ROBUST VIEW SYNTHESIS IN WIDE-BASELINE COMPLEX GEOMETRIC ENVIRONMENTS

Sheng Wang[1,2], *Ronggang Wang*[1,2*]

[1]School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
[2]Peng Cheng Laboratory
sheng.wang@pku.edu.cn, rgwang@pkusz.edu.cn

ABSTRACT

One of the most challenging problems of novel view synthesis is to predict the scene in complex geometric environments. Existing methods depend on either homography optimization or deep neural networks. In this paper, we provide a framework of view synthesis, which includes grid-based warp, depth refinement and ghost artifacts removal. The depth refinement method is our main contribution, which can be combined with any other warp operation to generate high quality images. To achieve it, the depth refinement method is combined with a shape-preserving warp operation together based on reliable, half-reliable and unreliable superpixel discrimination. We remove outliers in half-reliable superpixels by considering their neighboring reliable superpixels and distinguish half-reliable ones into reliable and unreliable parts. This step helps us to get more accurate depth information. Experimental results show that our view synthesis system has nearly 0.7dB gains in PSNR and 0.03 gains in SSIM compared with the state-of-the-art view synthesis algorithm.

Index Terms— View synthesis, depth refinement, image-domain-warping, view interpolation

1. INTRODUCTION

View synthesis is a complex system, which includes image signal transmission, 3D reconstruction, image inpainting and so on. As for different environments, different approaches are applied. Traditional view interpolation methods [1, 2] are designed for stereo-images captured by stereo-camera, which have small-baseline. These methods mainly depend on optical flow and are evaluated on Middlebury benchmark [3]. Baker et al. [3] have a detailed survey for them. For other environments, which have wide-baseline, there are two main approaches, depth-image-based-rendering (DIBR) and image-domain-warping (IDW). Both DIBR and IDW need depth information. However, DIBR [4] requires high accurate per-pixel depth, which is almost impossible in complex geometric environments. IDW [5, 6] is not sensitive to depth noise and can work well with sparse depth.

In most parts of environments, especially in texture-less regions and complex geometric environments, visual-based 3D reconstruction algorithms [7, 8] cannot get satisfactory depth information and some regions even have no depth information. Generating novel views in these environments is a challenging problem for both DIBR and IDW. Inaccurate depth leads to poor image quality at the virtual view position and causes ghosts, occlusions and holes.

Many approaches have been applied. Some approaches aim to refine homography and warp the image to the virtual view position, others aim to synthesize depth to achieve the goal. Nie et al. [9] propose an algorithm in wide-baseline environments. They compute the Approximate Nearest Neighbor Field (ANNF) between input images and optimize the homographies of superpixels as accurate and robust as possible. However, their algorithm cannot generate free-viewpoint walkthroughs, which is the biggest limitation. The work of [5] is the state-of-the-art view synthesis algorithm. It applies a depth-synthesis approach for poorly reconstruction regions with the help of PMVS [10, 11]. The approach of [5] can work well based on the assumption that the superpixels with enough homogeneous depth information are reliable. However, our experiment result shows that a region with dense depth information may also be disturbed and cannot be warped directly. This non-robust assumption is also the motivation of our work and we get a better result according to our depth refinement method.

Our depth refinement method is focusing on improving the view synthesis quality in complex geometric environments, where the estimated depth information is mixed with noises. Unlike other methods that require dense depth information, such as DIBR, our method only uses sparse depth points. We provide a framework of view synthesis that includes our depth refinement method in Figure 2. Firstly, We use COLMAP [7, 8] and SLIC [12], which are both widely used in IDW as preprocessing steps. Then, superpixels are divided into reliable, half-reliable and unreliable parts. A

Thanks to National Natural Science Foundation of China 61672063Shenzhen Research Projects of JCYJ20160506172227337 and GGFW2017041215130858, Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality ZDSY201703031405467.

three-step depth refinement operation is applied before warping the nearest three input images to the virtual view position. These warped images are blended and the ghosts are removed finally. The pipeline of our system can be seen in Figure 1, where the green block is our main contribution and Figure 2 has a detail flow chart about this step.

This paper is organized as follows. In Section 2, we describe our algorithm in detail and the following Section 3 presents the experimental results. Section 4 concludes the paper.

2. THE PROPOSED METHOD

Figure 1 shows the flow chart of our view synthesis system. Since 3D reconstruction and oversegmentation are usually regarded as preprocessing steps, we choose COLMAP [7,8] and SLIC [12] respectively, which are state-of-the-art algorithms. Then, most of the existing methods [5, 6, 13] project the point cloud into image coordinate to get sparse depth information and warp the superpixels to the virtual view position directly. It's not robust since even a region with dense depth information may be disturbed in wide-baseline complex geometric environments. Thus, we add a three-step depth refinement method into the traditional framework, like the green block in Figure 1. Our depth refinement method is described in section 2.1, the warp method is described in section 2.2 and the ghost removal method is mentioned in section 2.3.

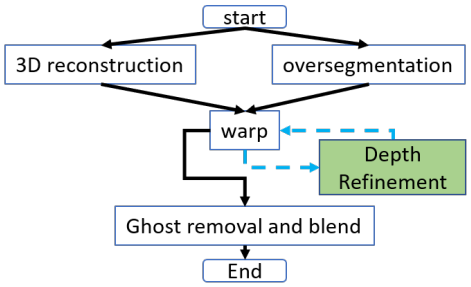

Fig. 1. Flow chart of the whole system.

2.1. Depth Refinement

As Figure 1 shows, the depth refinement method is an additional step, which can be combined with any other warp method. In other words, our depth refinement method is a warp-based method. This refinement method needs three steps, as shown in Figure 2.

For each texture image, we get the over-segmented image and sparse depth firstly. In the previous work [5], superpixels are mainly divided into two parts, which is coarse and inaccurate. Here, we follow three rules and distinguish superpixels into three parts: reliable, half-reliable and unreliable. Since we warp superpixels to the virtual view position by homography, we assume that the region in one superpixel is belonging to the same planar and the depth is regular. Let $Sp = \{Sp_i | i \in \{1...N\}\}$ be the set of the superpixels and $D = \{D_i | i \in \{1...N\}\}$ be the whole depth information in the image, where D_i is the depth information in Sp_i. For Sp_i, we have:

$$Sp_i = \begin{cases} unreliable & if\ D_i < 0.1 * \sum_{j=1}^{N} D_j \\ halfreliable & if\ D_i \geq 0.1 * \sum_{j=1}^{N} D_j \\ & and\ Regular(D_i) = false \\ reliable & if\ D_i \geq 0.1 * \sum_{j=1}^{N} D_j \\ & and\ Regular(D_i) = true \end{cases} \quad (1)$$

where $Regular(D_i)$ is a function that checks whether the distribution of D_i is regular or not. In our experiment, we set 20 uniform intervals in $[min(D),\ max(D)]$. If D_i belongs to one interval, this function returns true, otherwise returns false. After this step, we get S_r, S_g, S_b, representing unreliable, reliable and half-reliable superpixels respectively. We start to operate these superpixels in three steps.

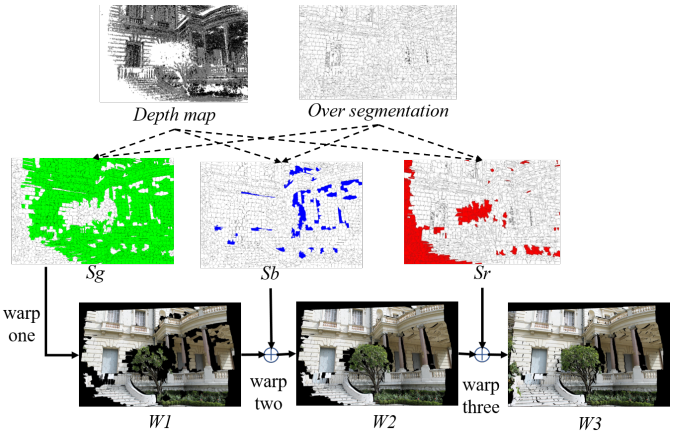

Fig. 2. Pipeline of our depth refinement method, which is combined with warp operation.

Firstly, we apply a grid-based warp for S_g and get the warped image W_1, as shown in Figure 2. There are many black holes and some background objects replacing the front objects, such as the pillars in the middle of the picture.

The next step is to deal with the half-reliable superpixels S_b, which is the main step. For a superpixel S_{bi} in S_b, we follow the method of [14] to build the color histogram, with 16 bins for each color channel and totally a 48D descriptor $\mathcal{H}_{Lab}(S_{bi})$. Then, we obtain the neighboring reliable superpixels' set of S_{bi}, regarded as $NS_{bi} = \left\{NS_{bi}^{j} | j \in \{1...n_{bi}\}\right\}$, where n_{bi} is the number of neighboring reliable superpixels. These reliable superpixels, which have right depth information, are regarded as the ground truth. We calculate the color similarity between

S_{bi} and NS_{bi} by the χ^2 distance and find the most similar superpixel $NS_{bi}^{\hat{j}}$. The $\hat{j}$ is defined as:

$$\hat{j} = \underset{j \in \{1...n_{bi}\}}{\arg\min}\ dist(\mathcal{H}_{Lab}(S_{bi}), \mathcal{H}_{Lab}(NS_{bi}^{j})) \quad (2)$$

We set L_{min} as the minimum distance, where $L_{min} = dist(\mathcal{H}_{Lab}(S_{bi}), \mathcal{H}_{Lab}(NS_{bi}^{\hat{j}}))$. If $L_{min} > T$, we ignore this $\hat{j}$ and put S_{bi} into S_r directly. Otherwise, we put S_{bi} into S_g. In our experiment, we set $T = 40$. Then, we use the mean depth value of $NS_{bi}^{\hat{j}}$ to filter S_{bi} and those depth values not in $[Mean(NS_{bi}^{\hat{j}}) - L_{min},\ Mean(NS_{bi}^{\hat{j}}) + L_{min}]$ are removed. This process is effective and robust even when most of depth values in S_{bi} are wrong. Now, we apply $warp\ two$ and get the warped image $W2$, as shown in Figure 2. After this step, S_b has been divided into S_r and S_g.

Finally, the operation for S_r is similar to that for S_b. For example, we set S_{ri} as one superpixel in S_r and find the most similar superpixel S_{gj} in S_g. Since no useful depth information can be used, we directly use the homography of S_{gj} to warp S_{ri} and get $W3$, as shown in Figure 2.

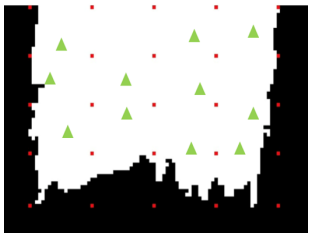

Fig. 3. A superpixel (white region) in grid-based warp, with projected depth points (green triangles) and the sampled points (red points)

2.2. Grid-based Warp

Our depth refinement method can be combined with any warp operation. In our experiment, we choose a grid-based warp, which is also a shape-preserving approach.

Some recent works [5, 13] create an axis-aligned bounding box for the image and get warped images by optimizing an energy function. In our experiment, we also create the bounding box for each superpixel. In Figure 3, the white region is a superpixel, the green triangles donate pixels that have depth information and the red points distribute regularly in the superpixel are sample points.

Let $P = \{P_i | i \in \{1...n\}\}$ be the set of sample points. n is the number of sample points and we set $n = 25$. As D is defined in section 2.1, here we let $D_i = \{D_{ij} | j \in \{1...m_i\}\}$ be the depth points in superpixel Sp_i. We calculate the Euclidean distance between each P_i and D_i in the image coordinate:

$$\hat{j} = \underset{j \in \{1...m_i\}}{\arg\min}\ dist(P_i, D_{ij}) \quad (3)$$

Then, we assign $D_{i\hat{j}}$ to P_i and project P from the input image coordinate to the virtual image coordinate according to the camera rotation and translation in the world coordinate. After that, we get the set of virtual points $P^v = \{P_i^v | i \in \{1...n\}\}$, which correspondences to P one by one in the virtual image coordinate. In fact, some points are not in the image coordinate. We remove them and calculate a new homography in the remaining correspondence points.

Now, we get a coarse-calibrated homography H_c. This matrix still has some noises and needs to be refined. By back-projecting P^v to the input image's coordinate, these matching points whose distances are bigger than 3 pixels are regarded as noised points. We remove them and calculate a fine homography H_n, which is more accurate than the coarse one. Finally, we warp each pixel in this superpixel to the virtual images coordinate by H_n.

Fig. 4. Ghost artifacts removal. The left one is the original image and the right one has removed the ghost artifacts.

2.3. Ghosts Removal and blend

Ghost artifacts usually consist in the edge of image. For this problem, we refer to the work of Mori et al. [4]. To blend the virtual image, we warp three neighbor input images to the novel view. Let $\{W3^1, W3^2, W3^3\}$ be the set of three warped images from different perspectives. We detect their depth edges E by checking the difference of depth value between two neighbor pixels. The detected pixels in the edge mostly have wrong positions and some pixels belonging to background are warped in the foreground and vice versa. For each pixel that belongs to E, we remove its neighboring 3 pixels in the horizon. This simple and effective operation provides dramatic performance gains, like the right image in Figure 4. Finally, we blend $\{W3^1, W3^2, W3^3\}$ by alpha blending [15]. The results are in Figure 5.

3. EXPERIMENTAL RESULTS

To evaluate our method, we use the datasets from [5,6], which include wide-baseline complex geometric environments. We

Fig. 5. Synthesized images. The top line is the result of [5]. The bottom line is our result.

Table 1. PSNR and SSIM

ID	1	2	3	4	5	6	7	8	9	10	11	12	Average
PSNR(dB) [5]	19.17	20.16	20.07	20.16	18.64	18.20	18.58	15.93	20.05	21.66	19.16	19.19	**19.25**
PSNR(dB)-Ours	19.57	20.66	20.84	21.05	19.07	18.51	18.76	16.76	21.02	22.56	20.73	20.25	**19.98**
SSIM [5]	0.591	0.626	0.633	0.650	0.613	0.596	0.615	0.541	0.639	0.683	0.702	0.645	**0.628**
SSIM-Ours	0.623	0.663	0.663	0.685	0.647	0.623	0.641	0.561	0.685	0.723	0.732	0.675	**0.660**
ΔPSNR(dB)	+0.40	+0.50	+0.77	+0.89	+0.43	+0.31	+0.19	+0.83	+0.98	+0.90	+1.57	+1.06	**+0.73**
ΔSSIM*10	+0.32	+0.31	+0.37	+0.35	+0.34	+0.27	+0.27	+0.20	+0.45	+0.40	+0.30	+0.31	**+0.32**

Fig. 6. Warped images. The left one is from [5] and the right one is ours.

mainly compare our method with the state-of-the-art view synthesis algorithm [5, 9].

As the section 2.2 says, we first warp several input images to the virtual view position and blend them to get the finial virtual image. Figure 6 shows the warped results, where the left image is the result of [5] and the right one is ours. We can find that our result has less black hole regions and more accurate relationships between the foreground and background in edge regions. One possible reason we think is as follows. The depth information in hole regions is inaccurate. According to [5], these regions are unreliable, but in our depth refinement method, they are half-reliable and the noises can be removed, which means these regions can be warped with accurate homography in $warp\ two$ step, as shown in Figure 2.

Figure 5 shows the finial synthesis images at the virtual view position. The top line of Figure 5 is the result of [5] and the bottom line is our result. We enlarge the white boxes in Figure 5. It is easy to find that images on the bottom line have fewer ghost artifacts and are clearer than those on the top line. One possible reason is that we get more accurate depth information than [5] after applying our depth refinement method.

We also calculate the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) as quantitative indexes. Table 1 shows the PSNR and SSIM of twelve images in three datasets between [5] and ours. We get nearly 0.7dB gains in PSNR and 0.03 gains in SSIM on average. In Table 1, the first four images belong to $museum$ data, the last four images belong to $tree$ data and the remaining images belong to $street$ data. More results are available at https://github.com/wsAndy/icassp_result.

4. CONCLUSION

In this paper, we propose a robust view synthesis system which includes grid-based warp, depth refinement and ghost artifacts removal. The depth refinement method can be combined with any other warp operation and be used to synthesize images at the virtual view position with only sparse depth information. This method considers the input depth information as much as possible and removes noises effectively. In our system, a grid-based warp method is also implemented and a ghost artifacts removal approach is applied. We nearly get 0.7dB gains in PSNR and 0.03 gains in SSIM on average compared to the state-of-the-art view synthesis algorithm.

5. REFERENCES

[1] Masayuki Tanimoto, Toshiaki Fujii, Kazuyoshi Suzuki, Norishige Fukushima, and Yuji Mori, "Reference softwares for depth estimation and view synthesis," *ISO/IEC JTC1/SC29/WG11 MPEG*, vol. 20081, pp. M15377, 2008.

[2] Seong Gyun Jeong, Chul Lee, and Chang Su Kim, "Motion-compensated frame interpolation based on multihypothesis motion estimation and texture optimization," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4497–4509, 2013.

[3] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.

[4] Yuji Mori, Norishige Fukushima, Tomohiro Yendo, Toshiaki Fujii, and Masayuki Tanimoto, "View generation with 3d warping using depth information for ftv," *Signal Processing: Image Communication*, vol. 24, no. 1-2, pp. 65–72, 2009.

[5] Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis, "Depth synthesis and local warps for plausible image-based navigation," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 3, pp. 30, 2013.

[6] Gaurav Chaurasia, Olga Sorkine, and George Drettakis, "Silhouette-aware warping for image-based rendering," in *Computer Graphics Forum*. Wiley Online Library, 2011, vol. 30, pp. 1223–1232.

[7] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision*. Springer, 2016, pp. 501–518.

[8] Johannes L Schonberger and Jan-Michael Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

[9] Yongwei Nie, Zhensong Zhang, Hanqiu Sun, Tan Su, and Guiqing Li, "Homography propagation and optimization for wide-baseline street image interpolation," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 10, pp. 2328–2341, 2017.

[10] Yasutaka Furukawa and Jean Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.

[11] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski, "Towards internet-scale multi-view stereo," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1434–1441.

[12] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, Sabine Süsstrunk, et al., "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[13] Feng Liu, Michael Gleicher, Hailin Jin, and Aseem Agarwala, "Content-preserving warps for 3d video stabilization," in *ACM Transactions on Graphics (TOG)*. ACM, 2009, vol. 28, p. 44.

[14] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa, "Efficient hierarchical graph-based video segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2141–2148.

[15] Thomas Porter and Tom Duff, "Compositing digital images," in *Conference on Computer Graphics and Interactive Techniques*, 1984, pp. 253–259.