IEIE Transactions on Smart Processing and Computing

Multiple Color and ToF Camera System for 3D Contents Generation

Yo-Sung Ho

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology / Gwangju, South Korea hoyo@gist.ac.kr

***** Corresponding Author:

Received March 15, 2017; Accepted April 30, 2017; Published June 30, 2017

** Short Paper*

** Invited Paper: This paper is invited by Donggyu Sim, the associate editor.*

** Review Paper: This paper reviews the recent progress possibly including previous works in a particular research topic, and has been accepted by the editorial board through the regular reviewing process.*

** Extended from a Conference: Preliminary results of this paper were presented at the IEEE VTC Fall 2012. This paper has been accepted by the editorial board through the regular reviewing process that confirms the original contribution.*

Abstract: In this paper, we present a multi-depth generation method using a time-of-flight (ToF) fusion camera system. Multi-view color cameras in the parallel type and ToF depth sensors are used for 3D scene capturing. Although each ToF depth sensor can measure the depth information of the scene in real-time, it has several problems to overcome. Therefore, after we capture low-resolution depth images by ToF depth sensors, we perform a post-processing to solve the problems. Then, the depth information of the depth sensor is warped to color image positions and used as initial disparity values. In addition, the warped depth data is used to generate a depth-discontinuity map for efficient stereo matching. By applying the stereo matching using belief propagation with the depth-discontinuity map and the initial disparity information, we have obtained more accurate and stable multi-view disparity maps in reduced time.

Keywords: Multi-depth generation, Multi-view camera, Time-of-flight, Depth sensor, 3D contents

1. Introduction

In recent years, three-dimensional television (3DTV) has been developed as the next-generation broadcasting system that can satisfy the growing demand for more realistic multimedia services [1]. In order to generate 3D video contents, we need multi-view images and depth information of the scene. Since the depth map represents the scene’s distance information, we can reconstruct intermediate-view images of the scene. They provide users wider and more natural 3D views. Therefore, generation of high-quality depth maps is necessary and their quality heavily influences the quality of 3D video contents.

In general, there are two categories when acquiring depth information of a scene: passive and active sensor-based methods. In the passive sensor-based method, stereo matching is one of the most popular methods [2]. Although it is efficient since obtaining data is easy, some difficulties, such as textureless or occluded regions, have remained.

The active range sensor-based methods generally employ measuring instruments to obtain the range information of the scene. Depth sensors based on a time-of-flight (ToF) technique are popular since they can obtain the depth information in real-time. However, ToF depth sensors have several problems to overcome, such as low spatial resolution and noisy acquisition, depending on the capturing environments.

Nowadays, there are several approaches proposed to obtain the depth information using both the passive and active range sensors. Those depth fusion systems are usually composed of stereo or multi-view cameras with ToF depth sensors. Gudmundsson et al. used a stereo camera with one ToF depth sensor, and they transferred the depth from the active range sensor to the color image position to initialize

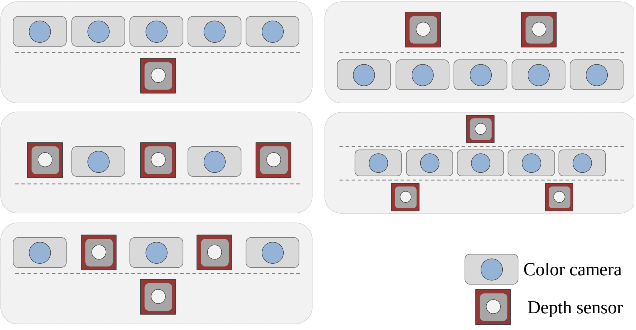


Fig. 1. Possible arrangements of color cameras and ToF depth sensors.

disparity for stereo matching [3]. Bartczak and Koch obtained high-definition multi-view color images and one low-resolution depth image. They also warped a low-resolution depth image to reference color views for depth map generation [4]. Other approaches mentioned the integration of multiple cameras and one ToF depth sensor to generate high-quality depth maps [5, 6].

In this paper, we propose a ToF fusion system for a multi-depth-generation. It is composed of multiple color cameras and up to three ToF depth sensors. We also propose a post-processing for captured images and multi-depth generation process. The contribution of this work is that we minimize the inherent error and distortion in the captured images, and generate multi-view depth information based on stereo matching using the post-processed color and ToF depth images. Since the ToF depth data are utilized as assistance for stereo matching, we can obtain more accurate depth information in reduced time. Therefore, the proposed ToF fusion system can be effectively used for 3D content generation and practical 3D applications.

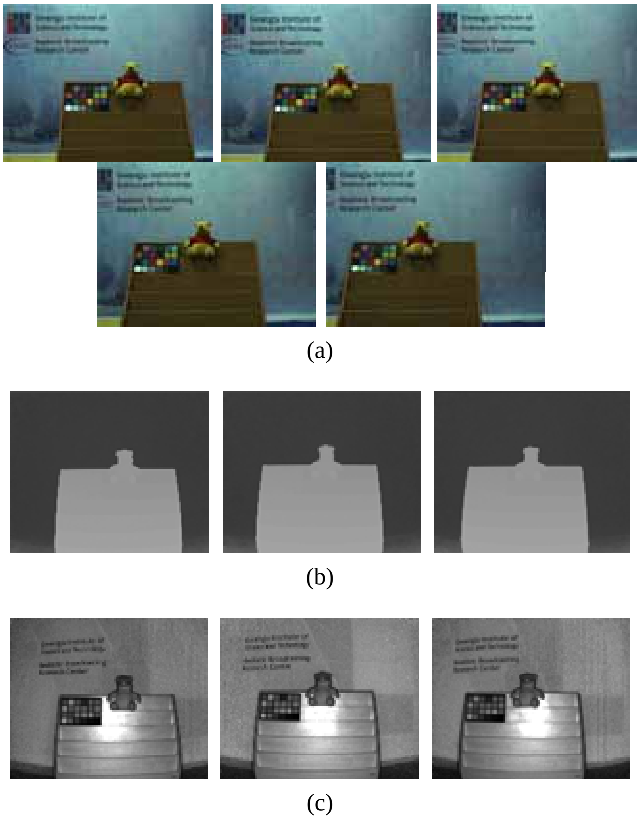
The remaining parts of this paper are organized as follows. In Section 2, we explain the proposed ToF fusion system and its problems. In Section 3 and Section 4, the proposed post-processing and multi-depth generation methods are explained. After showing the experimental results in Section 5, we conclude the paper in Section 6.

2. Multiple Color/ToF Fusion System

2.1 Camera Setup

Fig. 1 shows possible arrangements of multiple color cameras and ToF depth sensors. The color cameras are arranged in parallel, and the positions of the depth sensors are adjustable. The color cameras and the depth sensors are synchronized and connected to the control personal computers.

More than two color viewpoints are helpful for depth generation since it is hard to calculate the correct depth for occluded regions, weak-textured regions, and color-mismatched regions using two viewpoints. Therefore, the multiple color viewpoints help to solve this problem by referring to multiple adjacent-view images. In addition, the multiple viewpoints provide a wide viewing angle suitable for auto-stereoscopic 3D displays.


(a)

(b)

(c)

Fig. 2. Captured images (Image Set 1) (a) Multi-view color images, (b) multi-view depth images, (c) multi-view intensity images.

Moreover, the ToF depth sensors are used to assist multi-depth generation. Since the output resolution of the ToF depth sensor is usually too small, having multiple ToF depth sensors is advantageous for covering the views of all the high-resolution color images.

2.2 Color and Depth Capture

We obtain five-view high-resolution color images and three-view low-resolution depth and intensity images, as shown in Fig. 2. In the proposed ToF fusion system, we operate up to three ToF depth sensors, since they provide only three modulation frequencies and we need to avoid an interference problem among the emitted light signals [7].

2.3 Error and Distortion

Captured color and depth images have inherent errors and distortions. In color images, there are non-uniform disparities and vertical pixel mismatches among the views. Besides, they have different color representations for the same scene because of the camera's physical features from light effects. These problems can pose difficulties in the stereo matching process to generate the depth information.

Moreover, the output image of the ToF depth sensor not only has a large amount of lens distortion, as shown in Figs. 2(b) and (c), but also has boundary errors. In addition, there is a difference between measured depth values by the depth sensor and estimated depth values from the color cameras, since two types of cameras have different positions in the z-

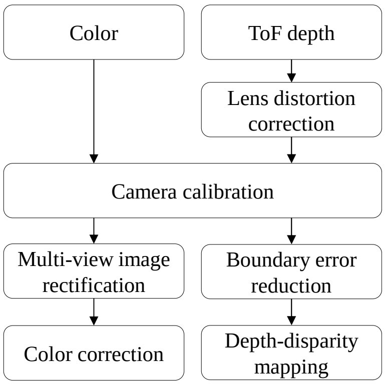
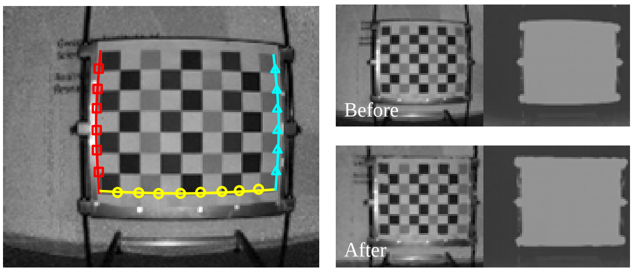


Fig. 3. Flowchart of the proposed post-processing.



(a) (b)

Fig. 4. Lens distortion correction (a) finding distorted lines, (b) before and after lens distortion correction.

direction.

3. Post-processing of Captured Images

In this section, we explain each step in post-processing the captured color and depth images. The purpose of the post-processing is to minimize the inherent errors and distortions in the captured images. Fig. 3 shows the flowchart of the proposed post-processing.

3.1 Lens Distortion Correction

The depth image of the ToF depth sensor has a large amount of lens radial distortion. Due to this distortion, there exists a shape mismatch between the color and depth images. Also, the lens distortion interferes with some feature point-based processing, such as camera calibration.

In order to reduce this lens radial distortion, we perform lens distortion correction on the captured depth and intensity images [8]. After finding three distorted straight-line components in the distorted image, as shown in Fig. 4(a), we estimate the distortion center and the distortion parameter. Then, we can reconstruct the original image from the distorted image. Fig. 4(b) shows the depth and intensity images before and after the lens distortion correction.

3.2 Camera Calibration

Camera calibration is the process to obtain the camera parameters from the captured image. In order to calibrate multiple cameras, we use a number of checker board images, and then extract several corner points to find the cameras' intrinsic and extrinsic parameters [9]. In the proposed method, we calibrate the depth sensors after lens distortion correction. Fig. 5 shows the lens distortion corrected checker board images for camera calibration.

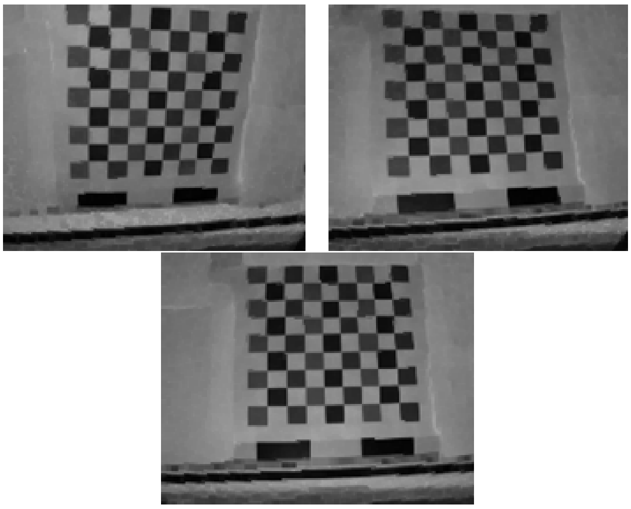


Fig. 5. Lens distortion corrected pattern images.

3.3 Image Rectification

The proposed ToF fusion system has mounted color cameras in parallel manually. In general, images from the manually mounted multiple cameras, shown in Fig. 6(a), have geometric errors. These errors appear as vertical pixel mismatches and irregular horizontal disparities between the corresponding points in each view, as shown in Fig. 6(c).

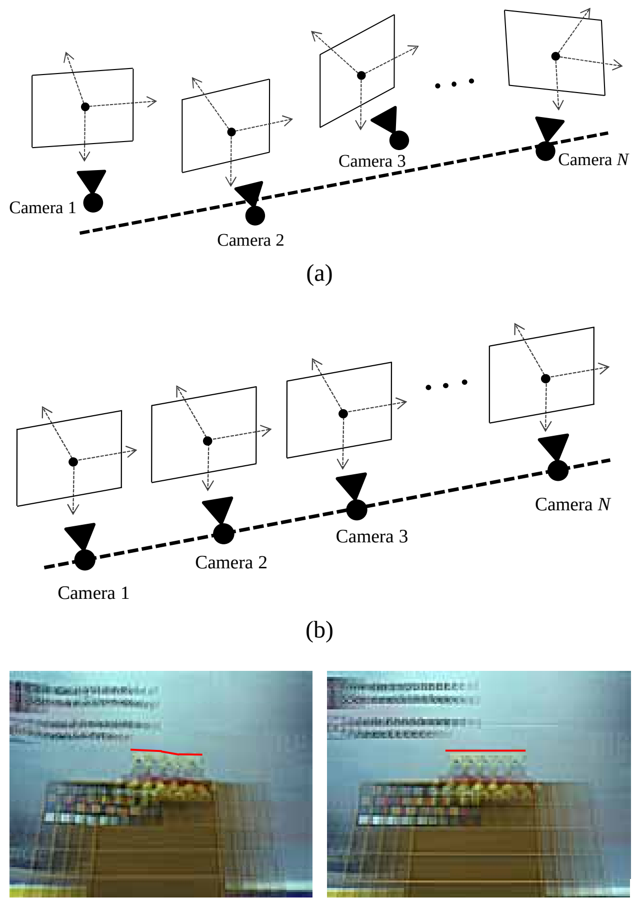
Geometric errors in the multi-view images not only decrease time efficiency and accuracy of stereo matching, but they also reduce the visual quality of the 3D video contents. Thus, it is essential to minimize the geometric errors in the multi-view images.

Multi-view image rectification is a transformation to minimize the geometric errors in the multi-view images. Based on the original camera parameters, we estimate the new camera parameters. These new camera parameters have the same intrinsic features, rotations, and camera intervals, as shown in Fig. 6(b). Using these two sets of camera parameters for each camera, we calculate the transform matrices and apply them for all the cameras. As shown in Fig. 6(d), the rectified images have the same vertical coordinates between the corresponding points, and the same horizontal distances to adjacent viewpoints. [10].

3.4 Color Correction

Although we use the same type of cameras in our system, there are color differences among multiple-view images because of the different color response of each camera and the different illumination conditions. Because stereo matching calculates the data cost based on the color consistency between the target and reference-view images, we need to solve the color mismatch problem.

In order to correct the color differences, we use the



(a)

(b)

(c)

(d)

Fig. 6. Multi-view image rectification (a) practical camera arrangement, (b) rectified camera arrangement, (c) before image rectification, (d) after image rectification.

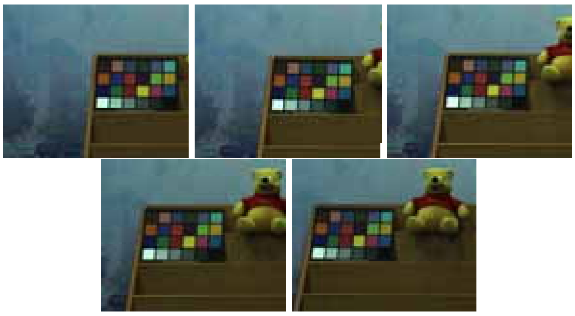


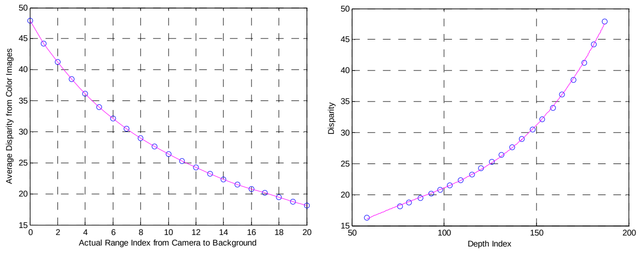
Fig. 7. Color chart images for color correction.

standard color chart, called a Macbeth chart. Each camera captures an image of the color chart, as shown in Fig. 7. After we define one view (usually the middle view) as the reference view, we apply a mapping function for each view based on the color chart samples [11].

3.5 Boundary Error Reduction

In this step, we minimize the boundary errors in the depth sensor images. These errors exist along object boundary pixels that have median depth values between two different depth planes. However, the ideal depth image has to divide the depth discontinuities without the median values. These median values are transferred to the wrong positions by 3D warping. Since the warped depth values are used as the initial disparity in the proposed method, the boundary errors decrease the performance of the multi-depth generation operation.

In order to solve this problem, we apply the shock filter to the depth camera image [12]. The shock filter can convert the smoothly increasing form to the form of step function. In addition, it flattens optical noises in homogeneous regions.



(a)

(b)

Fig. 8. Curves of depth-disparity mapping (a) disparity characteristics of the scene, (b) depth-disparity mapping function.

3.6 Depth-Disparity Mapping

Generally, disparity that is obtained in the rectified multi-view images of the scene, has a non-linear characteristic, as shown in Fig. 8(a). The data shown in Fig. 8(a) were obtained by capturing images of the checker-board patterns from several predefined positions within the background and camera, as shown in Fig. 9. The x-axis represents the distance from the camera position, and the y-axis indicates the disparity values in the pixels.

Due to the different representations of the actual range of the scene, it is necessary for us to correct the depth image. As shown in Fig. 8(b), we estimate a cubic curve between the disparity values and depth indexes for each camera. Using the estimated curves for each view, we can correct the depth index of each pixel of the depth image. Then, the corrected depth image from each view has the characteristics obtained from the corresponding color image.

4. Multi-depth Generation

In this section, we explain how to generate multi-view depth information. By using 3D warping, the ToF depth image is warped to the color image position, and it generates the depth-discontinuity map for the color image. After we estimate the initial disparity values, we can generate the disparity map of each view using the post-processed images.

4.1 3D Warping

For accurate stereo matching, we use the initial disparity values. We regard the depth information obtained by the depth sensors as the initial disparity of a multi-view image. In order to match the depth images that have different resolutions compared to the color images, and that are

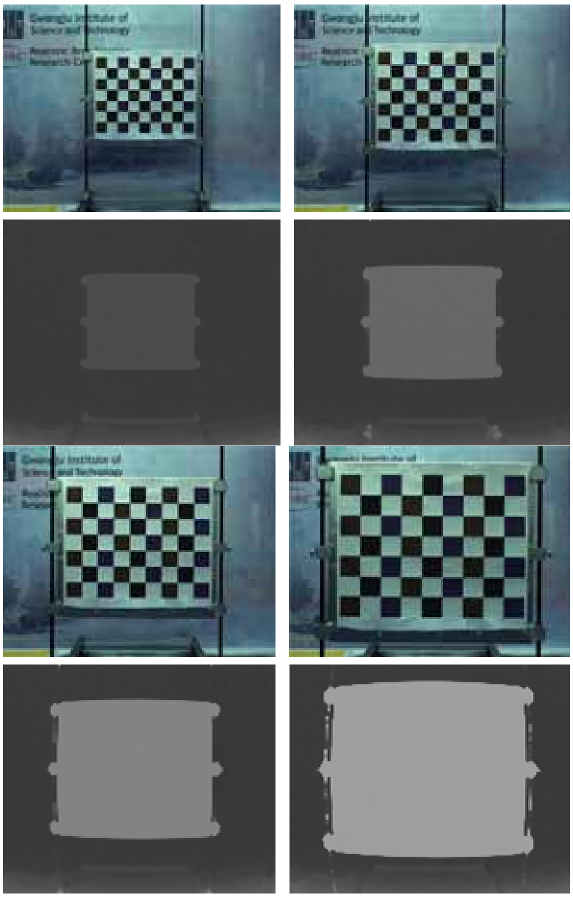


Fig. 9. Color and depth images for depth-disparity mapping.

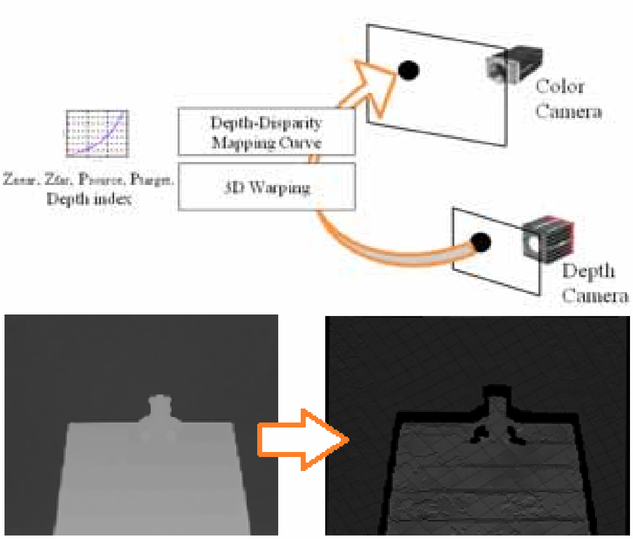

Depth-Disparity Mapping Curve
3D Warping
Znear, Zfar, Psource, Ptarget, Depth index
Color Camera
Depth Camera

Fig. 10. 3D warping from depth sensor to color camera.

captured at different positions from the color cameras, we back-project the ToF depth images to the world coordinates using 3D warping. Then, the depth information in the 3D space is re-projected onto each image plane of the video camera. In the 3D warping process, ToF depth values are mapped to the disparity values via the depth-disparity mapping curve. Fig. 10 shows that the ToF depth information is mapped to the color image plane.

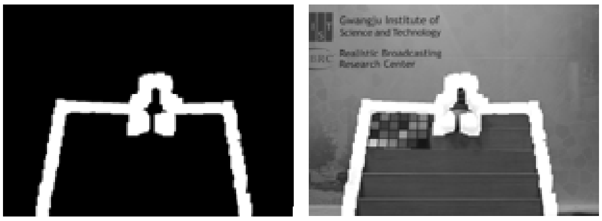


Fig. 11. Depth-discontinuity map.

When we perform 3D warping of the ToF depth image, we first calculate the edge of the ToF depth image using the Canny edge detector. The edge indicates the depth discontinuity in the scene. Then, we exclude the edge pixel for 3D warping since the edge pixels that have median depth values could be located in the wrong position, as explained in Section 3.5.

4.2 Generation of Depth-discontinuity Map

The edge map calculated in Section 4.1 is also warped to the color image position to generate the depth-discontinuity map. It is difficult to separate the depth-discontinuity and depth-homogeneous regions using only color information. However, the edge map of the ToF depth image directly gives the depth-discontinuity information.

Several dilation operations to the warped edge map of the ToF depth image produce a rough depth-discontinuity map. In order to ensure that the discontinuity region becomes wider, we perform a few erosion operations to the rough depth-discontinuity map. The amount of dilation and erosion could be adjustable according to the color image resolution. Fig. 11 shows the depth-discontinuity map that is suitable for the color image.

4.3 Stereo Matching

We design the energy function E(f) using the data cost, Dp(fp), smoothness cost, S(fp, fq), and ToF cost, Tp(fp), shown in Eq. (1), where p, q, fp and fq are the current pixel positions, neighboring pixel positions, and possible disparity values in the given search range.

$$E(f) = \sum_{p} D_p(f_p) + \sum_{p,q} S(f_p, f_q) + \sum_{p} T_p(f_p) \qquad (1)$$

The data term calculation uses the sum of absolute differences (SAD) between the current and target viewpoints. In our approach, we use both left and right views to get more accurate depth information.

In order to calculate the data term for stereo matching, we use the initial disparity information from the ToF depth image by 3D warping and depth-disparity mapping. Using the initial disparity information, we reduce the computation time and increase the accuracy of stereo matching.

There are many holes, which have no initial disparities, in the warped ToF depth image, because of the resolution differences between color images and depth camera images. We divide the warped ToF depth image into the following three cases based on conditions [13].

Case I shows that there is only one initial disparity value in the pixel window. The center pixel indicates the current pixel position. There are more than two different initial disparity values in Case II. Case III has no initial disparity values around the current pixel. In each case, we apply a different method to calculate the data term, considering the depth-discontinuity map.

At first, for each depth-homogeneous region, we trust the initial disparity from the ToF depth camera, since there is no depth discontinuity and no drastic change of depth values. For Case I, the data term of the current pixel has the lowest cost at the only one disparity value in the window, and we skip the cost calculation for the other disparity candidates. For Case II, the initial disparity d_i for the current pixel is calculated as the weighted sum of neighboring values.

$$d_i(x,y) = \sum_k w_k d_{k,m}(x,y) \qquad (2)$$

In Eq. (2), w_k is the weighting factor and $d_{k,m}$ is the average of the pixel values according to the distance from the current pixel, as seen in Eq. (3) where *n* indicates the number of initial disparity values at the distance *m*.

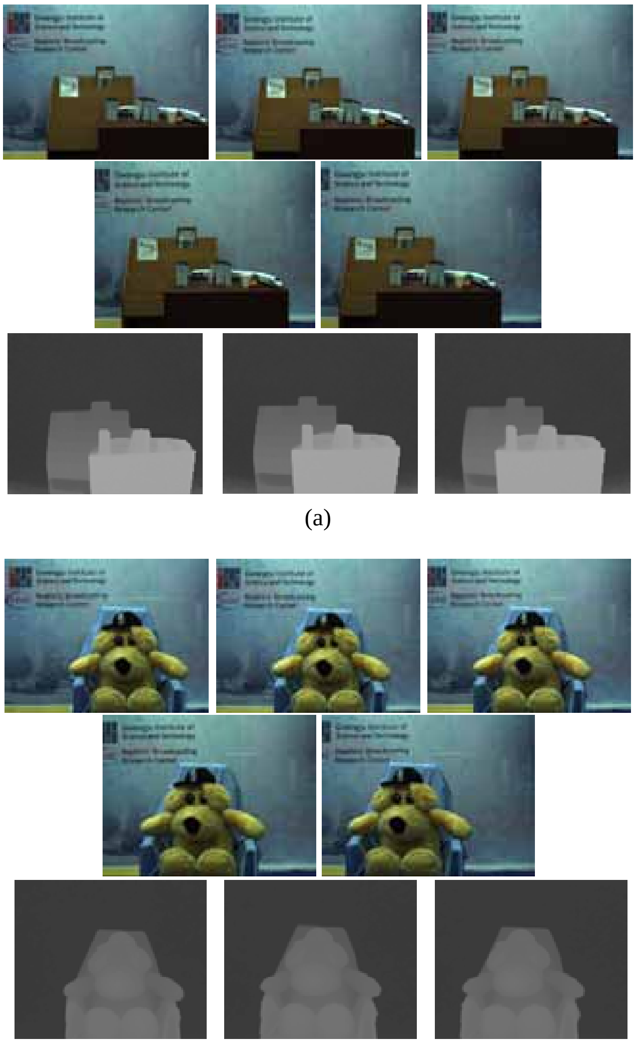
$$d_{k,m}(x,y) = \frac{1}{n}\sum_{i=-k}^{k}\sum_{j=-k}^{k} d(x+mi, y+mj) \qquad (3)$$

After calculating the initial disparity for the current pixel, we define a short search range for data term calculation. Then, the data costs of the initial disparity value and around the initial disparity are computed. We skip the cost calculation for the other disparity candidates outside the search range.

Secondly, for the depth-discontinuous regions, we calculate the data term more carefully, since there could be radical depth changes. In Case I, we define the short search range around the only one initial disparity in the window, as with Case II for the depth-homogeneous regions. The data costs are calculated for the disparity candidates in the search range. For Case II, the search range becomes wider. The initial disparity is the average value of all the values in the window. The search range goes from the minimum value to the maximum value in the window, and the data costs are calculated for the disparity candidates in this range.

Thirdly, we calculate the data costs for all the disparity candidates for Case III, regardless of the depth-discontinuity map. Since we have no initial disparity, the search range is defined as being from the minimum disparity to the maximum disparity of the scene.

In (1), the smoothness term is calculated based on upper, lower, left, and right pixel values. The ToF term T_p represents the difference between f_p and $d_{p,i}$ as Eq. (4), where $d_{p,i}$ is the initial disparity value at pixel *p*. However, for Case III, we set the ToF term to zero, since there is no initial disparity value. Finally, we optimize this function using belief propagation and obtain the final disparity map [14].



(a)

(b)

Fig. 12. Captured images (a) book, (b) dog.

$$T_p(f_p) = \begin{cases} |f_p - d_{p,i}| & (Case\ I\ \&\ II) \\ 0 & (Case\ III) \end{cases} \qquad (4)$$

5. Experimental Results

We captured five-view color images and three-view ToF depth images. The ToF depth sensors were located below the second, third, and fourth color cameras. The horizontal and vertical distances between two cameras were 6.5cm and 7.5cm, respectively. The resolution of the color images is 1280×960 and the ToF depth images have 176×144 pixels.

We captured several test image sets. One set is shown in Fig. 2, and the other two sets are shown in Fig. 12. The results of post-processing for the captured images are shown in the previous sections. Then, we used a 9×9 search window for the initial disparity determination, and an 11×11 window for data cost calculation. For Case I and Case II, we assigned the short search range as four pixels. We generated the disparity maps of view 2, view 3, and view 4. View 1 and view 5 were used as reference viewpoints for stereo matching.

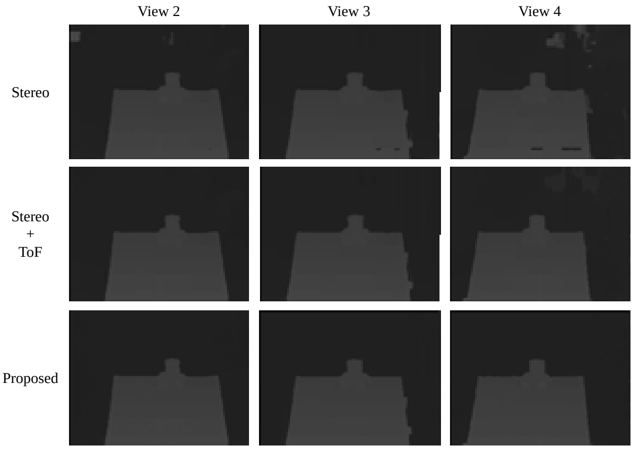


Fig. 13. Depth-discontinuity map.

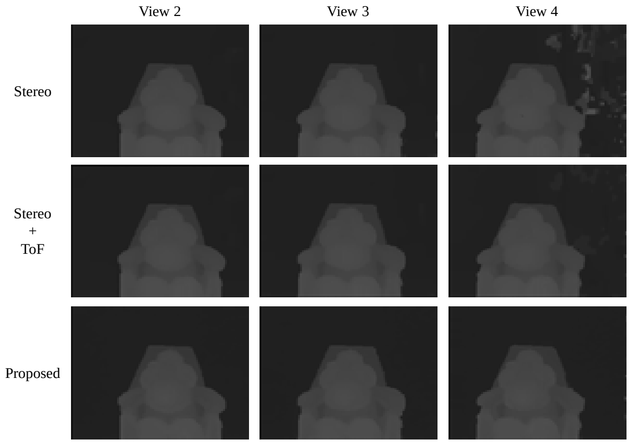


Fig. 14. Depth-discontinuity map.

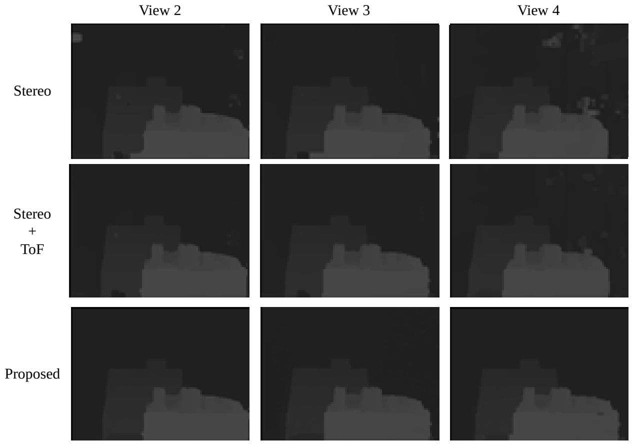


Fig. 15. Depth-discontinuity map.

Figs. 13-15 show the generated disparity maps of three images sets from stereo matching [14], stereo matching with ToF depth information [5], and from the proposed method, respectively. The disparity maps from stereo matching have wrong disparity values in the background regions especially for View 4. It is because of the weak texture in the background.

Table 1. Total average of vertical pixel mismatches.

Sequence		Stereo Matching	Stereo + ToF	Proposed
Image Set 1	PSNR (dB)	28.37	31.37	**32.28**
	Time (sec.)	102.28	93.79	**72.97**
Image Set 2	PSNR (dB)	31.99	34.33	**34.90**
	Time (sec.)	90.74	83.82	**62.99**
Image Set 3	PSNR (dB)	30.86	30.80	**33.33**
	Time (sec.)	89.21	84.26	**64.51**

The results of Zhu et al. [5] and the results of the proposed method are similar in the background regions and along the object boundaries. However, the proposed method generated more stable disparity values for the background regions in View 4. In addition, as shown in Fig. 14, the stereo-with-ToF method has inaccurate disparity values at the lowest part of the middle object in View 2 and View 3, and the object boundary is incorrect for the upper side of the toy train.

The proposed method outperform the other methods, as shown in Table 1. In order to compare their results, we synthesized the intermediate views at the position for View 3 using the color images and disparity maps of View 2 and View 4, and we calculated the peak signal-to-noise ratio (PSNR) between the original and synthesized images of View 3. For all test image sets, the results of the proposed method generated more accurate intermediate view images, taking less processing time. The processing time is the average of the disparity generation times of View 2, View 3, and View 4. The accuracy increment and time reduction were obtained by skipping the data cost calculations and decreasing the search range for matching based on the three cases and the depth-discontinuity information.

6. Conclusion

In this paper, we proposed a multi-depth generation method using a ToF fusion system. After capturing multi-view color and ToF depth images, we performed post-processing to reduce the inherent error and distortion in the original images. Then, the ToF depth images were warped to color image positions and used for generation of a depth-discontinuity map and the initial disparity values for stereo matching. Using the depth-discontinuity map and the initial disparity, we increased the accuracy of the disparity map and reduced the processing time. Since the proposed method processes the result based on the initial disparity values, and it considers the ToF term in the energy function, we obtain more accurate and stable disparity values in weak texture regions and occluded regions than the results from previous methods. We expect that the proposed ToF fusion system and multi-depth generation method could be useful for 3D content generation and practical 3D applications.

Acknowledgement

This work was supported 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].

References

[1] A. Smolic and P. Kauff, "Interactive 3D Video Representation and Coding Technologies," Proceedings of the IEEE, Spatial Issue on Advances in Video Coding and Delivery, vol. 93, no. 1, pp. 99-110, Jan. 2005.

[2] J. Sun, N.N. Zheng, and H.Y. Shum, "Stereo Matching Using Belief Propagation," IEEE Transactions of Pattern Analysis and Machine Intelligence (PAMI), vol. 25, no. 5, pp. 787-800, July 2003.

[3] S. A. Gudmundsson, H. Aanaes, and R. Larsen, "Fusion of Stereo Vision and Time-of-Filght Imaging for Improved 3D Estimation," International Journal of Intelligent Systems Technologies and Applications, vol. 5, no. 3, pp. 425-433, Nov. 2008.

[4] B. Bartczak and R. Koch, "Dense Depth Maps from Low Resolution Time-of-Flight Depth and High Resolution Color Views," Proc. of 5th International Symposium on Visual Computing, pp. 1-12, Nov. 2009.

[5] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of Time-of-Flight Depth and Stereo for High Accuracy Depth Maps," Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 231-236, June 2008.

[6] K. D. Kuhnert and M. Stommel, "Fusion of Stereo-Camera and PMD-Camera Data for Real-time Suited Precise 3D Environment Reconstruction," Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4780-4785, Mar. 2006.

[7] SR4000 User Manual Version 0.1.2.2, Mesa Imaging AG, pp. 18-19.

[8] A. Wang, T. Qiu, and L. Shao, "A Simple Method of Radial Distortion Correction with Centre of Distortion Estimation," Journal of Mathematical Imaging and Vision, vol. 35, no. 3, pp. 165-172, Nov. 2009.

[9] Z. Zhang, "A Flexible New Technique for Camera Calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 22, no. 11, pp. 1330-1334, Oct. 2000.

[10] Y.S. Kang and Y.S. Ho, "Geometrical Compensation for Multi-view Video in Multiple Camera Array," Proc. of International Symposium on Electronics and Marine (ELMAR), pp. 83-86, Sept. 2008.

[11] N. Joshi, B. Wilburn, V. Vaish, M. Levoy, and M. Horowitz, "Automatic Color Calibration for Large Camera Arrays," in UCSD CSE Tech. Rep. CS2005-0821, May 2005.

[12] G. Gilboa, N. Sochen, and Y.Y. Zeevi, "Regularized Shock Filters and Complex Diffusion", ECCV 2002, LNCS 2350, pp. 399-313, Springer-Verlag, May 2002.

[13] Y.S. Ho and Y.S. Kang, "Multi-view Depth Generation using Multi-Depth Camera System," Proc. of International Conference on 3D Systems and Application (3DSA), pp. 1-4, May 2010.

[14] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient Belief Propagation for Early Vision," International Journal of Computer Vision, vol. 70, no. 1, pp. 41-54, Oct. 2006.

Yo-Sung Ho received the B.S. and M.S. degrees in electronic engineering from Seoul National University, Seoul, Korea, in 1981 and 1983, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 1990. He joined the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea, in 1983. From 1990 to 1993, he was with Philips Laboratories, Briarcliff Manor, NY, where he was involved in development of the advanced digital high-definition television system. In 1993, he rejoined the Technical Staff of ETRI and was involved in development of the Korea direct broadcast satellite digital television and high-definition television systems. Since 1995, he has been with the Gwangju Institute of Science and Technology (GIST), Gwangju, Korea, where he is currently a Professor in the School of Electrical Engineering and Computer Science. Since August 2003, he has been Director of Realistic Broadcasting Research Center at GIST in Korea. His research interests include digital image and video coding, image analysis and image restoration, advanced coding techniques, digital video and audio broadcasting, 3-D television, and realistic broadcasting.