Analysis of Multi-view Generation from Stereoscopic Images using the Depth Map

Do-Young Kim, Woo-Seok Jang, and Yo-Sung Ho
Gwangju Institute of Science and Technology (GIST)
261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, Korea
E-mail: {kimdo, jws, hoyo}@gist.ac.kr

***Abstract* — In this work, we generate multi-view images from stereoscopic texture images. After we estimate a disparity map using a stereo matching method, we detect an occlusion area in the acquired disparity map and refine the initial disparity values of the occlusion area. The refined disparity map is enhanced by a joint bilateral filter for boundary matching to improve the quality of synthesized multi-view images. Then, with the 2-view video-plus-disparity images, a virtual view image is generated by two different methods for intermediate view synthesis; VSRS3.5 and VSRS-1D-fast renderer. We also compare sequential and hierarchical structures for view synthesis. Experimental results show the characteristics of the view synthesis tools and structures.**

Keywords*— *auto-stereoscopic, virtual view image, disparity map, occlusion detection, hierarchical structure.

I. INTRODUCTION

Recently 3D video services have attracted a lot of attention due to the success of several 3D films, such as 'Avatar' [1]. We can feel 3D experiences by seeing slightly different images using the left and right eyes. 3D video provides a more natural and realistic perception to users through 3D displays, such as auto-stereoscopic displays and free-view point TVs [2][3].

Multi-view images can be captured by multiple cameras [4]. However, it is difficult to send all the data through bandwidth limited transmission channels. Thus, we can send only two or three video-plus-depth images and generate intermediate view images using the reconstructed information at the decoder side.

Figure 1. Texture image and its corresponding depth map

As shown in Fig. 1, the depth map is an 8-bit gray-scale image containing the distance information between the camera and each pixel. To obtain the depth map, we can use different approaches: passive or active sensor methods, and hybrid sensor fusion methods. Most passive sensor methods obtain the disparity information by applying the stereo matching operation into two-view texture images [5]. Active sensor methods employ physical sensor systems, such as depth cameras, to measure the depth information directly [6][7]. Hybrid fusion methods combine advantages of passive and active sensor methods to obtain more accurate depth maps [8].

Since the disparity information is directly related to the depth map, we can calculate the depth map from the disparity or vice versa. Stereo matching is the most popular operation to obtain the disparity map from the stereoscopic images. In order to acquire the disparity information of each pixel, we need to find the corresponding points in both images, assuming that the corresponding points exist in the same horizontal line of the two images. Conventionally, in the depth map, a closer pixel to the camera has a higher depth value.

In this paper, we investigate several problems in generating multi-view images from the stereoscopic texture images. After we estimate the disparity map using stereo matching, we convert it to the depth map. Since the stereo matching method does not provide accurate disparity values for all the pixels, we need to apply a post-processing operation using the joint bilateral filter to enhance the disparity map. Then, we can generate virtual view images at intermediate view positions using the texture and depth images. Moreover, we explore two different structures for virtual view synthesis and compare the performance of two different view synthesis tools: VSRS 3.5 and VSRS 1D fast renderer.

II. DISPARITY ESTIMATION

A. *Initial Disparity Map Acquisition*

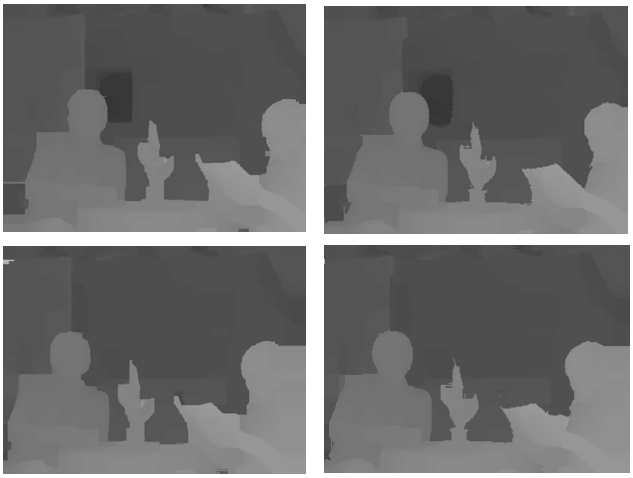
In order to obtain the depth map, we need to generate the disparity map. Using the stereo matching method, we can estimate the disparity information by finding the corresponding points in the stereoscopic texture images. There are generally two different methods: local and global methods. While the local method is inaccurate, the global method is complicated. In this paper, we extract the initial disparity map from the left and right texture images using the constant-space belief propagation (CSBP) method [9], a global method with reduced complexity of belief propagation considerably for practical use.

B. *Occlusion Handling*

Occlusion is a very difficult problem to handle in stereo matching. In order to obtain a more accurate disparity map, we detect the occlusion area in the initial disparity map using the

THIS RESEARCH IS SUPPORTED BY MCST AND KOCCA IN THE CULTURE TECHNOLOGY RESEARCH & DEVELOPMENT PROGRAM 2012.

cross check and warping constraints. Then, pixel values of the detected occlusion areas are filled with the background values using the potential energy function based on the difference of distance and color information [10]. At that time, we use only those pixel values that are not in the occlusion area. In other words, we do not use the initial disparity values in those areas because the pixel values are erroneous. Fig. 2(a) shows the results of the stereo matching method after occlusion handling.



(a) Before refinement (b) After refinement

Figure 2. Acquired disparity maps of *Newspaper*

III. DISPARITY REFINEMENT

The boundary of the acquired disparity map is not matched well with that of the corresponding texture image. The problem degrades the quality of the synthesized image. Thus, we employ a joint bilateral filter (JBF) to solve the unmatched boundary problem. Formally, the disparity value *D(x,y)* at the position *(x, y)* is computed by JBF as follows:

$$D(x,y) = \arg\min_{d \in \vec{d}_p} \frac{\sum_u \sum_v W(u,v) \cdot C(u,v)}{\sum_u \sum_v W(u,v)} \quad (1)$$

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

$$C(u,v,d) = \min(\lambda, |D(u,v) - d|) \quad (3)$$

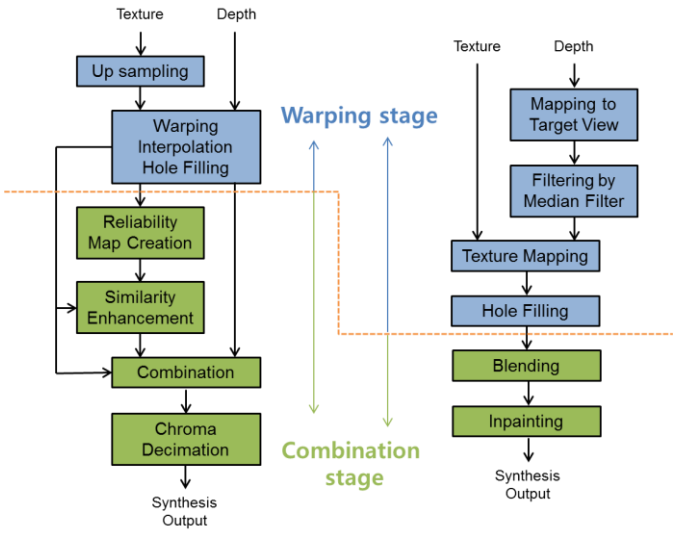
where *W(u,v)* consists of intensity and spatial weighting functions that are usually modeled by the Gaussian function, and σ_R and r are smoothing parameters of intensity and spatial weighting functions, respectively. Eq. (3) is a truncated linear model to allow depth discontinuities and λ is a constant to reject outliers.

In Eq. (1), the center pixel is substituted by one of $\vec{d}_p$ = {*D(x-1,y), D(x,y-1), D(x, y), D(x+1,y), D(x,y+1)*} among its neighboring pixels. As shown in Fig. 1, we can enhance the depth image and obtain a refined edge map by applying JBF to the decoded depth image.

IV. VIRTUAL VIEW SYNTHESIS

A. View Synthesis Algorithms

In this paper, we used two different algorithms for virtual view generation. While the VSRS-1D-fast algorithm is based on HEVC, the VSRS3.5 algorithm was developed during the MPEG 3DV exploration experiment [11].



(a) VSRS-1D-fast (b) VSRS3.5

Figure 3.Block diagram of view synthesis methods

Fig. 3(a) depicts the VSRS-1D-fast method that employs the interpolation of the synthesized view from the left and right texture images with the corresponding depth maps. After the two texture images are extrapolated from the left and the right views to the virtual view, we create a reliability map for view combination. The similarity of two texture images is also enhanced before combining them to synthesized output images. Using the reliability map, the multi-view images are generated. In this method, we only use the disparity information in the horizontal direction.

The VSRS3.5 method takes two video-plus-depth images to generate virtual view images using the intrinsic and extrinsic camera parameters. In the general mode, virtual views are produced by the 3D warping operation [12].

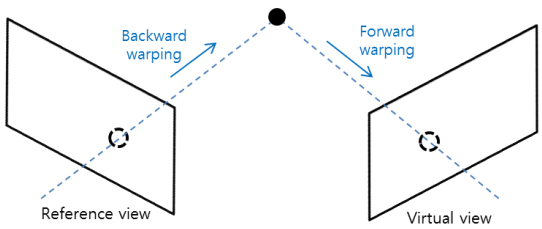


Figure 4. Principle of 3D warping technique

As shown in Fig. 4, this process includes two separated steps. After the reference view is projected into the 3D world space using corresponding reference depth map and Eq. (4), the 3D space points are projected into the virtual view image plane using Eq. (5). The intrinsic matrix is *A* and the extrinsic matrix consists of the rotation matrix *R* and the translation vector *T*.

$$[u,v,w]^T = R_D \cdot A_D^{-1} \cdot [x,y,1]^T \cdot D_s(x,y) + T_D \quad (4)$$

$$[x',y',z']^T = A_C \cdot R_C^{-1} \cdot [u,v,w]^T - T_C \quad (5)$$

The steps of the VSRS3.5 method, shown in Fig. 3(b), are briefly described below.

STEP 1: Two depth maps are mapped to the virtual view and the reference texture images are warped to the target view using the 3D warping operation.

STEP 2: Hole areas in the warped texture image, which are caused by occlusion, are filled by the same position pixels from the other warped texture image.

STEP 3: Two virtual images are blended with a weighting function based on the baseline distance. The closer virtual view to the reference view is assigned a higher weight.

STEP 4: Any remaining holes after blending are filled by an in-painting algorithm using a binary mask.

View Synthesis Structure

Virtual synthesis images are generated using the reference images. Depending on the usage of the reference images, there are two different structures for intermediate view synthesis.

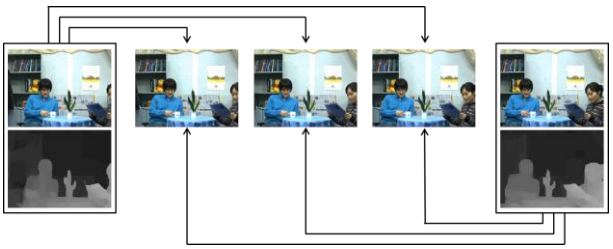


Figure 5. Sequential view synthesis structure

As shown in Fig. 5, the sequential view synthesis structure is very common and general. In this structure, only the left and right input images are employed as the reference views. Intermediate images are generated by a weighted sum of the reference images using Eq. (6). The weighting factor is calculated by the distance from the reference images.

$$View_{syn} = \alpha \cdot View_{left} + (1-\alpha) \cdot View_{right} \quad (6)$$

Fig. 2 shows the hierarchical view synthesis structure. In this method, the generated intermediate view image is used as the reference view image. In order to generate the intermediate view image, we need an intermediate depth image at the view point. A depth image is generated by the same process as the virtual texture image. Hole areas are filled by the other depth image and the remaining common hole areas are covered by the in-painting algorithm.

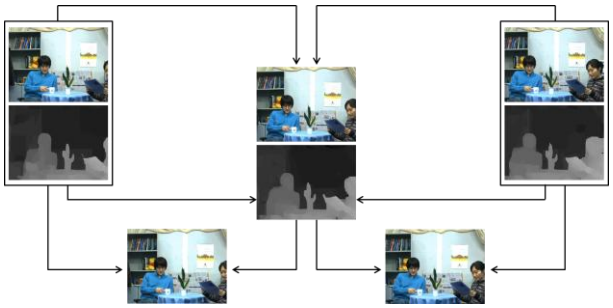


Figure 6. Hierarchical view synthesis structure

V. EXPERIMENTAL RESULTS

In order to analyze the performance and effect at each stage, we extract two kinds of synthesis images and compare them. As shown in Fig. (3), intermediate results from the warping stage are obtained. In the VSRS-1D-fast algorithm, the reference view is reprojected to the target view using horizontal pixel shifting and the hole area is temporarily filled with background values. In the VSRS3.5 algorithm, the target view is generated using pixel-by-pixel mapping based on the 3D warping operation from the reference view. Originally, the hole filling method in the VSRS3.5 algorithm refers to the other reference view; however, we utilize an in-painting algorithm for hole filling in this paper. Then, final results at the combination stage are compared.

In order to evaluate the quality of intermediate views and coding error compensation of each view synthesis algorithm, we have tested with the MPEG 3DV test sequences.

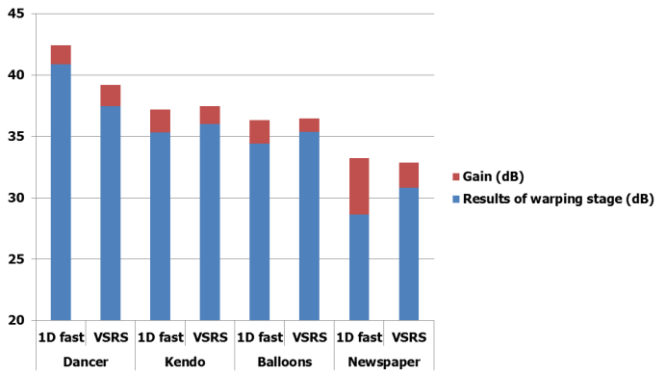


Figure 7. Performance comparison of view synthesis

Fig. 7 shows the performance of view synthesis methods in terms of PSNR values between the original and synthesized images using the original depth data. Although the VSRS-1D-fast (1D-fast) method had lower PSNR values than the VSRS3.5 (VSRS) method, the PSNR values were significantly improved after the combination stage. Thus, VSRS-1D-fast had more gains than VSRS3.5 by 2.48 dB for the *Newspaper* sequence, as shown in Fig. 2. This implies that VSRS-1D-fast

generates intermediate views that are more similar to the original view image.

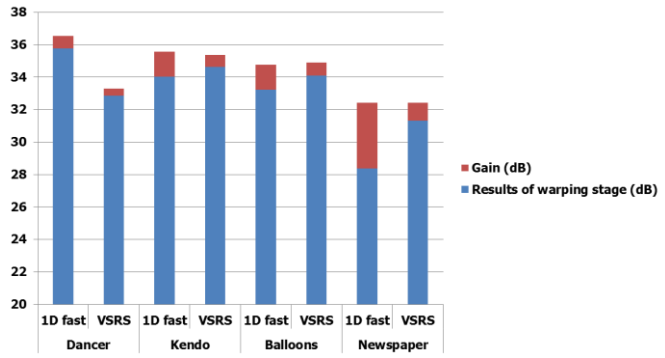


Figure 8. Comparison of coding error compensation

Fig. 8 shows the results for coding error compensation. In this experiment, the original and synthesized view images using the encoded depth data were compared in terms of PSNR values. Four test sequences were encoded with four different QP values: 25, 30, 35, and 40. We averaged the PSNR values of each view. In the warping stage, VSRS-1D-fast had lower PSNR values than VSRS3.5 because VSRS-1D-fast extended background pixels along the line for the dis-occluded area while VSRS3.5 utilized the in-painting algorithm. However, as shown in Fig. 8, VSRS-1D-fast compensated the coding error effectively at the combination stage using similarity enhancement and blending according to the reliability map. Final results of VSRS-1D-fast had higher PSNR values than VSRS3.5 by 3.27 dB for the *Dancer* sequence.

Table 1. Comparison of view synthesis structure

Sequence	PSNR (dB)	
	sequential	hierarchical
Dancer	35.71	37.30
Kendo	35.51	34.88
Balloons	35.62	34.97
Newspaper	30.77	29.11

We have also investigated two different structures for virtual view synthesis: sequential and hierarchical structures using the VSRS3.5 algorithm. We have tested them with the same test sequences.

Table 1 compares performances of the two view synthesis structures in terms of average PSNR values. In general, results from the sequential view synthesis structure had higher PSNR values than the hierarchical view synthesis structure by 0.6 dB on average. On the other hand, the result of the *Dancer* sequence is different from the others. Because the sequences acquired by the general camera have more inaccurate information than the sequences obtained by the computer graphics (CG), the error propagation problem is occurred during the process of intermediate depth map synthesis.

VI. CONCLUSION

In this paper, we compared VSRS-1D-fast and VSRS3.5 view synthesis algorithms to find better synthesis performance for multi-view image generation from the stereoscopic images. Even if the VSRS3.5 algorithm produced better results at the warping stage mostly, the VSRS-1D-fast algorithm compensated for coding errors of depth data and improved the quality of synthesized view images at the combination stage using similarity enhancement and blending according to the reliability map. Consequently, the VSRS-1D-fast algorithm was better than the VSRS3.5 algorithm. Besides, the sequential view synthesis structure in the VSRS3.5 was better than the hierarchical view synthesis structure because of the error propagation problem during the process of intermediate depth map synthesis.

REFERENCES

[1] A. Smolic, and D. McCutchen, "3DAV exploration of video-based rendering technology in MPEG," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 348-356, March 2004.

[2] ISO/IEC JTC1/SC29/WG11, "Applications and requirements on FTV," N9466, Oct. 2007.

[3] M. Tanimoto, "Overview of free viewpoint television," *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454-461, July 2006.

[4] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multi-View Imaging and 3DTV (Special Issue Overview and Introduction)," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 10-21, Nov. 2007.

[5] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7-42, April 2002.

[6] S. B. Gokturk, H. Yalcin, and C. Bamji, "A Time-Of-Flight Depth Sensor - System Description, Issues and Solutions," *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pp. 35-35, June 2004.

[7] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 195-202, June 2003.

[8] E. Lee and Y. Ho, "Generation of high-quality depth maps using hybrid camera system for 3-D video," *Journal of Visual Communication and Image Representation*, vol. 22, no. 1, pp. 73-84, Jan. 2011.

[9] Q. Yang, L. Wang, and N. Ahuja, "A constant-space belief propagation algorithm for stereo matching," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1458-1465, June 2010.

[10] W. Jang and Y. Ho, "Efficient disparity map estimation using occlusion handling for various 3D multimedia applications," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 4, pp. 1937-1943, Nov. 2011.

[11] ISO/IEC JTC1/SC29/WG11, "Test Model under Consideration for HEVC based 3D video coding v3.0," N12744, April 2012.

[12] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *SPIE Stereoscopic Displays and Virtual Reality Systems XI*, pp. 93-104, May 2004.