https://doi.org/10.2352/ISSN.2470-1173.2017.5.SD&A-367
© 2017, Society for Imaging Science and Technology

Real-time Depth Estimation Method Using Hybrid Camera System

Eu-Tteum Baek, Yo-Sung Ho
Gwangju Institute of Science and Technology (GIST)
123 Cheomdangwagi-ro, Buk-gu, Gwangju, 61005, South Korea
Email: {eutteum, hoyo}@gist.ac.kr

Abstract

In this paper, we present a new real-time depth estimation method using the stereo color camera and the ToF depth sensor. First, we obtain the initial depth information from the ToF depth sensor. Exploiting the initial depth information to narrow the disparity range by performing 3-D warping from the position of the ToF camera to the position of the stereo camera due to accelerating the algorithm. We construct the cost volume by calculating intensity difference and truncated absolute difference of gradients. After narrowing the disparity range, we aggregate the cost volume. Experimental results show that the proposed method can represent the disparity detail and improve the quality in the vulnerable areas of stereo matching.

Keywords: depth estimation, ToF camera, Stereo camera, stereo matching.

1. Introduction

Various stereo image-based depth estimation methods have been developed over the past several decades to obtain accurate depth information and involves many applications such as 3D movies, 3D printing, object detection, and 3D reconstruction. In general, depth information can be acquired by two approaches: active and passive sensor based depth estimation methods. Passive sensor based depth estimation methods estimate depth information by correlating the captured scenes from two or more cameras [2]. The advantages are low cost and flexible resolution. However, passive sensors acquire incorrectly calculated depth information in many types of regions, and it is too slow to use in real time. Active depth sensor acquires depth information with a physical sensor such as infrared ray (IR), laser and light pattern. These sensors emit their own light onto the scene, and derive its depth information [1]. Usually, the active depth cameras are more effective and efficient in generating high-quality depth data indoors than the passive sensors. However, they produce lower resolution images and generally require expensive devices compared to the color camera. In addition, due to the use of infrared (IR) lights for depth acquisition, there is a problem with low reflectance areas such as black areas. It also provides inaccurate depth values at object boundaries when the object is moving fast in the scene.

In order to enhance the quality of the depth map, hybrid sensor-based methods that combine video cameras and a ToF depth camera have been introduced [1, 2]. Hybrid methods integrate the active and passive methods to generate more accurate depth data and to cover their weaknesses. Recent approaches of fusing active and passive sensors have shown improvements in depth quality by making up for the weakness of each sensor method [3-5].

The objective of this paper is to obtain accurate depth information using depth and stereo images. Therefore, we propose an accurate disparity map acquisition method through stereo correspondence. We design disparity estimation systems to enhance the advantages of active and passive depth sensors and to complement weaknesses.

2. Real time rectification

The captured multi-view images by the stereo camera system have inconsistencies because the direction and internal characteristics are different from each other and even two camera centers do not have the same vertical coordinates.
In this case, the epipolar lines on each side are not parallel to each other as shown in Fig. 1(a). This mismatch results in a vertical pixel difference between the two images, reducing the correlation between each view, as shown in Fig.2(a).

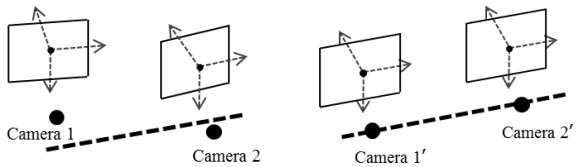
The discrepancies are also a major hurdle to stereo image processing and applications. Therefore, image rectification between two color images is required.

2.1 Camera Calibration

Before rectification, camera parameters should be estimated. In order to obtain relative camera information, we apply a camera calibration algorithm to each camera in the hybrid camera system. The projection matrix of the depth camera and each video camera can be obtained as follows

$$\mathbf{P} = \mathbf{K}[\mathbf{R} \mid \mathbf{t}] \tag{1}$$

where P is the projection matrix which is composed of the rotation matrix R and the translation vector t, indicating the camera orientation and position, respectively.



(a) Before image rectification *(b) After image rectification*

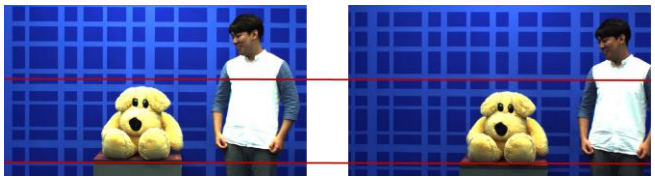
Figure 1. Stereo camera model

2.2 Rectification

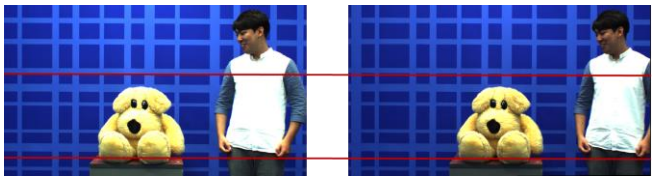
To perform stereo matching in real time, we need to reduce the execution time. Therefore, we separate the rectification method into the offline step and the online step. In the offline step, we calculate the 2D homography matrix using the projection matrices. The 2D homography matrix is defined as

$$\mathbf{H}_l = \mathbf{P}_{\mathbf{ref}}\mathbf{P}_{\mathbf{tar}}^{+} \tag{2}$$

where P_{ref} is the projection matrix of the reference image, P_{tar} is the projection matrix of the target image. In the online step, we use the GPU to perform the algorithm in real time. Each pixel was rectified using the homography matrix precomputed in the offline step. Figure 2(a) represents the unrectified stereo image, and Figure 2(b) show the rectified stereo image.



(a) Unrectified images



(b) Rectified images

Figure 2. Image rectification

3. Real time color correction

Even if we capture an image using the same type of camera, there is a color mismatch between the views. It is caused by the different color characteristic of the cameras and the capturing environment. Previous approaches to addressing color discrepancies fall into two categories: processing images after acquisition with and without pre-processing.

In order to perform the color correction in real time, the color correct is divided into the offline step and the online step. In the offline step, we estimate color transform matrix using a linear regression method. First, we apply 3d warping from ToF to color 1 and color 2, and we can match the correspondences as shown in Fig. 3. We construct a relation for values between source and target viewpoints. The vector equation is represented as

$$\mathbf{Y} = \mathbf{AX} \tag{3}$$

where Y and X are column vectors, and A is an m×n matrix. X and Y are shown as

$$\mathbf{X} = \begin{bmatrix} 1 & r_{s1} & r_{s1}^2 & r_{s1}^3 & r_{s1}^4 \\ 1 & r_{s2} & r_{s2}^2 & r_{s2}^3 & r_{s2}^4 \\ & & \vdots & & \\ 1 & r_{sn} & r_{sn}^2 & r_{sn}^3 & r_{sn}^4 \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} r_{r1} \\ r_{r2} \\ \vdots \\ r_{rn} \end{bmatrix} \tag{4}$$

We can find a translation matrix represented as

$$\mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{5}$$

In the online step, we use the GPU to perform the algorithm in real time. Figure 4 show the color corrected images.

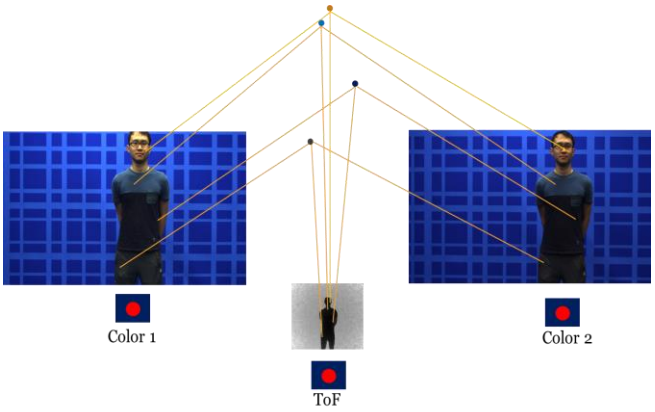
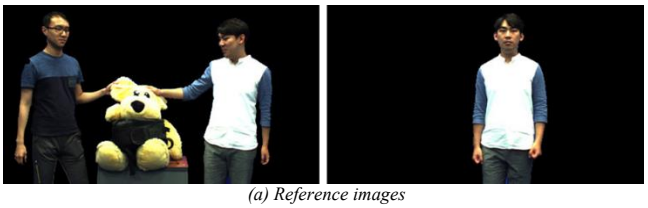
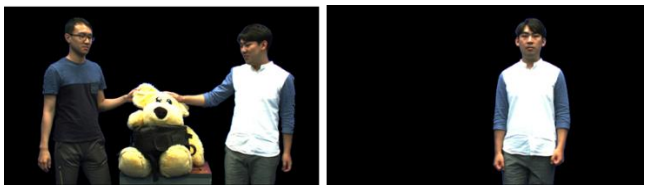


Figure 3. Correspondence matching



(a) Reference images

(b) Target images



(c) Color corrected images

Figure 4. Color correction

4. Hybrid stereo matching

4.1 ToF warping

Based on the pin-hole camera model, the pixel point m_l can be defined by the camera parameters as

$$\mathbf{m_l} = \mathbf{A_l} \cdot \mathbf{R_l} \cdot \mathrm{M} + \mathbf{A_l} \cdot t_l \quad \textbf{(6)}$$

The next step is to find the corresponding pixel position m_c in the view point of ToF 2. The point at the world coordinates M is projected using its camera parameters onto the view point of ToF 2 as

$$\begin{aligned} m_c &= \mathbf{A_c} \cdot \mathbf{R_c} \cdot \mathrm{M} + \mathbf{A_c} \cdot t_c \\ &= \mathbf{A_c} \cdot \mathbf{R_c} \cdot \mathbf{R_l^{-1}} \cdot \mathbf{A_l^{-1}} \cdot m_l - \mathbf{A_c} \cdot \mathbf{R_c} \cdot \mathbf{R_l^{-1}} \cdot t_l + \mathbf{A_c} \cdot t_l \end{aligned} \quad \textbf{(7)}$$

As in the case of ToF 1, the points of ToF 3 can be warped.

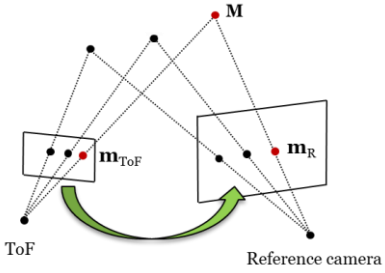


Figure 5. 3D warping

4.2 ToF upsampling

In order to obtain an initial high-resolution depth map, we use the mean filter to obtain a coarse result. Depth upsampling using mean filter is represented as

$$D_p^{MH} = \frac{\sum_{q \in S} D_p^L}{n} \quad \textbf{(8)}$$

Where D_p^{MH} the pixel is the value of the high-resolution depth map and D_p^L is the warped pixel value, n is the number of the warped pixel, and S is the size of the kernel.
ToF camera generates depth information instead of disparity, therefore, we need to change depth to disparity and vice versa. Using the initial disparity, we can reduce the disparity search range to speed up the stereo matching algorithm.

4.4 Hybrid stereo matching

4.4.1 cost matching

A 3D cost volume is generated by measuring matching costs for each pixel p at all possible disparity levels between the left image and the right image. In our implementation, we choose the modified census transform. The census transform is expressed as

$$D_s(d_s) = Hamming(I_c(p), \bar{I}_c(\bar{p}_d)) \quad (8)$$

where $I_c(p)$ and $\bar{I}_c(\bar{p}_d)$ are vectors converted using census transform, which is a non-parametric local transform method. Hamming distance is the number of differences between two vectors. Let $I_c(p)$ represents census transform of one point p. The center pixel's intensity value is replaced by the bit string composed of set of boolean comparisons such that in a square window and $I_c(p)$ is defined as

$$I_c(p) = \underset{q \in N_p}{\otimes} \xi(I_{MEAN}, I(q)) \quad (9)$$

where $\otimes$ denotes concatenation, N_p is neighboring pixels in a window, I_{MEAN} is the mean value from the window, and ξ denotes transform represented as

$$\xi(I_{MEAN}, I(q)) = \begin{cases} 0, if\ I_{MEAN} < I(q) \\ 1, \ otherwise \end{cases} \quad (10)$$

Census transform converts the relative intensity difference to 0 or 1 in 1 dimensional vector form. Figure 3(b) represents an example of the census transform of a window with respect to the center pixel.

4.4.2 Cost Aggregation

After constructing the cost volume, we exploit a smooth filter based on the guide image filter [6] to filter each slice of the cost volume in order. Using a guidance image I, the guided image filter can be used to calculate the cost volume as follows

$$C^g(p,d) = \sum_q W_{p,q} C(p,d) \quad (11)$$

where $C^g(p,d)$ denotes the aggregated cost using guided image filter, and $W_{p,q}$ is a filter weight. The filter weights are defined as

$$W_{i,j} = \frac{1}{|w|^2} \sum_{k:(i,j) \in wk} (1 + (I_i - \mu_k)(\sum_k + \varepsilon U)^{-1}(I_j - \mu_k)) \quad (12)$$

where $|w|$ is the total number of pixels in a window w_k centered at pixel k, and ε is a smoothness parameter. $\sum_k$ and μ_k are the covariance and mean of pixel intensities within w_k. I_s, I_t and μ_k are 3 ×1 vectors, while $\sum_k$ and the unary matrix U are of size 3 ×3. In order to speed up the algorithm, we exploit cross-shaped kernel.

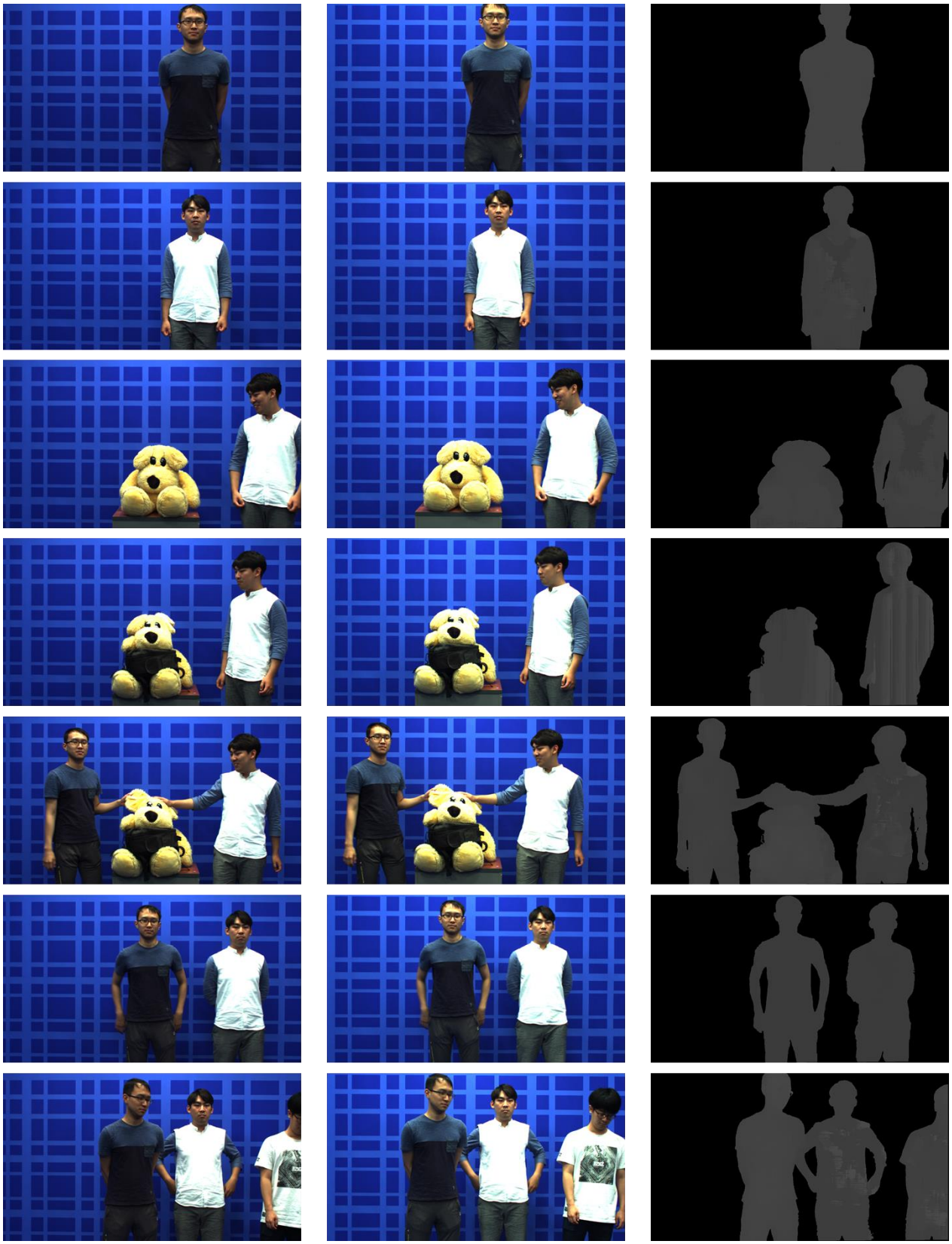


Figure 6. Results of proposed method.

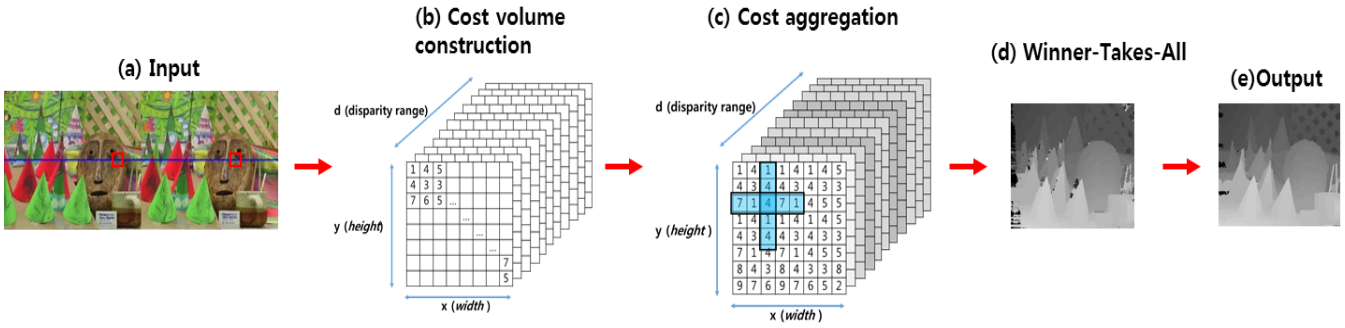


Figure 7. Process of proposed method

5. Experimental results

The stereo image and the depth map are captured by the hybrid stereo matching system. Our setup is composed of the two stereo cameras and the ToF camera. The stereo cameras installed horizontally, and the ToF camera is placed between the stereo cameras. The size of the image is 1080*720, and the size of ToF is 176*144. Before making color correction, we can notice distinct color differences due to inconsistencies in the points. However, after the proposed color correction, you can obtain more natural 3D models than the original 3D model as shown in Fig. 5.

(a) reference *(b) target* *(c) corrected*

Figure 5. Results of the color correction

Table 1 represents the result of the qualitative evaluation. We converted images from RGB to CIELab color space and calculated average values of Euclidean distance. The proposed method shows the closest distances for both target viewpoint 1 and 2. Through this result, the proposed method achieves the most consistent result among above results.

TABLE 1

Euclidean distance of color space

	result 1	result 2	result 3
original	13.7642	14.3441	13.3786
histogram matching	14.5035	15.1947	13.5124
linear regression	12.8332	14.0749	13.2916
global color transfer	14.7532	14.9127	13.4218
proposed	11.6847	13.2061	12.9725

In the modified census transform algorithm, we use 3*3 window for obtaining a mean value, and we use 5*5 census transform window to obtain the set of the boolean window. In the aggregation step, we exploit 19*19 cross kernel. The real data sets computed using the proposed hybrid stereo matching method are presented in Fig. 6, which shows that the proposed method generates the accurate result.

Table 2 shows the process time of the proposed method. According to the Table 1, it generates 1.7~3.3 fps depth images. The most time-consuming process is cost aggregation, and the second time-consuming process is the volume construction. To speed up the process, we need to increase both speeds.

6. Conclusions

In this paper, we proposed the real-time depth estimation method using the hybrid camera system. The proposed method divides the offline and the online steps. In the offline step, we perform complex algorithms in advance. In the online step, we use GPU to reduce the algorithm complexity Therefore, we can accelerate the depth estimation method. Experimental results show that our method generates the accurate disparity maps, and it generates 1.7~3.3 fps depth images for HD resolution.

TABLE 2

Process time of each algorithm

	figure 1	figure 2	figure 3	figure 4	figure 5	figure 6	figure 7
whole time	0.330	0.438	0.389	0.452	0.611	0.345	0.701
color correction	0.004	0.004	0.004	0.004	0.004	0.004	0.005
rectification	0.004	0.004	0.004	0.005	0.005	0.005	0.005
3D warping	0.001	0.001	0.001	0.001	0.001	0.001	0.001
CPU to GPU data transfer	0.033	0.036	0.034	0.036	0.031	0.034	0.024
volume construction	0.027	0.092	0.044	0.041	0.066	0.043	0.057
cost aggregation	0.156	0.095	0.232	0.250	0.490	0.182	0.470
CPU calculation	0.103	0.204	0.069	0.114	0.013	0.074	0.141

References

[1] G. Um, K. Kim, C. Ahn, and K. Lee, "Three-dimensional scene reconstruction using multiview images and depth camera," Proc. of 3D Digital Imaging and Modeling, pp. 271–280, 2005.

[2] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," Proc. of Advances in Neural Information Processing systems, pp. 291-298, 2005.

[3] E. K. Lee, and Y. S. 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, 2011.

[4] Y. S. Kang and Y. S. Ho, "Generation of Multi-view Images Using Stereo and Time-of-Flight Depth Cameras," International Conference on Embedded Systems and Intelligent Technology, pp. 104-107, 2013.

[5] W. S. Jang and Y. S. Ho. "Disparity fusion using depth and stereo cameras for accurate stereo correspondence," SPIE/IS&T Electronic Imaging International Society for Optics and Photonics pp. 93930T-93930T, 2015.

[6] K. He, J. Sun, and X. Tang. "Guided image filtering," European Conference on Computer Vision, pp. 1–14, 2010.

[7] R.C. Gonzalez, B.A. Fittes, "Gray-level transformations for interactive image enhancement," 2nd Conference on Remotely Manned Systems: Technology and Applications, pp. 17–19, 1975.

[8] J. Jung and Y. Ho, "Improved polynomial model for multi-view image color correction," The Journal of Korea Information and Communications Society, vol. 38, no. 10, pp. 881–886, 2013.

[9] E. Reinhard , M. Ashikhmin, B. Gooch , P. Shirley, "Color Transfer between Images," IEEE Computer Graphics and Applications, vol. 21 no.5, pp. 34-41, 2001.

[10] D. Shin, Y. Ho, "Color correction using 3D multi-view geometry," SPIE/IS&T Electronic Imaging. International Society for Optics and Photonics, pp. 93950O-93950O-6, 2015.

[11] Y. Kang, Y. Ho. "An efficient image rectification method for parallel multi-camera arrangement," IEEE Transactions on Consumer Electronics, vol. 57, no.3, pp.1041-1048, 2011.

Acknowledgment

This work was supported in part by the 'Cross-Ministry Giga KOREA Project' of the Ministry of Science, ICT and Future Planning, Republic of Korea (ROK). [GK16C0100, Development of Interactive and Realistic Massive Giga-Content Technology], and in part by the 'Brain Korea 21 Plus Project' of the Ministry of Education & Human Resources Development, Republic of Korea (ROK). [F16SN26T2205].

Author Biography

Eu-Tteum Baek received his B.S. degree in computer science and engineering from Chonbuk National University, Korea, in 2012 and M.S. degree in Information and Communication Engineering at the Gwangju Institute of Science and Technology (GIST), Korea, in 2015. He is currently working towards his Ph.D. degree in the Department of Information and Communications at GIST, Korea. His research interests are 3D digital image processing, depth estimation, and realistic broadcasting.

Yo-Sung Ho received his B.S. and M.S. degrees in electronic engineering from Seoul National University, Seoul, Korea (1981, 1983) and his Ph.D. in electrical and computer engineering from University of California, Santa Barbara, USA (1990). He worked at ETRI from 1983 to 1995, and Philips Laboratories from 1990 to 1993. Since 1995, he has been with Gwangju Institute of Science and Technology, Gwangju, Korea, where he is currently a professor. His research interests include video coding, 3D image processing, 3DTV, AR/VR, and realistic broadcasting systems.