Joint Multilateral Filtering for Stereo Image Generation Using Depth Camera

Yo-Sung Ho and Sang-Beom Lee

Abstract In this paper, we propose a stereo view generation algorithm using the Kinect depth camera that utilizes the infrared structured light. After we capture the color image and the corresponding depth map, we first preprocess the depth map and apply joint multilateral filtering to improve depth quality and temporal consistency. The preprocessed depth map is warped to the virtual viewpoint and filtered by median filtering to reduce truncation errors. Then, the color image is back-projected to the virtual viewpoint. In order to fill out remaining holes caused by disocclusion areas, we apply a background-based image in-painting process. Finally, we obtain a synthesized image without any visual distortion. From our experimental results, we realize that we obtain synthesized images without noticeable errors.

Keywords Depth image-based rendering • Kinect depth camera • Multilateral filtering • Three-dimensional television

1 Introduction

Three-dimensional television (3DTV) is the next-generation broadcasting system. Owing to advances in display devices, such as stereoscopic or multi-view displays, 3DTV provides users with a feeling of "being there", or presence, from the simulation of reality [1]. In this decade, we expect that the technology will be progressed enough to realize the 3DTV including content generation, coding, transmission, and display.

Y.-S. Ho (✉) • S.-B. Lee
Gwangju Institute of Science and Technology (GIST), 261 Cheomdan-gwagiro,
Buk-gu, Gwangju 500-712, Republic of Korea
e-mail: hoyo@gist.ac.kr

J.S. Jin et al., *The Era of Interactive Media*,
DOI 10.1007/978-1-4614-3501-3_31,

In 2002, the advanced three-dimensional television system technologies (ATTEST) project began the research for 3DTV [2]. ATTEST introduced a novel 3D broadcasting system including four main stages: 3D contents generation, coding, transmission, and rendering/display. While the previous approach dealt with two stereoscopic video streams—one for the left view and one for the right view—on the broadcasting system, ATTEST adopted two streams for monoscopic video and the corresponding depth map that is composed of per-pixel depth information.

The virtual image can be synthesized by a depth image-based rendering (DIBR) technique using the color video and the corresponding depth video [3]. We can deal with the depth map as 3D information of the real scene. The virtual image can be generated by following procedure. First, whole pixels of the color image of the original viewpoint are back-projected to the world coordinate using the camera geometry and the depth map. Then, the points in the world coordinate are reprojected on the image plane of the virtual viewpoint. This procedure is called "3D warping" in the computer graphics literature [4].

Although the DIBR technique is suitable for 3DTV, it has some problems. The most significant problem of the DIBR technique is that when we synthesize the virtual image, we can see newly exposed areas, which are occluded in the original view but become visible in the virtual images. These areas are called disocclusion. This disocclusion area is an annoying problem since the color image and the depth map cannot provide any information. Therefore, the disocclusion areas should be filled out so that the virtual image seems more natural.

In order to remove the disocclusion, several solutions were introduced. Those methods are mainly categorized by two approaches: filling out the disocclusion by using near color information such as interpolation, extrapolation, mirroring of background color, and preprocessing using a Gaussian smoothing filtering [3]. Recently, an asymmetric smoothing filtering is proposed for preprocessing [5]. This method reduces not only the disocclusion areas but also the geometric distortion that is caused by a symmetric smoothing filter.

While the disocclusion and the geometric distortion are mostly removed by the asymmetric depth map filtering, the synthesized view is deformed due to the distorted depth map. Recently, many solutions based on depth map filtering have been tried to solve the problem of the low depth quality. One of the solutions is the depth map filtering near the object boundary [6]. Although we can reduce the deformation of the depth map by restricting the filtered areas, the depth quality is still unsatisfactory.

In this paper, we propose a stereo view generation algorithm. The main contribution of this paper is that we synthesize the virtual image using the original color image and the preprocessed depth map and also we implement the entire process by aiming at the depth camera which utilizes the infrared structured light. The virtual image can be obtained by preprocessed depth map in virtual viewpoint and background-based image in-painting process.

2 Depth Image-Based Rendering (DIBR) Techniques

Color video and depth video can be used for synthesizing the virtual images in DIBR technique. The block diagram of DIBR technique is depicted in Fig. 1. Each process is explained in detail in this section.

2.1 Depth Map Preprocessing

When synthesizing the virtual image, we can find the disocclusion area. Since there is no information of the disocclusion area, we need to fill out it. One of the solutions is preprocessing of depth map using smoothing filter [3]. The main advantage of smoothing is that the sharpness of depth discontinuity is weakened and most disocclusion areas are filled with neighboring pixels.

Figure 2 shows various smoothing results for "Interview". As shown in Fig. 2b, the simple smoothing filter can fills out the disocclusion areas. However, it causes a geometric distortion that the vertical edges of the synthesized image are bent. This problem gives the discomfort to viewers. In order to reduce the geometric distortion, asymmetric smoothing method is proposed [5]. In this approach, the strength of filtering of a depth map in the horizontal direction is less than that in the vertical direction. Figure 2c shows the asymmetric smoothing result.

The synthesized image after asymmetric smoothing of the depth map has good subjective quality. However, the filtered depth map has many errors. It is desirable that the filter is applied so that the filtered areas are reduced through the prediction of the disocclusion areas. By aiming at this assumption, discontinuity-adaptive depth map filtering is proposed [6]. This approach assumes that the disocclusion area is detected nearby object boundaries and the depth map is filtered only near those regions. Therefore, the filtered region of the depth map is reduced. As shown in Fig. 2d, the deformation of the object is reduced.

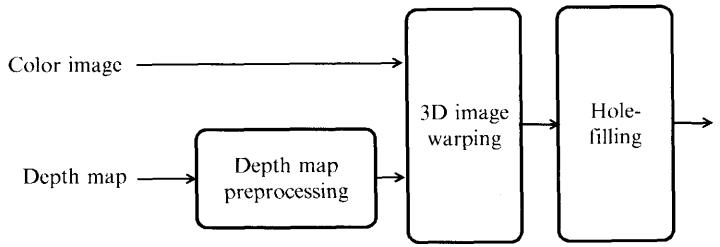

Fig. 1 Block diagram of depth image-based rendering technique

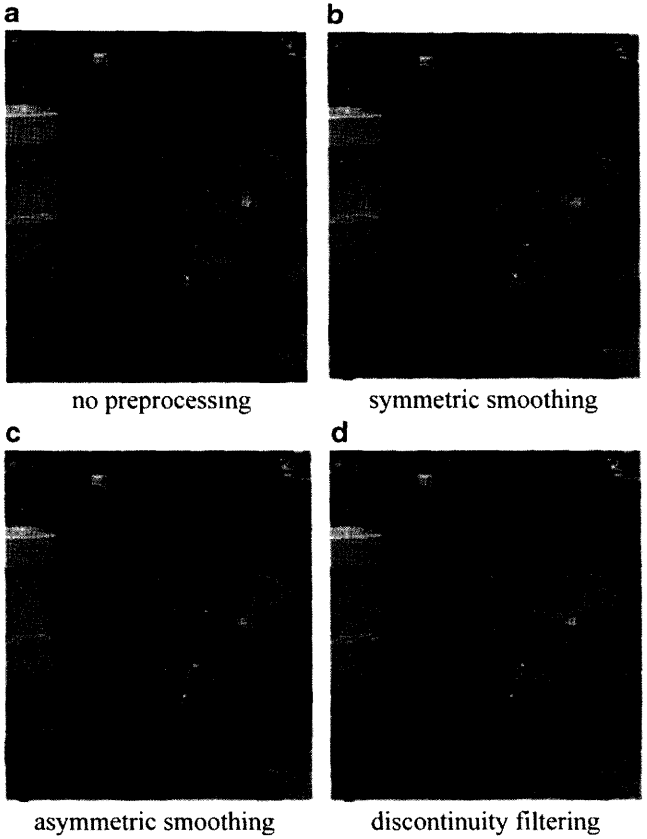

Fig. 2 Smoothing results for "Interview"

2.2 3D Image Warping

We assume that the camera configuration is parallel for simplicity. There are two approaches of stereoscopic image generation using DIBR technique. One is generating a virtual left image so that the original view is regarded as the right view. Another method is generating both the virtual left and right view by using original view. The first approach has the lowest quality of the left view since this view has the largest disocclusion areas compared to the second method. However, it gives us the highest quality for the right view. We adopt the first method since several conventional works proved that the binocular perception performance is determined by only one view which is higher quality than the other view [7].

Figure 3 shows the relationship of the pixel displacement and the real depth. The new coordinates (x_l,y) of the virtual viewpoint from the original coordinates (x_r,y) according to the depth value Z is determined by

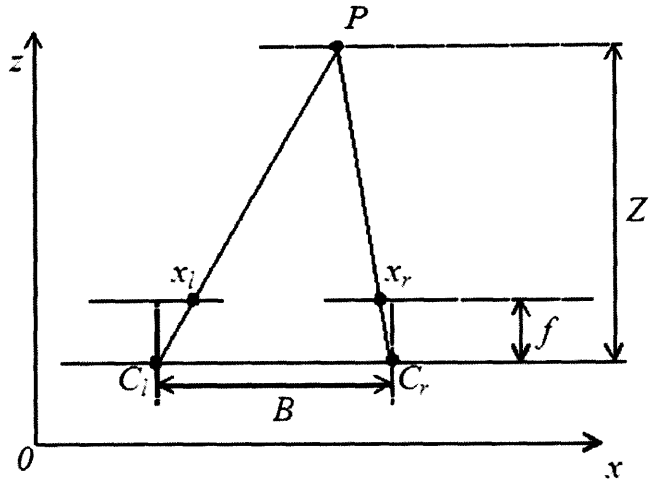

Fig. 3 Relationship between disparity and depth

$$x_l = x_r + \frac{fB}{Z} \tag{1}$$

where f represents the focal length of the camera and B represents the distance between cameras.

2.3 Hole-Filling

After depth map preprocessing and 3D warping, most unknown regions of the virtual image are filled out. Due to the truncation error in the 3D warping process, the small-sized holes are remained. Therefore, we need to fill those holes. The common method in this step is linear interpolation using neighbor pixels.

3 Proposed Stereo View Generation Algorithm

The proposed method exploits a depth camera, which interprets 3D scene information from a continuously-projected infrared structured light [8]. Figure 4 shows the overall block diagram of our algorithm. The first three steps are categorized by depth map preprocessing and remaining parts are included in the view synthesis operation.

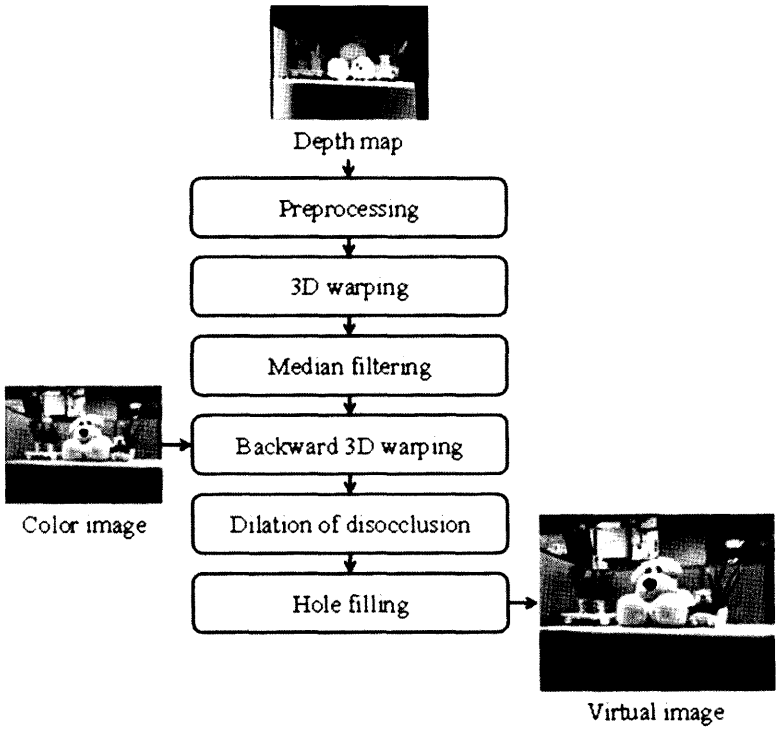

Fig. 4 Overall block diagram of the proposed method

3.1 Depth Map Preprocessing Using Image Inpainting

Since the position of the transmitter of infrared structured light and the receiver is different and there exist errors of the sensor itself, we obtain the depth map with some areas where the infrared sensor cannot retrieve the depths. Therefore, in the preprocess step, these areas are filled out by original image in-painting algorithm [9]. Figure 5 shows the depth preprocessing result using image in-painting algorithm.

After the image inpainting, the depth map is filtered by a temporal filtering using joint multilateral filter. We expand the conventional algorithm that exploits the joint bilateral filter to temporal domain [10]. The filter is designed by

$$D(x,y) = \arg\min_{d \in d_p} \frac{\sum_{u \in u_p} \sum_{v \in v_p} \sum_{w \in w_p} W(u,v,w) \cdot C(u,v,w,d)}{\sum_{u \in u_p} \sum_{v \in v_p} \sum_{w \in w_p} W(u,v,w)} \tag{2}$$

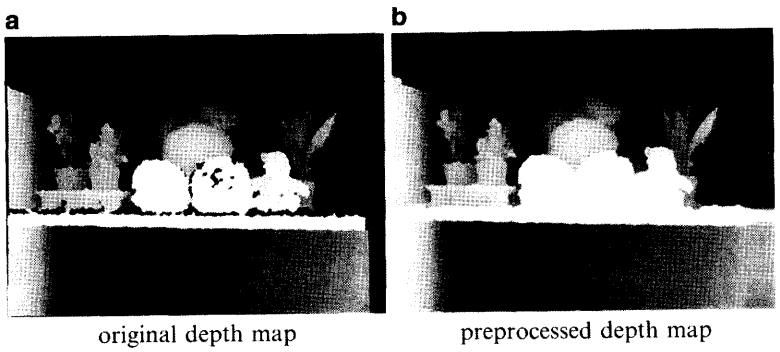

Fig. 5 Result of depth map inpainting

where $p=(x,y)$, $d_p=\{D(x-1,y,t),\ D(x+1,y,t),\ D(x,y-1,t),\ D(x,y+1,t),\ D(x,y,t-1),\ D(x,y,t+1)\}$, $u_p=\{x-r,\ldots,x+r\}$, $v_p=\{y-r,\ldots,y+r\}$, $w_p=\{t-r,\ldots,t+r\}$. Here, $W(u,v,w)$ and $C(u,v,w,d)$ is computed by

$$W(u,v,w) = \exp\left\{-\frac{\|I(x,y,t), I(u,v,w)\|^2}{2\sigma_R^2}\right\} \cdot \exp\left\{-\frac{(x-u)^2+(y-v)^2+(t-w)^2}{2r^2}\right\} \tag{3}$$

$$C(u,v,w,d) = \min(\lambda\Gamma, |D(u,v,w) - d|) \tag{4}$$

where λ is a constant to reject outliers.

We apply an outlier reduction operation in the temporal domain to avoid a motion estimation or optical flow technique for moving objects. Therefore, the temporal position w_p is selected by

$$w_{outlier\_reduction} = \{w_p \mid |I(x,y,t) - I(x,y,w_p)| < 2\lambda L \quad |D(x,y,t) - D(x,y,w_p)| < \lambda L\} \tag{5}$$

After the preprocessing, the 3D warping operation is performed using the depth map. During this step, the warped depth is truncated in integer value and as a result, the depth map includes truncation errors. These errors are easily removed by median filtering. Figure 6a shows the warped depth map and Fig. 6b shows the result of median filtering.

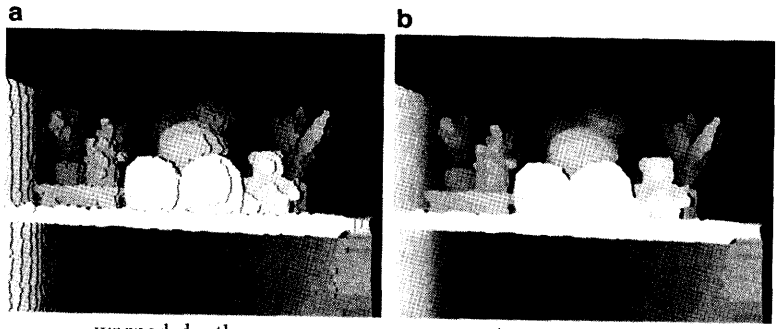

Fig. 6 Result of 3D warping

Fig. 7 Back-projected color image

3.2 *Virtual View Synthesis*

Using the warped depth map, the color image can be back-projected. It is computed by

$$I_{virtual}(x, y) = I_{original}\ \{x + D(x, y), y\} \tag{6}$$

where $D(x,y)$ represents the depth value at pixel position (x,y). The back-projected color image is shown in Fig. 7. As shown in Fig. 7, most of pixels are filled but remaining holes near the object are found.

In order to fill out those holes, we exploit the background in-painting operation. The in-painting algorithm first defines the region to be in-painted Ω and its boundary $\partial\Omega$ and the pixel p, the element of Ω is in-painted by its neighboring region $B_\varepsilon(p)$. In the proposed algorithm, we replace the boundaries facing the foreground with the corresponding background region located on the opposite side. This can be calculated by

$$p_{fg} \in \partial\Omega_{fg} \rightarrow p_{bg} \in \partial\Omega_{bg} \tag{7}$$

$$B_\varepsilon(p_{fg}) \rightarrow B_\varepsilon(p_{bg}) \tag{8}$$

where fg and bg represent the foreground and the background, respectively.

4 Experimental Results

We have evaluated the proposed algorithm with two aspects: visual quality and computational time. The resolution of the color image and the depth map is 640×480. The parameters for stereo view generation are set as follows: $B = 48$ mm for the distance between cameras and $f = 200$ mm for the focal length of the camera.

Figure 8 shows the view synthesis result. Figure 8a shows the original color image and Fig. 8b–d represents the synthesis results of asymmetric filter, discontinuity-adaptive filter, and the proposed algorithm, respectively. As shown in Fig. 8d, remaining holes are naturally removed compared to other methods since the proposed algorithm conducted the background-based image in-painting operation.

Figure 9 shows the enlarged figures of Fig. 8. As shown in Fig. 9a, b, there still remains the geometric errors in background. However, even though the proposed algorithm performed relatively unnatural hole-filling, it never deformed the depth map at all and caused geometric errors.

Table 1 shows the computational time of each process. From those results, the proposed system enabled the entire processing up to 18.87 fps. Without any techniques for real-time processing, such as GPU programming or fast algorithms, stereo video was easily generated in nearly real-time.

5 Conclusions

In this paper, we have proposed a stereo view generation algorithm using depth camera. The proposed scheme focused on the natural view synthesis. Therefore, we performed the depth preprocessing and view synthesis. The depth map is

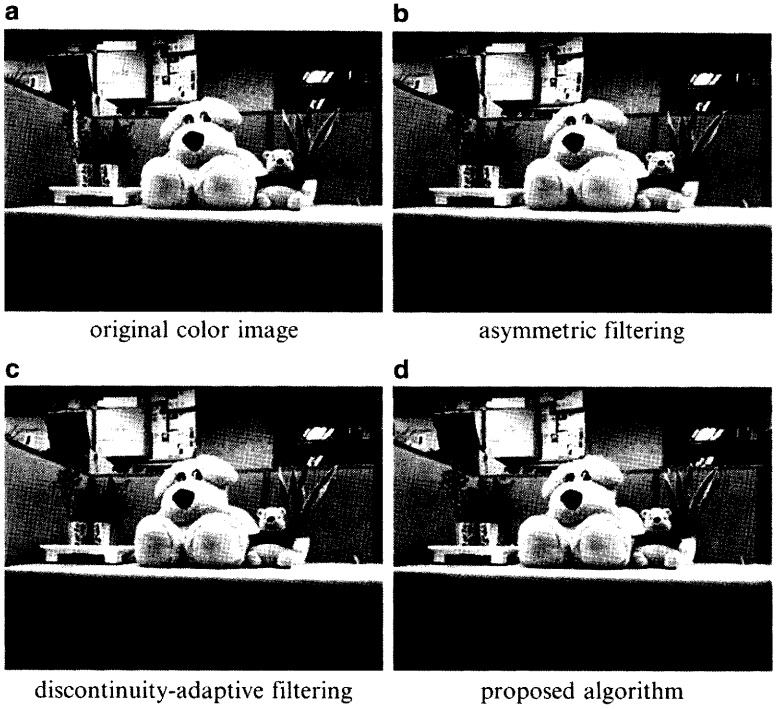

Fig. 8 Results of view synthesis

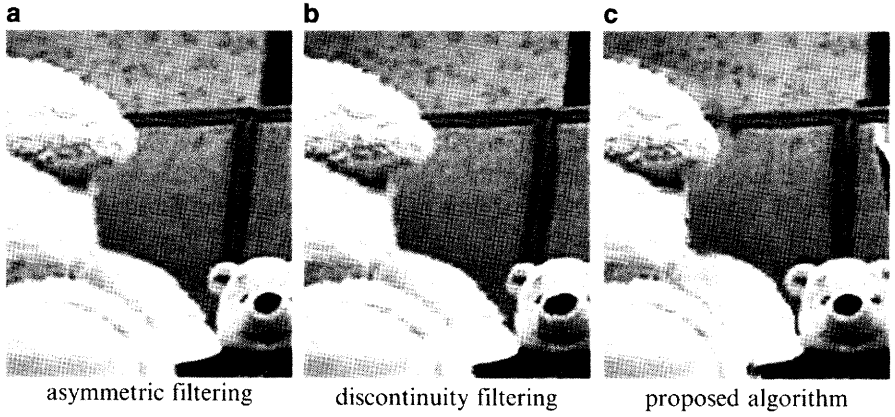

Fig. 9 Results of view synthesis

preprocessed by several image processing techniques and the synthesized image is obtained by background-based image in-painting operation. From experimental results, we noticed that we obtained the natural synthesized image.

Table 1 Computational time

Process	Computational time (ms)
Depth preprocessing	12.00
3D warping	11.00
View synthesis	5.00
Background in-painting	25.00
Total	53.00 (18.87 fps)

Acknowledgement This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0030822).

References

1. Riva, G., Davide, F., Ijsselsteijn, W.A.: Being There: Concepts, Effects and Measurement of User Presence in Synthetic Environments. Amsterdam, The Netherlands: Ios Press, (2003)
2. Redert, A., Op de Beeck, M., Fehn, C., IJsselsteijn, W., Pollefeys, M., Van Gool, L., Ofek, E., Sexton, I., Surman, P.: ATTEST-Advanced Three-Dimensional Television System Technologies. Proc. of International Symposium on 3D Data Processing, 313–319 (2002)
3. Fehn, C.: Depth-image-based Rendering (DIBR), Compression and Transmission for a New Approach on 3D TV. Proc. of SPIE Conf. Stereoscopic Displays and Virtual Reality Systems XI, Vol. 5291, 93–104 (2004)
4. Mark, W.R.: Post-Rendering 3D Image Warping: Visibility, Reconstruction, and Performance for Depth-Image Warping. PhD thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, (1999)
5. Zhang, L., Tam, W. J.: Stereoscopic Image Generation Based on Depth Images for 3D TV. IEEE Trans. on Broadcasting, Vol. 51, 191–199 (2005)
6. Lee, S., Ho, Y.: Discontinuity-adaptive Depth Map Filtering for 3D View Generation. Proc. of Immersive Telecommunications, T8(1–6) (2009)
7. Stelmach, L., Tam, W., Meegan, D., Vincent, A., Corriveau, P.: Human Perception of Mismatched Stereoscopic 3D Inputs. Proc. of International Conference on Image Processing, Vol. 1, 5–8 (2000)
8. PrimeSense, http://www.primesense.com/?p=487
9. Telea, A.: An Image Inpainting Technique based on The Fast Marching Method. Journal Graphics Tools, Vol. 9, 25–36 (2004)
10. Yang, Q., Wang, L., Ahuja, N.: A Constant-Space Belief Propagation Algorithm for Stereo Matching. Proc. of Computer Vision and Pattern Recognition, 1458–1465 (2010)