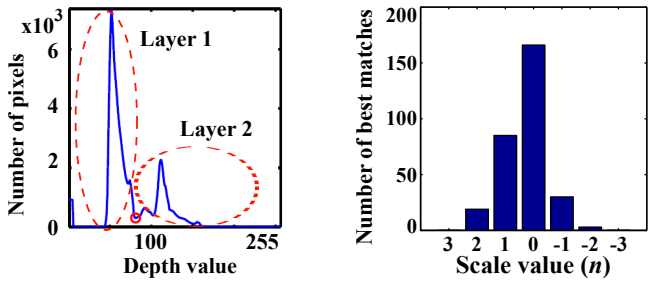


Fig. 2. (a) illustrates depth map histogram, $k = 2$ (b) shows bar graph representing number of best patches per scale value n for `Aloe` image.

TM within suitable depth layers. Here, SR provides side information on the suitable scaling values to resize candidate patches for each depth layer during TM.

The virtual texture view (V_T) and depth map (V_D) are synthesized from I_T and I_D via DIBR. Before filling disocclusion holes, both V_D and V_T are segmented with the same cut-off values D as discussed earlier. The following sub-section explains the depth layer selection method in our proposed disocclusion hole-filling algorithm.

1) Depth Layer Selection and Hole-filling: First, we select the target patch P_T and the corresponding depth target patch D_T. The order of selecting target patches for filling is very important, but is outside the scope of this paper; we will simply use the P_T selection method in [7]. The known values of D_T are used to determine the mean of depth values d_{mean}. Since disocclusion holes are missing pixels from *background* (BG) region [9], d_{mean} facilitates the selection of appropriate BG depth layer(s) Y_b as follows:

$$Y_b = \{y_i \in \boldsymbol{Y} \mid d_{mean} \leq d_i, d_i \in \boldsymbol{D}\} \ \ where\ i = 1, 2, ..., k. \quad (1)$$

The SR corresponding to depth layer(s) Y_b helps in generating multi-scale candidate search space $\boldsymbol{X}$ by rescaling the patches for given values in the range such that $\boldsymbol{X} = \{x_1, x_2, ..., x_q\}$ where q represents number of patches in $\boldsymbol{X}$. This search space is used for finding best candidate patch C_P as follows:

$$C_P = min\ MSE(x_j, P_T)\ where\ j = 1, 2, ..., q \quad (2)$$

The known pixels of selected C_P corresponding to the unknown (holes) pixels of P_T are then copied into P_T. This process repeats until all the disocclusion holes are filled.

IV. Experimental Setup and Results

The proposed algorithm was tested and evaluated using four Middlebury datasets [10]: `Aloe`, `Baby2`, `Books` and `Cones`. These datasets contain seven different captured views of the same static scene, as well as disparity maps for views #1 and #5. For each sequence, DIBR has been used to generate the reference view #3 using texture and the disparity map of view #1. To quantitatively and qualitatively evaluate the performance of proposed algorithm, the generated view #3 was inpainted using a patch-size of 9×9 pixels and compared with Criminisi [3] and JTDI [7] methods. Both the comparators, Criminisi [3] and JTDI [7] employs single-scale exhaustive TM to find C_P for filling disocclusion holes. For numerical analysis, the original view #3 of image datasets was used as the ground truth for all *peak signal-to-noise ratio* (PSNR) calculations, with the PSNR computed for both the whole image and hole regions. All experiments were performed on an Ubuntu 12.04 64-bit with 3.10 GHz Intel QuadCore and 4GB RAM, with all algorithms implemented in MATLAB.

A. Quantitative Result Analysis

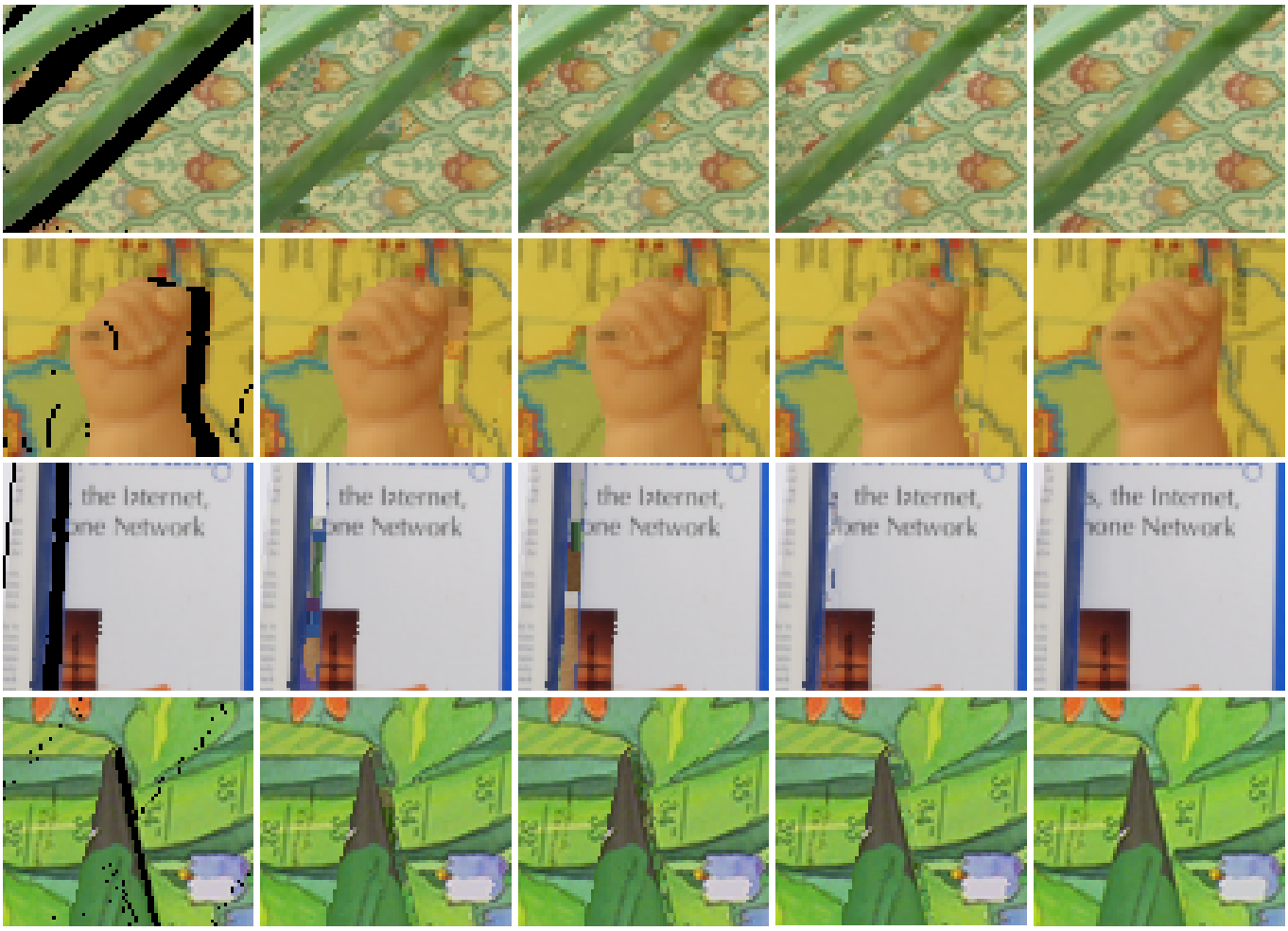
Table I shows the PSNR results for three inpainting methods, which reveal that proposed algorithm performs consistently better than [3] and [7] for all four datasets. In `Books` dataset for example, the PSNR increased by up to $2dB$ and $1.76dB$ for whole image as compared to Criminisi [3] and JTDI [7] respectively, while the corresponding results for only the hole regions, show an increase of $3.92dB$ and $3.44dB$. This confirms that characterization parameters provided in SR enhances the probability of finding a better match by searching C_P at two scales, i.e. at scale $n = 0$ and $n = -1$. The PSNR increase supports the fact that during the disocclusion filling process, there are cases where patches at scale $n = -1$ provide better matches than scale $n = 0$ (same scale). The computation cost due to multi-scale TM is compensated by layer based search during inpainting and the overall computation time remains either comparable or smaller than exhaustive TM. Similar observations can be made upon the results for `Aloe`, `Baby2` and `Cones` images using proposed algorithm.

B. Qualitative Result Analysis

From a perceptual quality perspective, Fig. 3 shows the qualitative comparison of proposed algorithm with [3] and [7] with example zoomed-in regions for `Aloe`, `Baby2`, `Books` and `Cones` in each row respectively. The proposed algorithm again provides improved visual quality and fewer artefacts by preserving the FG object boundaries as shown in Fig. 3(d) in comparison to Criminisi [3] (Fig. 3(b)) and JTDI [7] (Fig. 3(c)). For comparison of various techniques, the disocclusion holes and ground truth are shown in Fig. 3(a) and 3(e), respectively. Multi-scale TM reduces the artefacts and fills the disocclusion holes with enhanced perceptual quality.

TABLE I
PSNR Comparison For Texture Image Inpainting *(in dB)*

Dataset	Whole Image				Hole Regions		
	Criminisi [3]	JTDI [7]	Proposed n		Criminisi [3]	JTDI [7]	Proposed
`Books`	26.80	27.05	-1,0	**28.81**	15.14	15.62	**19.06**
`Baby2`	30.87	30.92	1, 0, -1	**31.80**	19.24	19.36	**21.75**
`Aloe`	26.83	27.83	1, 0	**28.14**	17.02	18.66	**19.50**
`Cones`	22.48	23.17	-1, 1, 2	**23.41**	17.45	19.92	**20.34**



(a) Disocclusion holes (b) Criminisi [3] (c) JTDI [7] (d) Proposed (e) Ground truth

Fig. 3. shows `Aloe` (row 1), `Baby2` (row 2), `Books` (row 3) and `Cones` (row 4) with corresponding (a) disocclusion holes, filled by (b) Criminisi [3], (c) JTDI [7], (d) Proposed method and respective (e) Ground truth.

V. Conclusion

While free viewpoint video enables the synthesis of novel virtual view images at decoder via DIBR, the synthesized images often contain disocclusion holes that require proper filling. In this paper, we propose a new disocclusion hole-filling algorithm that exploits multi-scale self-similarity during TM. Though searching for nonlocal patches of different scales entails a larger search space, we contain the resulting search complexity by performing TM only within designated depth layers−subset of the image with similar depth values. Experimental results show that our proposed hole-filling algorithm can outperform previous proposals by up to $3.9dB$ at comparable or smaller complexity.

References

[1] K. Takeuchi, N. Fukushima, T. Yendo, M. Panahpour Tehrani, T. Fujii, and M. Tanimoto, "Free-viewpoint image generation from a video captured by a handheld camera," vol. 7863, 2011, pp. 78 631N–78 631N–9.

[2] D. Tian, P.-L. Lai, P. Lopez, and C. Gomila, "View synthesis techniques for 3d video," *Proc. SPIE*, vol. 7443, pp. 74 430T–74 430T–11, 2009.

[3] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *Image Processing, IEEE Transactions on*, vol. 13, no. 9, pp. 1200–1212, Sept 2004.

[4] D. Tschumperle and R. Deriche, "Vector-valued image regularization with pdes: a common framework for different applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 4, pp. 506–517, April 2005.

[5] J.-F. Cai, R. Chan, L. Shen, and Z. Shen, "Simultaneously inpainting in image and transformed domains," vol. 112, no. 4. Springer-Verlag, 2009, pp. 509–533.

[6] I. Daribo and B. Pesquet-Popescu, "Depth-aided image inpainting for novel view synthesis," in *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, Oct 2010, pp. 167–170.

[7] S. Reel, G. Cheung, P. Wong, and L. Dooley, "Joint texture-depth pixel inpainting of disocclusion holes in virtual view synthesis," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, Oct 2013, pp. 1–7.

[8] D. De Silva, W. Fernando, H. Kodikaraarachchi, S. Worrall, and A. Kondoz, "Adaptive sharpening of depth maps for 3dtv," *Electronics Letters*, vol. 46, no. 23, pp. 1546–1548, November 2010.

[9] I. Ahn and C. Kim, "Depth-based disocclusion filling for virtual view synthesis," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, July 2012, pp. 109–114.

[10] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, June 2007, pp. 1–8.