LOW-SALIENCY PRIOR FOR DISOCCLUSION HOLE FILLING IN DIBR-SYNTHESIZED IMAGES

Bruno Macchiavello [*], *Camilo Dorea* [*], *Edson M. Hung* [*], *Gene Cheung* [#] *and Ivan Bajic* [$]

[*] Universidade de Brasilia, Brazil, [#] National Institute of Informatics, Japan
[$] Simon Fraser University, Canada

ABSTRACT

Although images as viewed from intermediate virtual viewpoints can be synthesized using texture and depth maps from nearby camera views via depth-image-based rendering (DIBR), the rendered images contain disocclusion holes—spatial regions that were not visible in the reference views due to foreground object occlusion—that requires proper filling. In this paper, we introduce a new signal prior into the hole filling problem formulation: given disocclusion holes are part of the background and background tends to have low visual saliency, the extrapolated signal into the holes must also be of low saliency. Mathematically, we add a low-saliency prior to an exemplar-based inpainting algorithm, so that the best-matched block has both small matching cost and is of low visual saliency. Moreover, we compute a suitable Lagrange multiplier value for the saliency cost term via analysis of the reference images. Experimental results show that using a low-saliency prior can improve performance by 0.5 dB over a previous hole filling scheme.

***Index Terms*—** Depth-image-based rendering, inpainting, visual saliency

1. INTRODUCTION

Towards the goal of *free viewpoint navigation* [1]—the ability for a receiver to freely choose any view from which to observe a dynamic 3D scene—it is now common in the literature to represent visual data of the 3D scene in *texture-plus-depth* format [2]. In a nutshell, it means texture maps (color images) and depth maps (per-pixel distance between objects in the 3D scene and the capturing camera) from multiple closely spaced viewpoints are captured and encoded at sender, so that synthesis of images at intermediate virtual views can be executed via *depth-image-based rendering* (DIBR) [3] at receiver. While DIBR is attractive for its low computation cost—it is essentially a pixel-to-pixel color mapping from reference view(s) to target view dictated by corresponding depth pixels—there exist *disocclusion holes* in the rendered images that can cause visual discomfort. Disocclusion holes are spatial areas that were not visible in the reference view(s) due to occlusion by foreground object(s), but became visible after the view-shift. We address the problem of how to fill disocclusion holes in a visually satisfactory manner in this paper.

Previous attempts [4, 5, 6, 7] at disocclusion hole filling leverage on inpainting techniques developed in the computer vision community such as Criminisi's exemplar-based matching algorithm [8] (called `CR` in the sequel). Specifically, a common observation is that disocclusion holes tend to be part of the background, so direction of signal propagation into a disocclusion hole should emanate from the background side. While this observation enables these proposals to often produce more reasonable fillings than original `CR`, the occasional errors they produce can be visually disturbing.

In this paper, we introduce one more significant insight into the problem of disocclusion hole filling: *background scenery tends to draw less visual attention than foreground objects in typical 3D scene*. That means that during hole filling, we know *a priori* that the extrapolated signal tends to have low visual saliency [9]. Mathematically, we express this knowledge as a *low-saliency prior*, so that during exemplar-based block matching, we can include it as an additional term in addition to the matching criteria. Moreover, we compute a suitable Lagrange multiplier value for the saliency cost term via a simple analysis of the reference images. Experimental results show that using the low-saliency prior we can outperform a previous hole filling scheme by up to 0.5 dB in PSNR in disoccluded regions.

The outline of the paper is as follows. We first review related work in Section 2. We then overview our chosen visual saliency model and how it was used in our previous works in Section 3. We discuss how the low-saliency prior is applied to disocclusion hole filling in Section 4. Finally, we present experimental results and conclusion in Section 5 and 6, respectively.

2. RELATED WORK

Inpainting of missing pixel patches in an image has been studied in computer vision for well over a decade, with approaches including partial differential equations (PDE) [10], exemplar-based matching [8] and sparse representation [11].

This work was partly supported by CNPq grants 476176/2013-1 and 310375/2011-8 and NSERC grant RGPIN 327249.

Exemplar-based matching, in particular, has gained popularity due to its conceptual and implementation simplicity; for example, [12, 13] pursued extensions where linear combination of multiple similar patches are sought instead of just the single best-matched patch. We also follow the exemplar-based matching paradigm in our work.

Previous works on disocclusion hole filling in DIBR-synthesized images can be broadly divided into two categories: i) signal extrapolation based on spatial correlation [4, 5, 6, 7], and ii) extrapolation based on temporal correlation [14, 15]. In principle, our proposed low-saliency prior can also be used for disocclusion hole filling of DIBR-synthesized video, where the saliency computation will include in addition low-level temporal features such as flicker and motion [16]. However, we focus on applying the low-saliency prior for inpainting of DIBR-synthesized images in this paper, and leave the video extension as future work.

In our previous work, we applied the low-saliency prior to error concealment in loss-corrupted streaming video [17] and view synthesis of loss-corrupted free viewpoint video [18]. To the best of our knowledge, this is the first work that incorporate the low-saliency prior to the disocclusion hole filling problem in DIBR-synthesized images.

3. LOW-SALIENCY PRIOR

Visual saliency refers to the propensity of visual stimuli to draw attention to themselves. Contrast in various low-level features such as intensity, color, orientation and motion, mediated by the center-surround mechanism, is known to attract attention [9]. Bayesian surprise, measured as the difference between prior and posterior distribution of a certain feature following an observation, has also been linked to saliency [19]. As computational models for saliency become more accurate, they become new tools to improve various visual signal processing tasks.

In particular, the low-saliency prior has been found to be useful in video error concealment [17], where it was used to promote blocks with low saliency relative to the neighborhood during the concealment process. The benefit was twofold. First, if high-saliency Region-Of-Interest (ROI) is protected more than the remainder of the frame, which is a reasonable design approach, the low-saliency prior is the correct side information (SI) and focuses the search to a smaller feasible region around the correct solution. Second, the low-saliency requirement leads to concealment blocks that are less attention grabbing, so that resulting errors are less noticeable.

In the present application of disocclusion hole filling, while there is no guarantee that the newly-revealed background will always be of low saliency relative to its immediate neighborhood, we have observed empirically that this is in fact the case in most frames of free viewpoint test sequences. See Fig. 1 for an example of saliency maps where clearly the foreground objects attract visual attention much more so than the background. Hence in this application too, the low-saliency prior provides the correct SI during exemplar-based patch matching in the typical case.

Fig. 1. Example of a color image (left) and its corresponding saliency map using the model in [20].

4. DISOCCLUSION HOLE-FILLING ALGORITHM

4.1. System Overview

We first overview a generic free viewpoint video streaming system in which our disocclusion hole filling algorithm is applicable. We assume that at sender, multiple cameras synchronously capture a dynamic 3D scene from different viewpoints in texture (color) and depth maps of the same spatial resolution. For bandwidth efficiency, texture and depth map pairs from at most two camera viewpoints nearest to the requested virtual view are compressed at sender for transmission to receiver. The received texture and depth maps are used for virtual view synthesis via DIBR [3]. For better RD performance, recent proposals [21, 22] call for transmission of texture and depth map pair from a *single* viewpoint for view synthesis at receiver. This results in larger disocclusion holes in general, and our proposed hole filling method becomes more important.

DIBR [3] is a pixel-based image synthesis procedure, where each color pixel in the texture map of a camera-captured view (*reference view*) is copied to a pixel location in the virtual view image; the copied location is determined by camera parameters and corresponding depth pixel value. If two pixels from the same reference view are mapped to the same location, then the pixel with the smaller depth value is kept. If two pixels from two different reference views are mapped to the same location, then a linear combination of the two pixel values (*pixel blending*) is computed. Disocclusion hole is a location in the virtual view image where no color pixels are mapped from the reference texture map(s), due to occlusion by foreground objects in the reference view(s). We focus on the filling of disocclusion holes next.

4.2. Exemplar-based Matching

We now overview the exemplar-based patch matching strategy proposed in [8]. Let the *source region* (known pixel region) be $\Phi = I - \Omega$, where I and Ω are input image and *target*

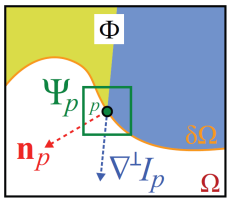

Fig. 2. Illustration of Criminisi's algorithm [8].

region (disocclusion holes), respectively. Let the boundary between the source and hole region be $\delta\Omega$. See Fig. 2 for an illustration.

Let a $N \times N$ patch with center at pixel p be denoted by Ψ_p. [8] proposed to always select a *target patch* Ψ_p with center pixel p on the boundary, *i.e.* $p \in \delta\Omega$, for exemplar-based matching. Mathematically, the matching is written as:

$$\min_{\Psi_q \in \Phi} d(\Psi_p, \Psi_q) \tag{1}$$

In other words, the most similar patch Ψ_q to target Ψ_p in source region Φ, in terms of the difference between known pixels in Ψ_p and corresponding pixels in Ψ_q, is sought. The idea is that images tend to be *self-similar*, so the target patch with missing pixels will likely reappear in the source region.

At any given time in the inpainting process, there can be many potential target patches Ψ_p, $p \in \delta\Omega$. [8] stressed that the order in which the patches are selected as target is important; the order proposed was according to a *priority term* $P(p)$:

$$P(p) = C(p)\, D(p) \tag{2}$$

where $C(p)$ and $D(p)$ are the *confidence* and *data* terms, defined as:

$$C(p) = \frac{\sum_{q \in \Phi_p \cap \Phi} C(q)}{|\Psi_p|}, \qquad D(p) = \frac{|\nabla I_p^\perp \cdot n_p|}{\alpha} \tag{3}$$

where $|\Psi_p|$ counts the number of known pixels in Ψ_p, n_p is the unit vector orthogonal to $\delta\Omega$ at p, $\nabla I_p^\perp$ is the isophote (direction and intensity) at p, and α is a normalization factor. $C(p)$ gives higher priority to patches with more known pixels. $D(p)$ encourages propagation of linear structures. See [8] for details.

4.3. Saliency-cognizant Exemplar-based Matching

We are now ready to discuss how we introduce a low-saliency prior into exemplar-based patch matching. Essentially, we restrict candidate matching patches Ψ_q to ones with saliency values $S(\Psi_q)$ less than a threshold value, *i.e.*:

$$\min_{\Psi_q \in \Phi} d(\Psi_p, \Psi_q) \quad \text{s.t.} \quad S(\Psi_q) \le \bar{S} \tag{4}$$

Saliency is a relative term, and $S(\Psi_q)$ is computed relative to the known pixels in a local neighborhood center at q for computation efficiency. $\bar{S}$ can be computed, for example, via observed saliency values of background regions in reference frames.

As traditionally done in the literature [23], instead of solving the original constrained optimization (4), we solve instead the corresponding unconstrained Lagrangian problem with multiplier λ:

$$\min_{\Psi_q \in \Phi} d(\Psi_p, \Psi_q) + \lambda S(\Psi_q) \tag{5}$$

We discuss selection of an appropriate λ in the next section.

Besides the actual patch search for given target Ψ_p in (5), we also optimize the selection of suitable target patch Ψ_p given target region Ω. In particular, DIBR-synthesized image contains (partial) per-pixel depth information that we can exploit for target path selection. Specifically, we use available depth values in the target patch Ψ_p to compute an average depth $\bar{Z}_p$ and inverse depth variance $L(p)$ for inclusion into the priority computation [24]:

$$p(p) = (C(p) + D(p) + L(p)) \times f(\bar{Z}_p) \tag{6}$$

where $f(Z)$ is a monotonically increasing function of input Z. The main idea is that patches with largest average depth will be selected first, and among those with the similar average depth, ones with smallest depth variance will be selected first. This ensures background information will be propagated to the disocclusion holes, as described in the Introduction.

4.4. Selection of Lagrange Multiplier

In a typical Lagrangian minimization, the optimal selection of the appropriate multiplier value λ is a difficult task [23]. In our specific case of low-saliency prior, however, we can compute an appropriate λ as follows. Using a reference texture map, we first identify portions of boundary background regions horizontally next to foreground objects—regions likely similar to disoccluded region in the virtual views. We then perform patch search using (1), where in this case the target region Ψ_p has no unknown pixels. Suppose the best-matched patch is Ψ_{q^*}, *i.e.*,

$$\Psi_{q^*} = \arg\min_{\Psi_q \in \Phi} d(\Psi_p, \Psi_q) \tag{7}$$

which represents the best solution using the exemplar-based framework under ideal condition. The first constraint for λ is to ensure that even with the low-saliency prior, q^* can still be selected, *i.e.*, $\forall \Psi_q \in \Phi \mid S(\Psi_q) < S(\Psi_{q^*})$,

$$\begin{aligned} d(\Psi_p, \Psi_{q^*}) + \lambda S(\Psi_{q^*}) &\le d(\Psi_p, \Psi_q) + \lambda S(\Psi_q) \\ \lambda &\le \frac{d(\Psi_p, \Psi_q) - d(\Psi_p, \Psi_{q^*})}{S(\Psi_{q^*}) - S(\Psi_q)} \end{aligned} \tag{8}$$

The second constraint on λ is that it has to be large enough to make a difference. In other words, when only *half* of the

pixels are used for distortion computation (denoted as d'), Ψ_{q^*} is still the optimal solution. Mathematically we write: $\forall \Psi_q \in \Phi \mid d'(\Psi_p, \Psi_q) \leq d'(\Psi_p, \Psi_{q^*})$,

$$\begin{aligned} d'(\Psi_p, \Psi_{q^*}) + \lambda S(\Psi_{q^*}) &\leq d'(\Psi_p, \Psi_q) + \lambda S(\Psi_q) \\ \lambda &\geq \frac{d'(\Psi_p, \Psi_{q^*}) - d'(\Psi_p, \Psi_q)}{S(\Psi_q) - S(\Psi_{q^*})} \end{aligned} \quad (9)$$

Performing the above calculation for a given patch Ψ_p yields a range R_p for λ. We repeat the calculation for all patches in the estimated region $\mathcal{E}$, and the λ selected is in the intersection of the largest set $\mathcal{S}$ of ranges R_p's without having the intersection as empty set. In other words:

$$\begin{aligned} \lambda &\in \bigcap_{p \in \mathcal{S}} R_p \\ \mathcal{S} &= \arg\max_{\mathcal{S} \subseteq \mathcal{E}} |\mathcal{S}| \quad s.t. \quad \bigcap_{p \in S} R_p \neq \emptyset \end{aligned} \quad (10)$$

The final chosen λ is the middle value of the intersection of ranges R_p's in $\mathcal{S}$.

5. EXPERIMENTATION

Experimental results are reported for four test sequences, *Ballet*, *Breakdancers* [25], *Akko & Kayo* [26] and *Poznan Street* [27] under various camera setups. We assume a virtual viewpoint is synthesized via DIBR using texture and depth map pair from a single camera viewpoint, resulting in disocclusion holes. For *Breakdancers*, we synthesized view 2 using view 1 as reference and synthesized view 2 using view 3, as specified in Table 1. For *Akko & Kayo*, we synthesized view 48 from 47, for *Ballet*, we synthesized view 2 from 3 and for *Poznan Street*, we synthesized view 5 from 3. In all cases the first frame of each sequence is used.

Saliency was computed according to the method in [20]. To improve performance of the regular exemplar-based hole filling algorithm [8], referred to as $EB1$ in our discussion, we restricted the search area for candidate patches to regions around holes which are opposite the direction of DIBR projection. For example, the search area is limited to a 36 pixel band to the left of the holes depicted for *Breakdancers 2-1* shown in Fig. 3(a). In this manner, the background regions are generally applied towards hole filling. In addition, we test an improved exemplar-based algorithm [28], referred to as $EB2$, using a level regularity term based on depth information. Note that depth information of the virtual view will also contain disocclusion holes. In $EB2$ as well as in our proposed scheme, these depth holes are filled jointly with the corresponding texture filling mechanism. In other words, after the best candidate is selected for texture, its corresponding depth patch is also used to fill the virtual depth map. For $EB1$ and $EB2$ the patch size was set to 11×11, while the proposed scheme uses variable patch sizes from 9×9 to 13×13. In this case, the best candidate is selected among the best-matched patch of each size using a size-normalized version of (5).

Table 1. Luma PSNR results in dB within disocclusion areas for hole filling algorithms.

Image	$EB1$	$EB2$	**Proposed**
Breakdancers 2-1	22.14	22.93	23.43
Breakdancers 2-3	22.42	22.61	22.75
Akko & Kayo 48-47	15.01	15.33	15.64
Ballet 2-3	22.10	22.29	22.30
Poznan Street 5-3	27.79	27.51	27.87

Table 1 presents PSNR for the hole filling algorithms computed for luminance components with respect to the original views (ground truth) within the disocclusion areas. For the tested images, our proposed scheme achieved higher PSNR, outperforming $EB2$ by up to 0.5 dB. As illustrated in Fig. 3, we note further that our proposed method fills background holes with lower saliency content, resulting in more visually pleasant images.

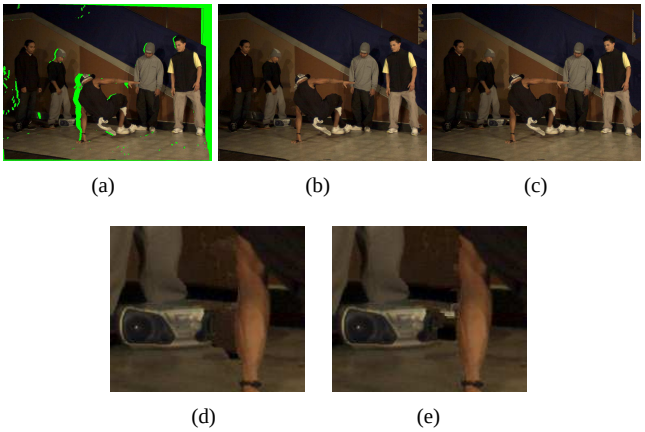

Fig. 3. (a) Disocclusion holes (in green) for *Breakdancers* view 2 synthesized from view 1 and hole filling results of (b) $EB2$ and (c) proposed method. Detail crops from (d) $EB2$ and (e) proposed.

6. CONCLUSION

We presented a method for filling of disocclusion holes—spatial regions that were occluded in the reference view(s) but became visible after a view-switch. Given background tends to draw less visual attention than foreground in typical 3D scene, the key idea is to include a *low-saliency prior* during exemplar-based patch matching, so that the selected patch in the source region has both small matching cost and low saliency value. Experimental results show that the addition of a low-saliency prior together with variable patch size outperformed a previous implementation of disocclusion hole filling method by 0.5 dB in PSNR.

7. REFERENCES

[1] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," in *IEEE Signal Processing Magazine*, January 2011, vol. 28, no.1.

[2] P. Merkle, A. Smolic, K. Mueller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *IEEE International Conference on Image Processing*, San Antonio, TX, October 2007.

[3] D. Tian, P.-L. Lai, P. Lopez, and C. Gomila, "View synthesis techniques for 3D video," in *Applications of Digital Image Processing XXXII, Proceedings of the SPIE*, 2009, vol. 7443 (2009), pp. 74430T–74430T–11.

[4] K.-J. Oh, S. Yea, and Y.-S. Ho, "Hole-filling method using depth based in-painting fore view synthesis in free viewpoint television (FTV) and 3D video," in *Picture Coding Symposium*, Chicago, IL, May 2009.

[5] I. Daribo and B. Pesquet-Popescu, "Depth-aided image inpainting for novel view synthesis," in *IEEE Multimedia Signal Processing Workshop*, Saint-Malo, France, October 2010.

[6] O. Le Meur, J. Gautier, and C. Guillemot, "Exemplar-based inpainting based on local geometry," in *IEEE International Conference on Image Processing*, Brussels, Belgium, September 2011.

[7] I. Ahn and C. Kim, "Depth-based disocclusion filling for virtual view synthesis," in *IEEE International Conference on Multimedia and Expo*, Melbourne, Australia, July 2012.

[8] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," in *IEEE Transactions on Image Processing*, September 2004, vol. 13, no.9, pp. 1–13.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, November 1998, vol. 20, no.11, pp. 1254–1259.

[10] D. Tschumperle, "Fast anisotropic smoothing of multi-valued images using curvature-preserving PDE's," in *International Journal on Computer Vision*, November 2006, vol. 68, no.1, pp. 65–82.

[11] Z. Xu and J. Sun, "Image inpainting by patch propagation using patch sparsity," in *IEEE Transactions on Image Processing*, May 2010, vol. 19, no.5, pp. 1153–1165.

[12] A. Wong and J. Orchard, "A nonlocal-means approach to exemplar-based inpainting," in *IEEE International Conference on Image Processing*, Atlanta, GA, October 2006.

[13] C. Guillemot, M. Turkan, O. Le Meur, and M. Ebdelli, "Image inpainting using LLE-LDNR and linear subspace mappings," in *IEEE International Conference on Acousitics, Speech and Signal Processing*, Vancouver, Canada, May 2013.

[14] K.-W. Hung and W.-C. Siu, "Depth-assisted nonlocal means hole filling for novel view synthesis," in *IEEE International Conference on Image Processing*, Orlando, FL, September 2012.

[15] W. Sun, O. Au, L. Xu, Y. Li, and W. Hu, "Novel temporal domain hole filling based on background modeling for view synthesis," in *IEEE International Conference on Image Processing*, Orlando, FL, September 2012.

[16] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," in *IEEE Transactions on Image Processing*, October 2004, vol. 13, no.10, pp. 1304–1318.

[17] H. Hadizadeh, I. Bajic, and G. Cheung, "Saliency-cognizant error concealment in loss-corrupted streaming video," in *IEEE International Conference on Multimedia and Expo*, Melbourne, Australia, July 2012.

[18] B. Macchiavello, C. Dorea, M. Hung, G. Cheung, and W. t. Tan, "Saliency-cognizant robust view synthesis in free viewpoint video streaming," in *IEEE International Conference on Image Processing*, Melbourne, Australia, September 2013.

[19] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, 2009.

[20] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Ssstrunk, "Frequency-tuned Salient Region Detection," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 2009, pp. 1597 – 1604.

[21] I. Daribo, G. Cheung, T. Maugey, and P. Frossard, "RD optimized auxiliary information for inpainting-based view synthesis," in *3DTV-Conference*, Zurich, Switzerland, October 2012.

[22] Y. Gao, G. Cheung, and J. Liang, "Rate-complexity tradeoff for client-side free viewpoint image rendering," in *IEEE International Conference on Image Processing*, Melbourne, Australia, September 2013.

[23] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, September 1988, vol. 36, no.9, pp. 1445–1453.

[24] S. Reel, G. Cheung, P. Wong, and L. Dooley, "Joint texture-depth pixel inpainting of disocclusion holes in virtual view synthesis," in *APSIPA ASC*, Kaohsiung, Taiwan, October 2013.

[25] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Transactions on Graphics (Proc. SIGGRAPH 2004)*, vol. 3, no. 23, pp. 600–608, 2004.

[26] "Nagoya university ftv test sequences," in *http://www.tanimoto.nuee.nagoya-u.ac.jp/*.

[27] M. Domanski, T. Grajek, K. Klimaszewski, M. Kurc, and O.Stankiewicz, "Poznan multiview video test sequences and camera parameters," in *ISO/IEC JTC1/SC29/WG11, M15386, 2009*.

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