VIEW GENERATION WITH 3D WARPING USING DEPTH INFORMATION FOR FTV

Yuji MORI, Norishige FUKUSHIMA, Toshiaki FUJII and Masayuki TANIMOTO

Graduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, JAPAN

ABSTRACT

Free viewpoint images can be generated from multi-view images using Ray-Space method. Ray-Space data requires ray interpolation so as to satisfy the Plenoptic Function. Ray interpolation is realized by estimating view-dependent depth. Depth estimation is usually costly process, thus it is desirable that this process is skipped from rendering process to achieve real-time rendering. This paper proposes a method to render a novel view image using multi-view images and depth maps which are computed in advance. Virtual viewpoint image is generated by 3D warping, which causes some problems that have not occurred in the method with view dependent depth estimation. We handled these problems by projecting depth map to virtual image plane first and perform post-filtering on the projected depth map. We succeeded in obtaining high quality arbitrary viewpoint images from relatively small number of cameras.

Index Terms— Depth Image Based Rendering, 3D warping, 3D reconstruction

1. INTRODUCTION

Due to recent computer development, advanced image technology has become an active research field. Image Based Rendering (IBR), a method to synthesize a novel view image from captured images, is an example. IBR can render photorealistic images like actually captured images.

A light in 3D space is parameterized with seven parameters, $f(x, y, z, \theta, \phi, \lambda, t)$. Ray-Space [1], Light Field[2], and the Lumigraph[3] were proposed around the same time as the ray description method with four parameters in an assumption that ray travels in 3D space straight without attenuation. Multi-camera images are reset in new 4D space as rays. Virtual views are synthesized by loading ray information in 4D parameter space. However, since captured rays are limited, ray interpolation is needed when the requested ray is unavailable. To generate omnifocal images from a small number of cameras, depth of each pixel in images are estimated as a depth map. Fukusima et al. [4] have succeeded in estimating depth map of the view to generate, called view-dependent depth map, in quasi-real-time. This method brought most of its mighty to real-time optimization. While, high quality depth estimation methods, for example, color segmentation based one has been proposed [5]. These methods outperform real-time estimation in quality especially at edge and occluded area.

In this research, we render a novel view image with 3D warping[6] using depth maps computed in advance, which enable us to use high quality depth map without optimization. However, the way to render a novel view differs from view centered method: each pixel is projected respectively to virtual image plane, and some problems which does not appear in view centered method happens. We have resolved some problems by projecting depth map and performing post-filtering on that.

2. PINHOLE CAMERA MODEL

This section describes pinhole camera model. Let $\tilde{\boldsymbol{M}} = [X, Y, Z, 1]^\top$ be a world point and $\tilde{\boldsymbol{m}} = [u, v, 1]^\top$ be its projection in a camera with projection matrix $\boldsymbol{P}$ in homogeneous coordinates. A camera is modeled by usual pinhole; relationship between $\tilde{\boldsymbol{M}}$ and $\tilde{\boldsymbol{m}}$ is given by

$$\boldsymbol{P}\tilde{\boldsymbol{M}} = s\tilde{\boldsymbol{m}} \tag{1}$$

where s is non-zero scalar and 3×4 matrix $\boldsymbol{P}$ is called camera matrix. Camera matrix $\boldsymbol{P}$ can be decomposed as

$$\boldsymbol{P} = \boldsymbol{K}\left[\begin{array}{cc} \boldsymbol{R} & -\boldsymbol{Rt} \end{array}\right] \tag{2}$$

where 3×3 orthogonal matrix $\boldsymbol{R}$ represents the orientation and 3-vector $\boldsymbol{t}$ represents the position. $\boldsymbol{K}$ is 3×3 upper triangular matrix given by

$$\boldsymbol{K} = \begin{bmatrix} f_u & \gamma & u0 \\ 0 & f_v & v0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

with focal length f, skew parameter γ, and principal point (u_0, v_0) Matrix $\boldsymbol{K}$ is called intrinsic matrix and represents the inner structure of camera. Matrix $\left[\begin{array}{cc} \boldsymbol{R} & -\boldsymbol{Rt} \end{array}\right]$ is called extrinsic matrix and relates the world coordinate system to the camera coordinate system. Note that this model ignores non-linear distortion.

Yuji Mori:email mori@tanimoto.nuee.nagoya-u.ac.jp

978-1-4244-1755-1/08/$25.00 ©2008 IEEE

3. PROPOSED ALGORITHM

Proposed algorithm consists of four steps. Each step is explained in more detail as follow.

3.1. Depth map projection with 3D warping

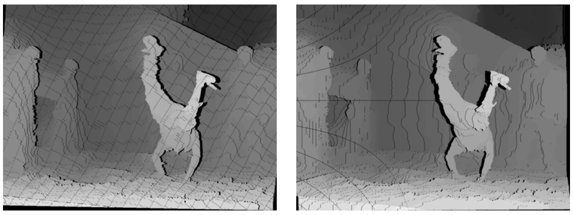
3D warping projects one image to other image plane. As described in equation (1), 3D point $\tilde{\boldsymbol{M}}$ can be reconstructed from image point $\tilde{\boldsymbol{m}}$ using projection matrix $\boldsymbol{P}$ and depth Z. Then reconstructed 3D point is projected to virtual image plane with projection matrix of virtual camera. This step projects depth map of two nearest cameras to virtual image plane. A number of pixels can be projected to same pixel on virtual view image, the nearest point is to be adopted then (Equation 4).

$$\mathrm{Z}(u,v) = \mathop{\arg\min}_{\tilde{\boldsymbol{M_{u,v}}}} Z \tag{4}$$

where $\mathrm{Z}(u,v)$ is depth value at (u,v) on virtual image plane, $\tilde{\boldsymbol{M_{u,v}}}$ is 3D point whose projection is to (u,v).

This method uses two nearest cameras to render a virtual view image: left side and right side. Projected depth maps are shown in Figure 1(a) and Figure 1(b).With the assumption that depth should change smoothly inside same object and the fact that many blank points appeared in the projected depth map, the depth map should be smoothed.

The following step is to perform post-filtering and smooth the depth map.



(a) left side

(b) right side

Fig. 1. projected depth map from two nearest cameras

3.2. Post-filtering on depth map

This step aims at smoothing the depth map and make the image to render clear. Smoothing depth map reduces noise on the result image, and does not blurs the result compared with smoothing the result directly. First, we consider blank points appeared on the projected depth map. There are two reasons that make blank points on the map: round off error and depth discontinuity.

By the Equation (1), image coordinates (u, v) are calculated decimally, but rounded off to integer value. It causes one pixel width blank appears with constant interval. This blank region can be filled by median filter. Depth discontinuity causes lump blank in depth map. These blanks can not be filled by median filter. However, most of discontinuity in depth map is caused by occlusion, so these areas can be filled from opposite-side camera.

3D warping causes not only blank area, irregular change of depth in same object also appears. These irregularities cause unnatural pixel in the rendered image, so it is desirable to smooth away these irregularities. They can be smoothed away by low-pass filter. However, edge region should be preserved since low-pass filtered edge in depth map blurs edge of objects in renderd image. Considering that, low-pass filter which does not smooth edge area is preferable, so we adopted bilateral filtering[7], defined as Equation (5).

$$\mathrm{h}(\mathrm{x}) = k^{-1} \iint_D f(\xi) c(\xi - \mathrm{x}) s(f(\xi) - f(\mathrm{x})) d\xi \tag{5}$$

where k is normalization constant, D is filtering domain. This is shift-invariant Gaussian filtering, in which both the closeness function c and similarity function s are Gaussian functions. More specifically, c is radically symmetric:

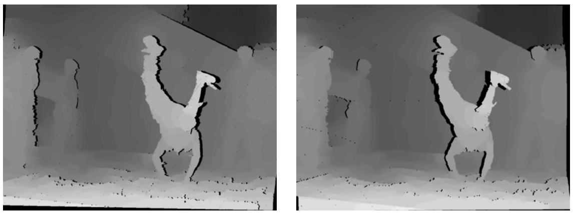
$$c(\xi - \mathrm{x}) = \exp\left(-\frac{1}{2}\left(\frac{|\xi - \mathrm{x}|}{\sigma_d}\right)^2\right) \tag{6}$$

where σ_d is the variance of Euclidean distance. The similarity function s is perfectly analogous to c:

$$s(\xi - \mathrm{x}) = \exp\left(-\frac{1}{2}\left(\frac{|f(\xi) - f(\mathrm{x})|}{\sigma_r}\right)^2\right) \tag{7}$$

where σ_r is the variance of color space. By introducing this factor, far point in color space would be less weighted from filtering kernel, which preserves edge region from smoothing.

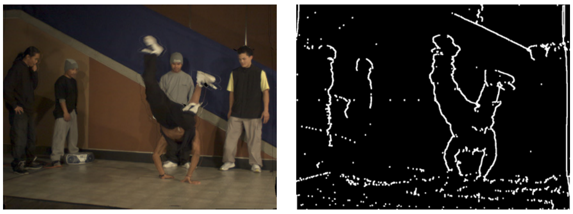
The following step is to render virtual view image with depth map shown in Figure 2(a) and 2(b)



(a) left side

(b) right side

Fig. 2. post-filtered depth maps



(a) blended image (b) boundary region

Fig. 3. blended image and boundary region

3.3. Boundary matting and inpainting

After performing post-filtering on depth maps, these two depth maps are projected to each real camera respectively. Then, virtual view image shown in Figure 3(a) is rendered by blending two neighboring images as described in Equation (8).

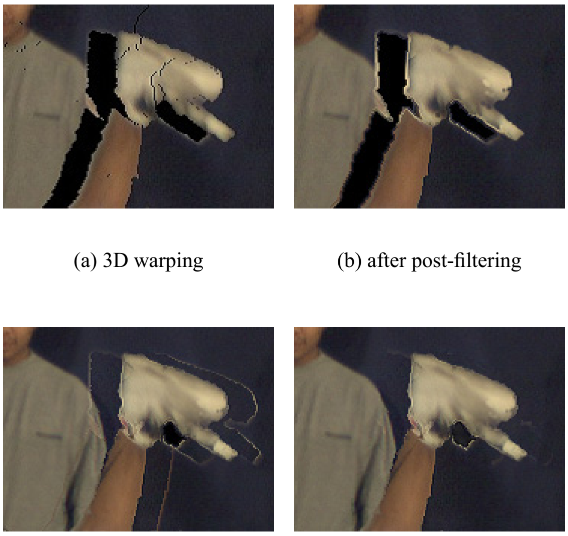
$$I(u,v) = \begin{cases} (1-\alpha)I_L(u_L,v_L)+\alpha I_R(u_R,v_R) \\ I_L(u_L,v_L) \\ I_R(u_R,v_R) \\ 0 \end{cases} \quad (8)$$

$$\alpha = \frac{|\boldsymbol{t}-\boldsymbol{t_L}|}{|\boldsymbol{t}-\boldsymbol{t_L}|+|\boldsymbol{t}-\boldsymbol{t_R}|} \quad (9)$$

with $I(x,y)$ being pixel value at (x,y) virtual image plane, I_L, I_R being reference image plane, $(u_L,v_L)$and(u_R,v_R) being projected point to reference camera from (u,v) on virtual image plane, $\boldsymbol{t}$ being translation vector of extrinsic matrix.

Where both projected images had a value, the pixel value of blended image would be computed by adding two values weighted by coefficient α defined in Equation(9). Where only one projected image had a value, the area would be considered as occlusion and pixel value would be copied from one image, and where neither images had a value, the pixel value of blended image would be remain 0.

Due to miss-focus and half-pixel problem, object can have ill-defined at borders. It makes depth estimation difficult and as a result, unnatural pixel appears around the boundary region. To reduce this unnaturalness, we conducted boundary matting. Figure 3(b) shows the boundary region. This matting actually expanded the occlusion to direction for background, meaning the area whose pixel value is copied from one reference image. It erases the mixture of foreground and back ground colors. The remaining blank area was filled with inpainting [8], which is usually used, for example, to mend damaged part of image or erase subtitle. Figure 4 shows the closeup of images obtained at some rendering steps.



(a) 3D warping (b) after post-filtering

(c) blended image (d) final output

Fig. 4. closeup sample results at rendering steps

4. EXPERIMENTAL RESULTS

For the experiment of our algorithm, we used a sequence named beakdancers, generated and distributed by Interactive Visual Group at Microsoft Research. This data includes a sequence of 100 images captured from 8 cameras. Depth maps, computed from stereo, are also included for each camera along with the calibration parameters. The captured images have a resolution of 1024×768. Camera configuration is shown in Figure 5. More detailed description is in [9]. Figure 7 is one view synthesis result.

To calculate PSNR (Peak Signal-to-Noise Ratio), we generated the image of camera4 which locates at center of camera array and the origin of world coordinates. Figure 6 shows the result of calculating PSNR with changing distance between

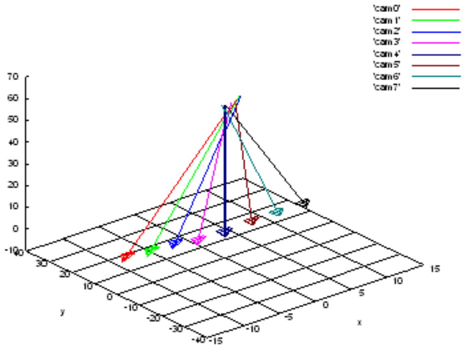


Fig. 5. Camera configuration

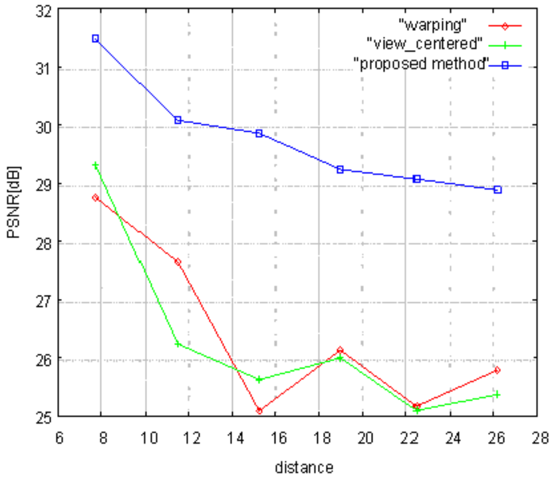


Fig. 6. PSNR with changing distance

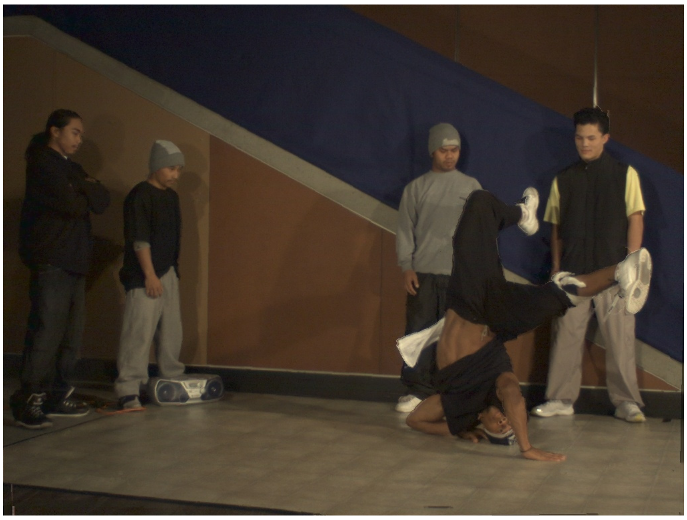


Fig. 7. View synthesis result

cameras given by $|\boldsymbol{t} - \boldsymbol{t_L}| + |\boldsymbol{t} - \boldsymbol{t_R}|$.

As a comparison, we also calculated with images generated by view centered method and 3D warping. Our method improved on PSNR by up to 4dB from the other two methods.

5. CONCLUSION

In this paper we proposed novel free viewpoint image generation method that solves some of DIBR (Depth Image Based Rendering) problems. This method renders view-dependent depth map first and perform post-filtering from the assumption that the depth value inside same object changes smoothly. As a result, unnaturalness which depth warping causes can be soothed.

As future work we will run our algorithm on parallel stereo sequence and compare with other algorithms. Studying more about depth warping and finding more suitable post-filtering process, furthermore, finding pre-filtering which reduces blank points on projected depth map.

6. ACKNOWLEDGEMENTS

We would like to thank Interactive Visual Media Group, Microsoft Research for distributing multi-camera video with fine quality depth data and camera parameters.

7. REFERENCES

[1] T. Fujii, T. Kimoto, and M. Tanimoto, "Ray space coding for 3d visual communication," in *Proc. Picture Coding Symp.*, 1996, vol. II, pp. 447–451.

[2] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. ACM SIGGRAPH'96*, 1996, pp. 31–42.

[3] S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Cohen, "The lumigraph," in *Proc. ACM SIGGRAPH'96*, 1996, pp. 43–54.

[4] N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "Free viewpoint image generation using multi-pass dynamic programming," in *Proc. of SPIE Stereoscopic Displays and Virtual Reality Systems*, 2007, vol. XIV, pp. 460–470.

[5] A. Klaus, M. Sormann, and K. Kamer, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proc. ICPR'06*, vol. III.

[6] W. Mark, L. Mcmillan, and G. Bishop, "Post-rendering 3d warping," in *Proc. Symposium on I3D Graphics*, 1997, pp. 7–16.

[7] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. ICCV'98*, 1998, p. 839.

[8] A. Telea, "An image inpainting technique based on the fast marching method," in *J. Graphics Tools*, vol. IX.

[9] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layerd representation," in *Proc. ACM SIGGRAPH and ACT Trans. on Graphics Los Angeles, CA*, Aug 2004, pp. 600–608.