Virtual View Synthesis Using Moving Multi-Camera Array

Su-Min Hong and Yo-Sung Ho
Gwangju Institute of Science and Technology (GIST)
123 Cheomdan-gwagiro, Buk-gu, Gwangju, 500-712, Republic of Korea
Telephone +82-62-715-2258, Fax: +82-62-714-3164
Email:{sumin,hoyo}@gist.ac.kr

Abstract— **Considerable improvements of information technologies (IT) allow consumers to enjoy various forms of multimedia services. Recently, 3D video has been established as a new format that provides more realistic and natural experiences to users by free-viewpoint TV (FTV). One of the main challenges is about rendering continuous viewpoint images using color and depth information. Thus, image interpolation, which generates virtual view-point images, is a key part in 3D display systems. In this paper, we propose a virtual view synthesis method for a moving multi-camera array using stereo images. Our method synthesizes virtual viewpoint images using both spatial and temporal information. The proposed method consists of two parts: depth estimation and virtual viewpoint image generation. In order to evaluate the view synthesis method, the peak signal-to-noise ratio (PSNR) of depth images are estimated. Our results show that the PSNR results from the proposed method outperforms the conventional method.**

Keywords— FTV; view synthesis; depth map upsampling; moving camera array

I. Introduction

Virtual view synthesis is one of the most important techniques to make high quality videos in the free viewpoint television and three-dimensional (3D) video. Unfortunately, efficiently making virtual view of complex real world locations is a challenging task. Depth image-based rendering (DIBR) is generally used to synthesize a virtual view in free viewpoint television (FTV) and 3D video [1].

In the previous works, most efforts were focused on static multi-camera system. Many static multi-camera systems were presented to capture the scene and generate 3D video contents. Most of them are composed of fixed multiple color cameras with parallel arrangement. In this case, if an object moves widely, the number of camera should be increased. As a result, camera setup and computation cost of the algorithm will be increased. Alternatively, we can reduce the number of cameras by increasing the camera interval. However, in this case, only sparse information can be used in the view synthesis, so that the quality of the generated free-viewpoint image significantly degrades and it is difficult to expect high quality of the 3D video.

The contributions of this paper include virtual view synthesis method in the moving multi-camera array, and acquire high quality depth maps [2]. To achieve this goal, first we propose a moving camera array. This camera systems captures multi-view color videos at 30 frame per second and color resolution is 1280×720. To capture the moving object, our proposed multi-camera array can move left and right sides. Second, we propose a depth map refinement method to improve a virtual views quality. The proposed algorithms is based on noise aware filter for depth upsampling (NAFDU) [3]. Also, we consider reliability of color and depth pixels to solve an edge blurring problem. Third, we propose a virtual view synthesis method in the moving multi-camera array. In our method, we synthesis virtual viewpoint images using both spatial and temporal information. Once the disparity maps are determined from left and right image sequences, we are able to synthesis a virtual view flexibly. After that, we search the reference image using the color differences between the virtual view and candidate images. Then we divide the moving and static areas differences virtual view image and reference image. Finally, we obtain the improved virtual view images using reference information.

II. View Synthesis Using Moving Camera Array

So far, the focus in virtual view synthesis was concentrated on static camera array. In the static camera array, range of virtual viewpoint image generation is limited. In particular, the important environments are often actively in use, containing moving objects, such as player in speed skating. In this case, the number of camera should be increased. Therefore, the system complexity and cost of camera system setup increase. To reduce the complexity and cost, the number of cameras can decreased by increasing the camera interval. But, in this case, only sparse information can be acquired, so that the quality of generated virtual viewpoint image significantly degrades.

To solve this problem, our proposed method using the moving multiple camera array. By using moving multiple camera array, we can follow an object of interest and also obtain a dense ray information of the scene. So moving multiple camera array can generate the virtual viewpoint images in a wider range in comparison with the static camera array. Fig. 1 shows range of capturing space in moving multiple camera array.

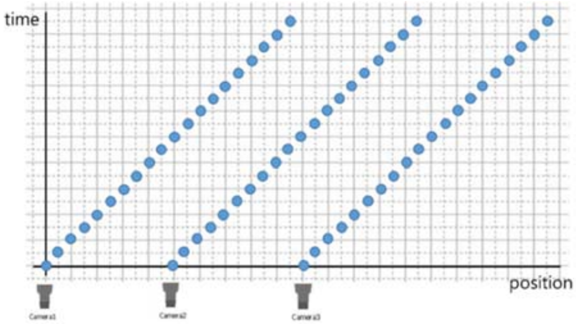


Fig. 1. Range of capturing space in moving camera array

A. Depth map estimation

When we make the virtual viewpoint image, depth or disparity information is required. And synthesized view quality is highly depends on the depth information. In general, there are two types of techniques for acquiring depth information of the scene. One is based on passive range sensors and the other is based on active range sensors, such as stereo matching. Stereo matching can be categorized into local, global and semi-global methods. Local methods are processed by windows based on correlation where the disparity is assumed to be equal for all pixels within the correlation window. In global methods, the task of computing disparities is cast as an energy minimization problem. Generally, global methods are computationally complex even for low resolution images with a small disparity range. The execution time leaves still room for improvement. There is a third category of stereo algorithms lying between global stereo and local stereo. These hybrid stereo algorithms are known as semi-global matching algorithms (SGM) [4]. To make the high quality disparity map fast, we use the SGM method.

B. Depth map refinement

When we make the virtual viewpoint image, depth or disparity information is required. And synthesized view quality is highly depends on the depth information. Unfortunately, stereo matching remains a difficult vision problem. So our proposed method use the multilateral filter to improve quality of disparity maps. The proposed algorithms is similar to our previous work [5].

First of all, we compute the standard deviation in filter kernel. In statistics, the standard deviation (σ) is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A standard deviation close to 0 indicates that the data points tend to be very close to the mean of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values. Standard deviation in the filter kernel is defined by (1).

$$\sigma_p = \sqrt{\frac{\sum_{i=1}^{n}(q_i - M_p)^2}{n}} \tag{1}$$

where p denote the center pixel in the filter kernel and q represent the neighbor pixels in the filter kernel. n is the number of pixels in the filter kernel. Also, we can get the mean value in the filter kernel using (2).

$$M_p = \frac{1}{n}\sum_{i=1}^{n} q_i \tag{2}$$

The standard deviation is a measure of how spreads out pixel values are. In addition, it can represent the dispersion of a set of pixel values from the filter kernel. So, we can decide which region is the edge by using standard deviation. Our proposed method use the adaptive weighting functions based on edge information. Blending function is given by Eq. (3).

$$\alpha(\mathrm{x}) = \frac{1}{1 + e^{-\varepsilon(x-\tau)}}, \mathrm{x} = \frac{\sigma}{\sigma_{max}} \tag{3}$$

blending function is the distribution of the 0 to 1 [3]. Here x is the ratio between the result of Eq. (1) and the maximum standard deviation in each kernel. The largest standard deviation possible will be found where the largest possible gaps are. So, we can calculate the largest standard variation very easy. The maximum standard deviation is calculated as in (4).

$$\sigma_{max} = \sqrt{(M - \min(\Omega))(\max(\Omega) - M)} \tag{4}$$

where Ω represent the spatial neighborhood around *p*. *M* indicates the mean in the Ω and min, max represents the minimum depth value and maximum depth value in the Ω. The standard deviation distribution is normalized in 0 to 1 by consider maximum standard deviation. Also, it can represent how much edge information is included in each kernel. If standard deviation is very small in some region then the x have almost zero. Otherwise, x is the close to one when the standard deviation is almost same with maximum standard deviation. So, we can determine the edge in the filter kernel by using the x value.

After that, we consider reference pixel reliability in edge region. In the conventional bilateral filter, if the depth map and the color image differ significantly in appearance, the use of the color image to upsample the depth map leads to artifacts. To reduce these artifacts, we are considering reference pixel reliability. First, we decide the edge region when x is greater than threshold. In the edge region, we calculate the mean value of the color image and depth map, via the Gaussian weighted sum of neighbors within a filter kernel.

$$M_p = \frac{1}{n}\sum_{i=1}^{n} G_\sigma(||p - q_i||)q_i \tag{5}$$

After that, we choose the reference pixels in the color image and depth map which difference between Gaussian weighted mean value and neighbor pixel is smaller than threshold. In our implementation, we empirically set it equal to 10.

$$|M_p - q_i| < th \tag{6}$$

So, the edges can be preserved by considering reference pixels reliability. Assume that there are input low quality disparity map I, output disparity map $\tilde{S}$, high resolution color image $\tilde{I}$ and low quality target depth map I_p. After blending function is determined, depth value $\tilde{S}_p$ is defined by Eq. (7).

$$\tilde{S}_p = \frac{1}{k_p}\sum_{q_\downarrow\in\Omega} I_{q_\downarrow} f(\|p_\downarrow - q_\downarrow\|)[\alpha(x)g(\|\tilde{I}_p - \tilde{I}_q\|) + (1-\alpha(x))h(\|I_{p_\downarrow} - I_{q_\downarrow}\|)] \qquad (7)$$

If the x is increased, α is close to 1 and proposed method refined quality of input disparity map while protecting the edge information. Otherwise, If the x is decreased, α is close to 0 and proposed method use the depth map information.

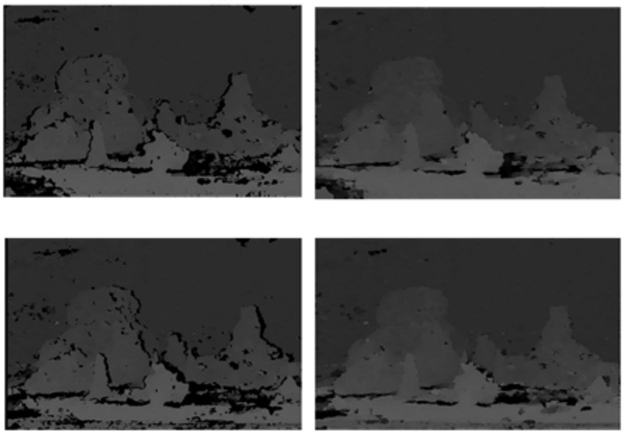


Fig. 2. Result of depth map filtering

C. Virtual view synthesis

Our method makes two left and right disparity map and warp virtual images respectively. Then, two virtual images are summed by weighting function. Although it has heavy complexity due to two disparity estimation process, it generates good quality virtual image. The static camera array has two or more static cameras in which a viewpoint corresponds to video. Static camera array stay still in time. All cameras are synchronized and can acquire dynamic scene from several viewpoints. The static camera array should be set so that the cameras can cover a required range of 3-D space.

Based on refined disparity maps over stereo image sequences, virtual views can be synthesized at any virtual camera position. It is represented by the ratio of base line, which is distance between the left and right cameras. Our proposed method using the linear interpolation to make the virtual viewpoint images. Virtual viewpoint image I' is defined by Eq. (8).

$$I'(x') = (1-\alpha)I_L(x+\alpha d) + \alpha I_R(x+(1-\alpha)d) \qquad (8)$$

where I_L and I_R are the left and the right images, respectively. *x* and *x*' are pixel positions, d is disparity vector and α is the ratio of baseline. In the moving multiple camera array, we can synthesize virtual viewpoint images using conventional linear interpolation method.

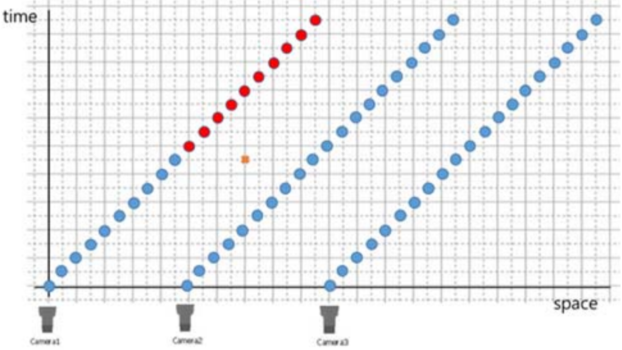


Fig. 3. Candidates for reference

After virtual view synthesized, our proposed method find the reference image to improve quality of synthesized view. In the conventional method, synthesized views were made by left and right images information. In our method, we use left, right and reference images for virtual viewpoint image generation. Reference images means most similar view with virtual viewpoint images in the different times.

As shown in Figure 3, candidates (red dots) for reference image are located in the different times. If the synthesized image(orange dot) and synthesized image are located in the same space, they will have small color differences. So, we can choose the reference image using the color differences. Fig. 4 illustrates color differences between candidates and synthesized image. In this case, we choose the smallest color differences image (red box).

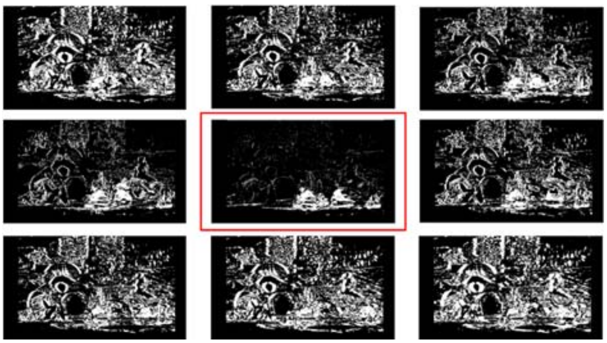


Fig. 4. Color differences between candidates and synthesized image

In the reference image, position of the scenes are same. It means we can use the information of reference image to improve quality of synthesized images. Note that the moving object has different positions between reference image and synthesized image. Therefore, we should divide moving and static areas in images. In the previous step, we calculate the color differences between reference image and synthesized image. This information represent moving object area. So moving object mask can generated by color differences between reference image and synthesized image. As shown in Figure 5, mask of color differences have some noise. The noise in the generated mask is removed by dilation and erosion, so that a better mask is obtained for the moving and static areas. Finally, given the mask of the moving and static areas, the moving areas are synthesized by using left and right images, which include blending and conventional inpainting. Virtual view in the location of the static area is generated by reference image.

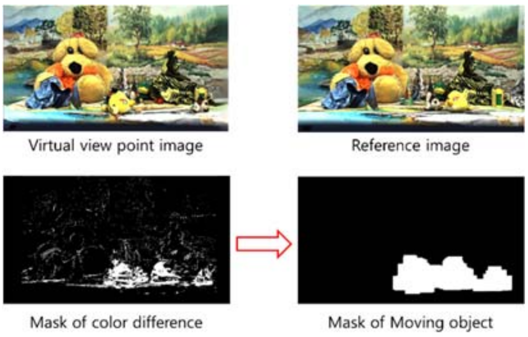
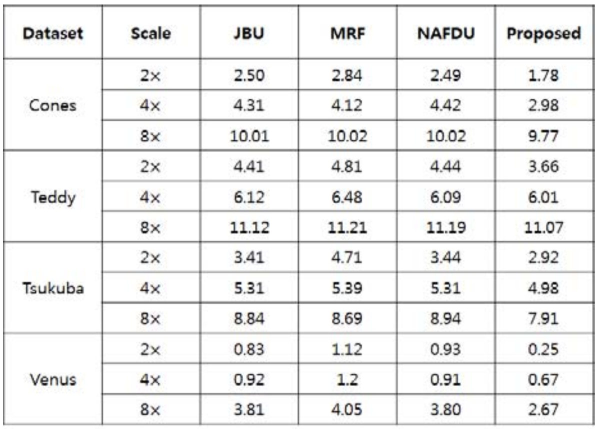


Fig. 5. Mask of moving object

III. Experimental Results

In order to evaluate our system, we generate a moving camera array sequence using stop motion animation. We captured 20 frames by moving the object gradually, while the camera array is shifted 1 cm per frame and distance between camera positions were set 10cm. For an objective evaluation of the proposed multilateral filtering, proposed method is compared with the joint bilateral upsampling (JBU), the noise-aware filter for depth upsampling (NAFDU) and the MRF-based depth upsampling. For an objective evaluation of the depth maps, the bad pixel rate (BPR), whose absolute difference is greater than 1, was used.

TABLE I. Performance Comparison of Multilateral Filtering



Dataset	Scale	JBU	MRF	NAFDU	Proposed
Cones	2×	2.50	2.84	2.49	1.78
	4×	4.31	4.12	4.42	2.98
	8×	10.01	10.02	10.02	9.77
Teddy	2×	4.41	4.81	4.44	3.66
	4×	6.12	6.48	6.09	6.01
	8×	11.12	11.21	11.19	11.07
Tsukuba	2×	3.41	4.71	3.44	2.92
	4×	5.31	5.39	5.31	4.98
	8×	8.84	8.69	8.94	7.91
Venus	2×	0.83	1.12	0.93	0.25
	4×	0.92	1.2	0.91	0.67
	8×	3.81	4.05	3.80	2.67

Table 1 shows the result of the BPR(%) comparison. According to the results, the proposed multilateral outperforms the conventional algorithms in terms of the BPR for all the scaling factors and all the tested images.

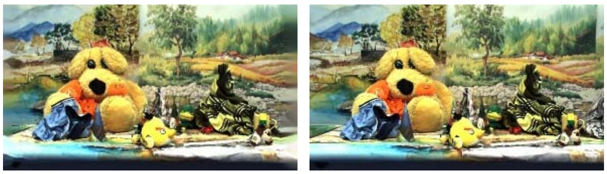
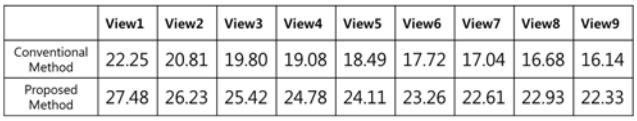


Fig. 6. Result of synthesized images

Figure 6 represents the virtual viewpoint image using the conventional method and the proposed method respectively. Synthesized image using conventional method have boundary noise around structures and background. To reduce these problems, we use left, right and reference image for virtual viewpoint image generation. Reference image means most similar view with virtual viewpoint images in the different times. After applying the proposed method the quality of static areas has been improved.

Also Table 2 shows comparison of PSNR values of the 9 virtual viewpoint images. The term peak signal-to-noise ratio (PSNR) is an expression for the ratio between the maximum possible value (power) of a signal and the power of distorting noise that affects the quality of its representation. The proposed method improved quality of the synthesized images by 5.67 dB.

TABLE II. PSNR Comparison of Synthesized Images



	View1	View2	View3	View4	View5	View6	View7	View8	View9
Conventional Method	22.25	20.81	19.80	19.08	18.49	17.72	17.04	16.68	16.14
Proposed Method	27.48	26.23	25.42	24.78	24.11	23.26	22.61	22.93	22.33

IV. Conclusion

In an ideal parallel fixed camera array, cameras are located on a certain line which is called the baseline. Also, all the cameras have the same distance to the adjacent cameras. In this case, range of virtual viewpoint image generation is limited. To solve this problem, our proposed method using the moving multiple camera array. We apply the semi global matching and multilateral filtering to get the high quality disparity map and synthesis virtual viewpoint images using both spatial and temporal information. As shown in the experimental results, we can obtain a high quality of the virtual viewpoint images for moving camera array. For most parts of the comparison, our proposed method showed the better results with the results of the conventional methods.

V. Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(2011- 0030079)

VI. References

[1] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in Proc. SPIE Conf. Stereoscopic Displays and Virtual Realistic Systems XI, vol. 5291, pp. 93-104, 2004.

[2] T. Yokoi, N. Fukushima, and T. Yendo, “Novel view synthesis for dynamic scene using moving multi-camera array,” in Proc. SPIE Conf. Stereoscopic Displays and Virtual Realistic Systems XXII, vol. 7863, pp. 24-30, 2011.

[3] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, “A Noise-aware Filter for Real-time Depth Upsampling,” ECCV Workshop on Multi-camera and Multimodal Sensor Fusion Algorithms and Applications, pp. 1–12, Oct. 2008.

[4] H. Hirschmueller, “Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information,” in Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 807-814, 2005.

[5] S. M. Hong and Y. S. Ho, “Adaptive Weighted Upsampling of Depth Map using Edge Information,” in Proc. Korean Institute Smart Media vol. 4, no. 1, pp. 146-149, 2015