High-resolution Depth Map Generator for 3D Video Applications Using Time-of-flight Cameras

Yunseok Song, *Student Member, IEEE* and Yo-Sung Ho, *Fellow, IEEE*

***Abstract*—3D video applications are easily accessible to consumers in this era. In general, high quality depth maps are crucial for interpolating virtual view images that are vital in 3D video applications. Many applications adopt time-of-flight (ToF) cameras due to their real-time distance capturing. Yet, the raw ToF image exhibit limitations such as inaccurate distance values in object boundaries and areas that can absorb the infrared ray. This paper presents a method that handles this issue. The existing methods do not focus on this problem. The proposed method is tested on a camera system that consists of a color camera and a ToF camera. Assuming immersive applications such as teleconferencing or virtual broadcasting, the object of interest is captured. The errors existing in the ToF images are modified as a preprocessing step. Afterward, the low-resolution ToF image is warped to the viewpoint of the color camera; empty areas are filled considering the neighbor information. The depth map results show improvements over the state-of-the-art method. The proposed method can increase the overall quality of the 3D video system, ultimately making 3D video consumer applications more marketable.**

***Index Terms*—3D video system, Time-of-Flight (ToF), Depth map, 3D warping.**

I. Introduction

3D video technologies have been improved significantly in the last decade. With the development of 3D displays and high-end televisions, the efficiencies of 3D and multi-view video applications are expected to enhance as well [1], [2]. For 3D image processing, depth images are crucial; they are mainly used for interpolating virtual view images. In the 3D video system, as shown in Fig. 1, depth data are acquired directly from depth cameras or from depth estimation using multi-view cameras [3]. Yet, various factors cause inaccurate depth maps. Depth cameras may induce flickering and mixed pixels while depth estimation is prone to false estimation around object boundaries [4], [5]. Furthermore, distortion occurs as a result of compression [6], [7].

Manuscript received October 1, 2017; accepted November 15, 2017. Date of publication December 19, 2017. This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government(MSIP)(No. 2011-0030079). *(Corresponding author: Y. Song.)*

Y. Song and Y.S. Ho are with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea (e-mail: ysong@gist.ac.kr and hoyo@gist.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCE.2017.015096

As depicted in Fig. 1, the 3D video system transmits compressed N views of color and depth video data. At the receiver end, M views are generated based on the N decoded views and synthesized views. Hence, the number of output views is always greater than that of input views [8]. For view synthesis, decoded color and depth video data are used as input in depth image based rendering. Thus, the quality of decoded color and depth data directly affects the quality of rendered images.

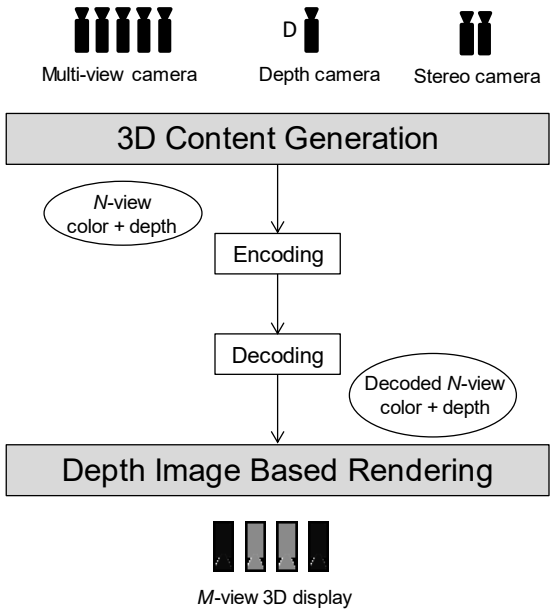


Fig. 1. Framework of a 3D video system including 3D content production and depth image based rendering.

Depth images present the distance between the camera and the object. Generally, depth images are produced by depth cameras or estimated by stereo matching. Depth cameras allow fast data acquisition but cost can be an issue. In addition, interference must be checked which is caused by frequency overlaps. Stereo matching does not have limitations of depth cameras [9], [10]; however, it can be time-consuming, thus, not suitable for practical applications.

This paper presents a depth map generation system that consists of a color camera and a time-of-flight (ToF) camera. The goal is to generate high quality depth maps that correspond to the color camera view. This system can be effective in various multi-view video applications that use synthetic background. Using the system in question, an object in interest

is captured in front of a chromakey screen. 3D image warping and hole filling are conventional tools. This paper adds a novel method that improves the quality of the raw ToF image. This enhances the quality of the resulting depth map, especially in the object region.

The remainder of this paper is organized as follows. Section II describes the conventional framework of depth map generation by means of ToF images [11]. In Section III, the distance error reduction technique for ToF images is explained, which is the novel contribution. Section IV analyzes the results and Section V concludes the paper.

II. Generation of Depth Map at Color Camera View

A. 3D Image Warping

ToF cameras are useful for obtaining object distances in a scene [12]. Yet, they can produce low resolution images only. In order to match the resolution of depth images and color images, ToF-to-color view 3D warping is applied using ToF and color camera parameters. Intrinsic and extrinsic camera parameters are necessary for this process [13]. Intrinsic camera parameters include focal length, principal point, and skew coefficients. In addition, extrinsic camera parameters represent rotation and translation characteristics of the camera.

In the ToF image, each pixel is projected to a 3D point, i.e., world coordinate, then this 3D point is projected to the destination image. For ToF-to-color 3D warping, the ToF camera and the color camera are source and destination, respectively. The 2D image coordinate at the destination image is acquired by the 2D image point at the source image and its intrinsic and extrinsic camera parameters.

B. Filling of Empty Areas in the Warped Image

From depth image warping, the depth image corresponding to the color camera view is acquired. These images contain empty pixels due to the resolution difference between ToF and color cameras. Joint bilateral filter (JBF) is used to fill such areas [15]. The filter is applied to the object region only, assuming the approximate depth value range of the object is known. JBF is an extension of the bilateral filter [14], which is widely used for edge preservation.

$$D(x,y)=\frac{\sum_u\sum_v W(u,v)\cdot D_i(x,y)}{\sum_u\sum_v W(u,v)} \tag{1}$$

In (1), $D_i(x, y)$ and $D(x, y)$ denote the value at (x, y) coordinate in the warped depth image and the final depth image which is to be filled, respectively. $D_i(x, y)$ is available as a result of ToF-to-color warping. (u, v) is the neighbor coordinate of (x, y). r represents the kernel size. In the experiments, the kernel size r is 11. W represents the weight, which is zero if the pixel value in the warped depth image is zero. Otherwise, the weight is a multiple of spatial weight $f(u, v)$ and range weight $g(u, v)$. This is represented in (2).

$$W(u,v)=\begin{cases}0 & ,\text{if } D_i(x,y)=0\\ f(u,v)\cdot g(u,v) & ,\text{otherwise}\end{cases} \tag{2}$$

The spatial weight is based on the intensity difference. When computing the intensity difference, JBF uses the color image while the bilateral filter uses the depth image itself. JBF produces more reliable spatial weights since color data difference can be more specific. The range weight is the same in both JBF and the bilateral filter. These weights are explained by (3) and (4). σ_f and σ_g are sigma values for the spatial and range weights which are Gaussian values; their values are 2 and 8, respectively, in the experiments.

$$f(u,v)=\exp\left\{-\frac{|I(x,y)-I(u,v)|^2}{2\sigma_f^{\,2}}\right\} \tag{3}$$

$$g(u,v)=\exp\left\{-\frac{(x-u)^2+(y-v)^2}{2\sigma_g^{\,2}}\right\} \tag{4}$$

III. Object Distance Correction in ToF Images

The proposed method replaces erroneous values in the ToF image with newly generated values using neighbor values which possess valid distance data. Three steps are employed: object boundary filtering, iterative outlier elimination, and iterative min/max averaging. The distance data of the foreground object is altered as a result. This technique can be implemented on static camera systems. Video applications that require real-time accurate distance sensing on various surface types would benefit from the proposed method.

A. Object Boundary Filtering

In the original ToF image, an intermediate distance value is stored on the boundary between foreground and background. In reality, nothing exists at this distance. This error causes problems if other applications use this distance data. To remove this intermediate value, filtering is performed using a 1-tap 3-pixel wide window. First, it is checked whether the pixel is boundary or not. If the pixel does not belong to background and one of its neighbors is in the background region, this pixel is determined as a boundary value. Since the background is still, the background distance is known.

The boundary value is replaced by the neighbor which belongs to the foreground. The neighbor with the smaller distance is the neighbor that belongs to foreground. Fig. 2 depicts this process in the horizontal direction. The same procedure is executed in the vertical direction as well. Fig. 3 shows the effect of object boundary filtering with red circles emphasizing noticeable changes. Since the ToF images are 16-bit data, they are shown in 8-bit images with maximized contrast for display purpose in this paper. In real applications, the 16-bit distance data would be employed without modification.

B. Iterative Outlier Elimination

After object boundary filtering, there still exist distance values that should be regarded as outliers in the foreground object. These outlier distance values are either greater than the maximum distance or less than minimum distance. The rough values of maximum and minimum object distances are

determined when capturing. The outliers are caused by inaccurate distance sensing. In particular, black objects such as hair tend to absorb the infrared ray.

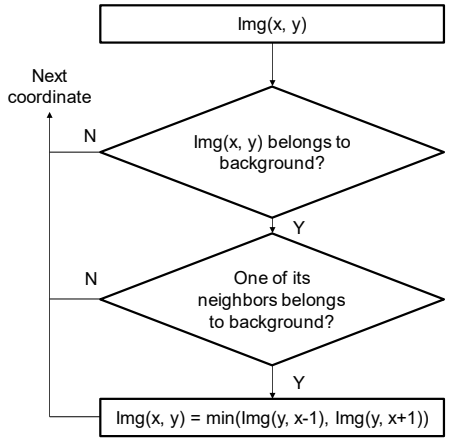
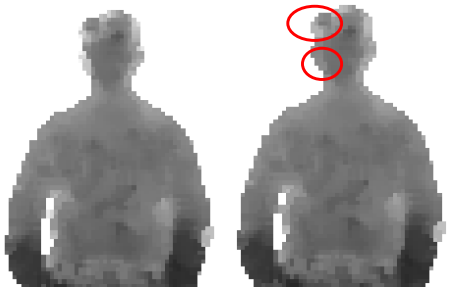


Fig. 2. Flowchart of object boundary filtering in the horizontal direction.

(a) Before boundary filtering (b) After boundary filtering

Fig. 3. Result of object boundary filtering.



(a) Before outlier elimination (b) After outlier elimination

Fig. 4. Result of iterative outlier elimination.

C. Iterative Min/Max Averaging

The objective of the final step is to smooth out the abrupt distance changes within the foreground. For this, two images are generated. First, using a 3×3 window, the center is replaced by its neighbor that possesses the highest distance value. This is defined to be a maximum neighbor filter. Similarly, a minimum neighbor filter returns the smallest neighbor value. The average of these two images show smoothed effects. This is beneficial for reducing sudden distance increase or decrease which usually occurs due to inaccurate depth sensing. Such a method is repeated since a single execution will not show much difference. However, too many executions will exhibit enormous blurriness in the image. In the experiments, five iterations are carried out for this step. Fig. 5 displays the averaging of minimum and maximum neighbor filtered images. As noted in the second step, similarly in this step, a larger window can be used; however, this can lead to losing details within the object.

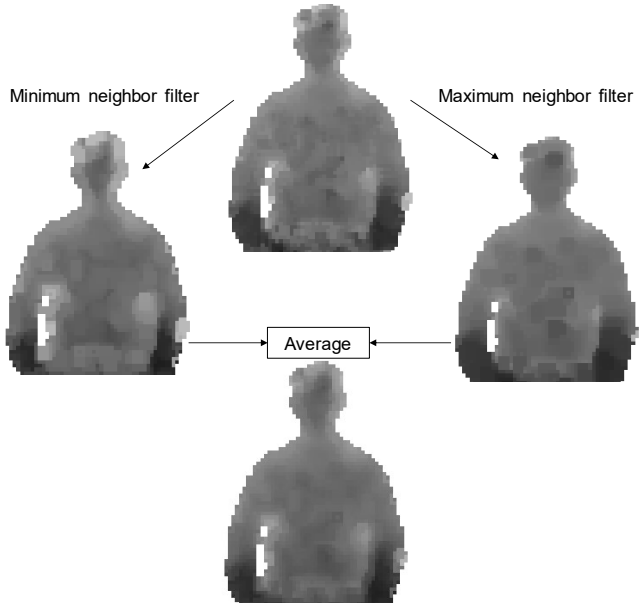
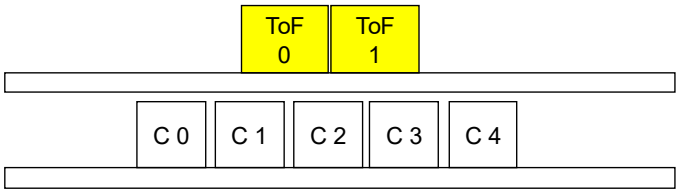


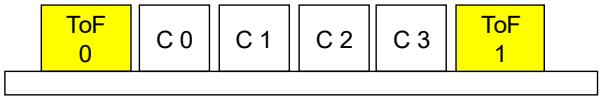
Fig. 5. Averaging of minimum and maximum neighbor filtered images.

D. Implementation on Camera Systems

The proposed method requires at least a color camera and a ToF camera. When incorporating it to a traditional multi-view camera setup, a rig for ToF cameras may or may not be needed. Fig. 6(a) shows a setup where ToF cameras are placed above color cameras using an additional rig for ToF cameras. In Fig. 6(b), color cameras are positioned in between ToF cameras; thus, additional rig is not required which can help reduce cost.



(a) ToF cameras location above color cameras



(b) Color cameras positioned in between ToF cameras

Fig. 6. Types of camera system setups that can incorporate the proposed object distance correction.

E. Relevance to Consumer Electronics

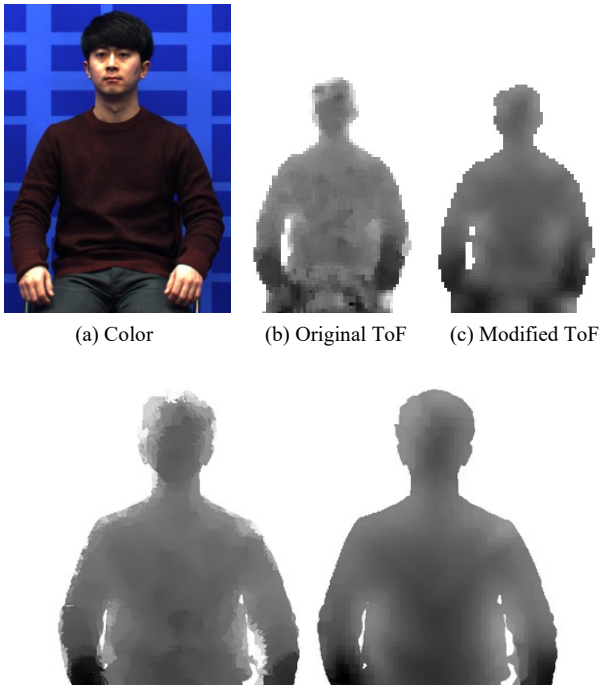
With regards to consumer electronics, the proposed method can be applied to multimedia applications that can equip ToF cameras, e.g., teleconferencing and virtual broadcasting. In these types of systems, the ToF camera can be placed adjacent to the color camera. The ultimate goal is to represent the scene

in 3D by using the depth map consisting of distance data. While the conventional methods are susceptible to inaccurate distance values in hair regions and boundaries, the proposed method is able to reduce the errors and improve the quality of depth map. Thus, more realistic 3D video service can be achieved, which can attract consumers.

IV. Experiment Results

Experiments were conducted on four image sets: "Sitting", "Bag grab", "Reader", and "Holding". The chromakey blue screen is 3.5 meters away from ToF and color cameras. The image resolution of ToF and color images are 176×144 and 1280×720, respectively. The ToF camera is positioned horizontally adjacent to the color camera as in Fig. 6(b). An external synchronization module guarantees the synchronization between the ToF camera and color cameras.

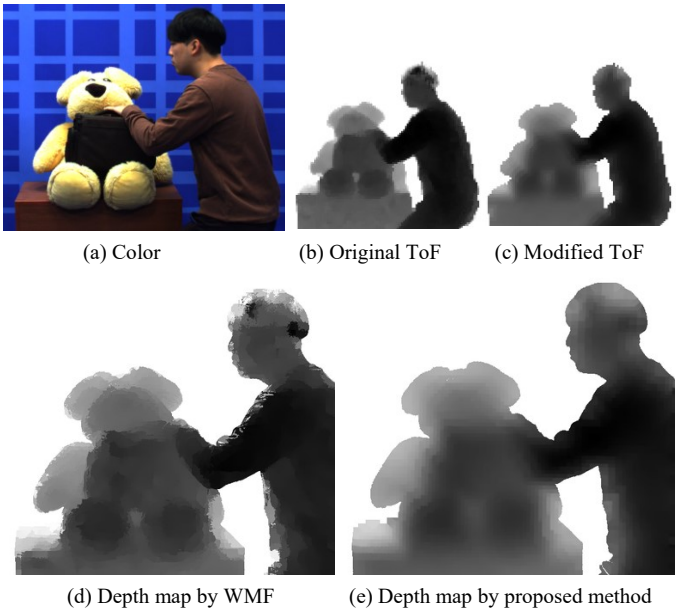
The depth maps generated at color camera view are 16-bit data since 16-bit ToF images are used in the process. Fig. 6, Fig. 7, Fig. 8, and Fig. 9 exhibit the original color image, original ToF image, modified ToF image and high-resolution depth map results. For these figures, color and ToF images are captured at C0 and ToF0 positions from Fig. 6(b). The generated depth map results correspond to C 0. The figures show only the foreground objects. We compare the results by proposed method with those generated by the weighted mode filter (WMF) [15]. WMF employs a joint histogram and optimizes the solution by L_1 norm minimization. This filter is used for depth video filtering as well as ToF image based depth map generation.



(a) Color (b) Original ToF (c) Modified ToF

(d) Depth map by WMF (e) Depth map by proposed method

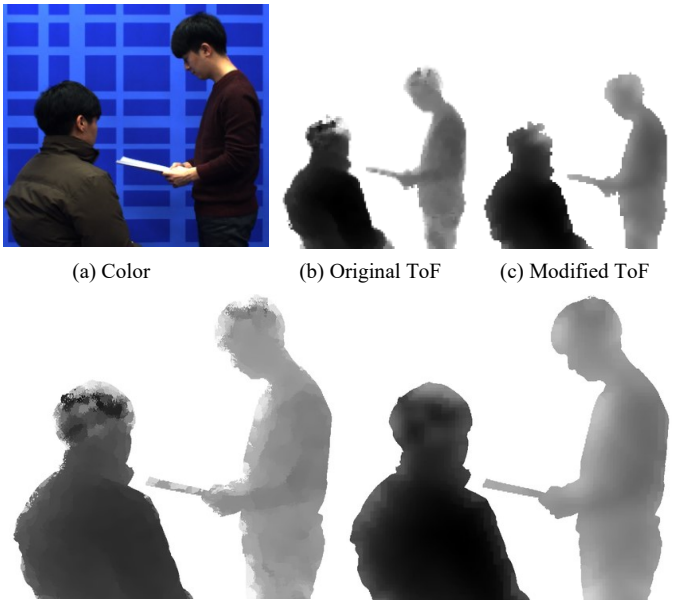
Fig. 7. ToF modification and its application to depth map generation at color view. Results on "Sitting".



(a) Color (b) Original ToF (c) Modified ToF

(d) Depth map by WMF (e) Depth map by proposed method

Fig. 8. ToF modification and its application to depth map generation at color view. Results on "Bag grab".

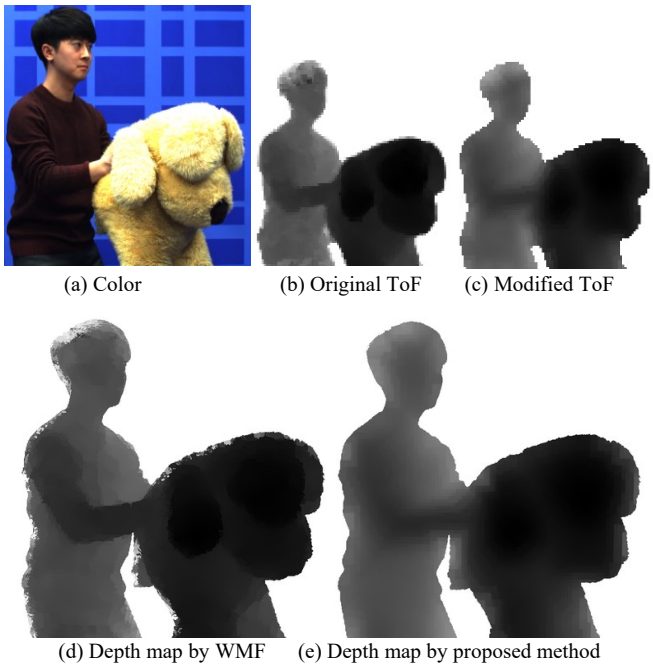


(a) Color (b) Original ToF (c) Modified ToF

(d) Depth map by WMF (e) Depth map by proposed method

Fig. 9. ToF modification and its application to depth map generation at color view. Results on "Reader".

Pixels that possess darker values are closer to the camera than those that have brighter values. When comparing the ToF images, the modified ToF images show that lost data in the original ToF are filled with reasonable values. The available neighbor data is taken into account. This is particularly noticeable in hair regions. Furthermore, errors near the boundaries are removed. These effects can also be seen in the high resolution depth maps. Comparing with the results by WMF, the results by proposed method show more consistency and less variation. Since errors in the low resolution image propagate to the high resolution depth map results, error reduction of ToF images as a preprocessing step is very advantageous.



(a) Color (b) Original ToF (c) Modified ToF

(d) Depth map by WMF (e) Depth map by proposed method

Fig. 10. ToF modification and its application to depth map generation at color view. Results on "Holding".

For objective evaluation of the depth map, we synthesized a virtual view at C1. The synthetic image is obtained by warping the color image at C0 to C1 position. Then, the synthetic image is compared with the image captured at C1 position camera. Similarly, evaluation of generated depth at C1 is carried out by comparing synthetic C2 and original C2. When computing the PSNR, only the foreground region was concerned. Table I and Table II represent objective evaluation of generated depth at C0 and C1, respectively. For depth at C1, the proposed method showed an improvement of 0.51 dB on average. In addition, the average gain is 0.43 dB in case of depth at C1 evaluation. Thus, the effectiveness of the proposed method is confirmed.

TABLE I
PSNR OF SYNTHETIC C1 AND ORIGINAL C1 (EVALUATION OF DEPTH AT C0)

Image	WMF [15]	Proposed Method
Sitting	26.88 dB	27.07 dB
Bag grab	25.60 dB	26.26 dB
Reader	27.82 dB	28.43 dB
Holding	22.01 dB	22.56 dB
Average	25.57 dB	26.33 dB

TABLE II
PSNR OF SYNTHETIC C2 AND ORIGINAL C2 (EVALUATION OF DEPTH AT C1)

Image	WMF [15]	Proposed Method
Sitting	27.35 dB	27.89 dB
Bag grab	25.29 dB	25.58 dB
Reader	26.07 dB	26.55 dB
Holding	21.14 dB	21.54 dB
Average	24.96 dB	25.39 dB

V. CONCLUSION

This paper presented a depth map generation system using a color camera and a ToF camera. The generated depth map represents the distance information of the object at the view of the color camera. Its resolution is the same as the color camera. The main contribution is distance error reduction for the ToF image. For this, object boundary filtering, iterative outlier elimination, and iterative min/max averaging are exploited. Consequently, the modified ToF image is warped to the color camera view and the empty areas are filled considering the neighbor information. The result depth maps appear more natural, particularly showing vast improvement in the hair region and object boundaries. The effectiveness of the proposed method is confirmed objectively as well. This ToF camera based depth map generating system is expected to be beneficial in various fixed-space 3D video consumer applications such as teleconferencing or virtual broadcasting.

REFERENCES

[1] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3-D TV," in *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XI*, 2004, vol. 5291, no. 2, pp. 93-104.

[2] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3D warping using depth information for FTV," *Signal Processing: Image Communication*, vol. 24, no. 1, pp. 65-72, Jan. 2009.

[3] Y. S. Kang and Y. S. Ho, "An efficient image rectification method for parallel multi-camera arrangement," *IEEE Trans. Consum. Electron.*, vol. 57, no. 3, pp. 1041-1048, Aug. 2011.

[4] Y. Song, C. Lee, and Y. S. Ho, "Adaptive depth boundary sharpening for effective view synthesis," in *Proc. Picture Coding Symposium*, 2012, pp. 73-76.

[5] Y. Song and Y.S. Ho, "Depth map boundary filter for enhanced view synthesis in 3D video," *Journal of Signal Processing Systems*, DOI 10.1007/s11265-016-1158-x, pp. 1-9, Aug. 2016.

[6] J. Y. Lee and H. W. Park, "Efficient synthesis-based depth map coding in AVC-compatible 3D video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 1107-1116, June 2016.

[7] B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Reducing 3D video coding complexity through more efficient disparity estimation," *IEEE Trans. Consum. Electron.*, vol. 60, no. 1, pp. 74-82, Feb. 2014.

[8] Y. J. Jung, H. Sohn, S. I. Lee, and Y. M. Ro, "Visual comfort improvement in stereoscopic 3D displays using perceptually plausible assessment metric of visual comfort," *IEEE Trans. Consum. Electron.*, vol. 60, no. 1, pp. 1-9, Feb. 2014.

[9] W. S. Jang and Y. S. Ho, "Efficient disparity map estimation using occlusion handling for various 3D multimedia applications," *IEEE Trans. Consum. Electron.*, vol. 57, no. 4, pp. 1937-1943, Nov. 2011.

[10] Y. Zhan, Y. Gu, K. Huang, C. Zhang, and K. Hu, "Accurate image-guided stereo matching with efficient matching cost and disparity refinement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1632-1645, Sept. 2016.

[11] C. Lee, H. Song, B. Choi, and Y. S. Ho, "3D scene capturing using stereoscopic cameras and a time-of-flight camera," *IEEE Trans. Consum. Electron.*, vol. 57, no. 3, pp. 1370-1376, Aug. 2011.

[12] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proc. IEEE International Conference on Computer Vision*, 1999, pp. 666-673.

[13] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics.*, vol. 26, no.3, pp. 1-5, Aug. 2007.

[14] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE International Conference on Computer Vision*, 1998, pp. 839-846.

[15] D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Trans. Image Process.*, vol. 26, no.3, pp. 1176-1190, March 2012.

Yunseok Song (S'17) received his B.S. degree in Electrical Engineering from Illinois Institute of Technology in 2008, M.S. degree in Electrical Engineering from University of Southern California in 2009. He is pursuing a Ph.D. degree in the School of Electrical Engineering and Computer Science at Gwangju Institute of Science and Technology (GIST), Korea. His research interests include 3D image processing, video processing, and realistic broadcasting.

Yo-Sung Ho (M'81–SM'06–F'17) received both B.S. and M.S. degrees in Electronic Engineering from Seoul National University, Korea, in 1981 and 1983, respectively, and Ph.D. degree in Electrical and Computer Engineering from University of California at Santa Barbara in 1990. He joined Electronics and Telecommunications Research Institute (ETRI) of Korea in 1983. From 1990 to 1993, he was with Philips Laboratories, Briarcliff Manor, New York, where he was involved in the development of the advanced digital high-definition television (AD-HDTV) system. In 1993, he rejoined the technical staff of ETRI and was involved in development of the Korea direct broadcast satellite (DBS) digital television and high-definition television systems. Since 1995, he has been with Gwangju Institute of Science and Technology (GIST), where he is currently a professor in the School of Electrical Engineering and Computer Science. His research interests include digital image and video coding, image analysis and image restoration, advanced coding techniques, digital video broadcasting, 3D television, and realistic broadcasting.