GENERATION OF MULTIPLE DEPTH IMAGES FROM A SINGLE DEPTH MAP USING MULTI-BASELINE INFORMATION

Eun-Kyung Lee, Seung-Uk Yoon, and Yo-Sung Ho

Gwangju Institute of Science and Technology (GIST)
1 Oryong-Dong, Buk-Gu, Gwangju, 500-712, Korea
E-mail: {eklee78, suyoon, hoyo}@gist.ac.kr

ABSTRACT

In this paper, we propose an algorithm for generation of multiple depth images from one depth map using multi-baseline information. Although depth information is essential in three-dimensional (3-D) applications, the depth estimation from stereo matching is still time consuming and inaccurate. Since the real-time application like free viewpoint TV (FTV) require multiple depth information, stereo matching algorithms can have limitations in terms of speed and accuracy. Thus, it is necessary to construct an effective and fast depth map generation algorithm. This paper assumes that there is an active sensor depth camera, which can obtain a depth map in real-time, and supportive multiple cameras. In order to generate multiple depth maps, we use a single depth map acquired from the depth camera and geometric relationships between cameras. Therefore, the proposed algorithm is faster and more accurate than other methods based on traditional stereo matching. Finally, our method is robust to inaccurate camera parameters since it uses baseline information only.

Keywords: depth map generation, free viewpoint TV, stereo matching

GENERATION OF MULTIPLE DEPTH IMAGES FROM A SINGLE DEPTH MAP USING MULTI-BASELINE INFORMATION

Eun-Kyung Lee, Seung-Uk Yoon, and Yo-Sung Ho

Gwangju Institute of Science and Technology (GIST)
1 Oryong-Dong, Buk-Gu, Gwangju, 500-712, Korea
E-mail: {eklee78, suyoon, hoyo}@gist.ac.kr

ABSTRACT

In this paper, we propose an algorithm for generation of multiple depth images from one depth map using multi-baseline information. Although depth information is essential in three-dimensional (3-D) applications, the depth estimation from stereo matching is still time consuming and inaccurate. Since the real-time application like free viewpoint TV (FTV) require multiple depth information, stereo matching algorithms can have limitations in terms of speed and accuracy. Thus, it is necessary to construct an effective and fast depth map generation algorithm. This paper assumes that there is an active sensor depth camera, which can obtain a depth map in real-time, and supportive multiple cameras. In order to generate multiple depth maps, we use a single depth map acquired from the depth camera and geometric relationships between cameras. Therefore, the proposed algorithm is faster and more accurate than other methods based on traditional stereo matching. Finally, our method is robust to inaccurate camera parameters since it uses baseline information only.

Keywords: depth map generation, free viewpoint TV, stereo matching

1. INTRODUCTION

As demands for high-quality visual services are increasing, three-dimensional (3-D) information is widely used in various multimedia applications, such as free viewpoint TV (FTV), 3DTV, games, and educational tools. Since these applications provide viewers with depth impression and an immersive sense of reality, depth acquisition is one the most essential issues in these 3-D applications. Usually, there are two ways to acquire depth information: depth from active sensor depth camera system and depth estimation from stereo matching. The latter takes long time and complex. In spite of its complexity, it does not guarantee the accuracy of estimated depth. On the contrary, as sensor technologies for obtaining depth information are developed rapidly, we can capture more accurate per-pixel depth information from real scenes directly using the depth camera system. However, the depth camera system has disadvantages: the high cost and limited viewing range. Therefore, we need a multi-view camera system to solve these problems.

The multi-view camera system captures the same scene at different viewpoints. If we acquire multi-view images from multiple cameras, we can generate scenes at arbitrary view positions. It means that users can change their viewpoints freely and can feel visible depth with view interaction. This feature is very useful for the FTV system since it requires depth maps for every viewpoint [1]. In this case, the problem is how to generate multiple depth maps. One possible way is to use stereo matching for every stereo pairs with additional information from multi-view data like multi-baseline stereo [2]. However, it is time consuming and heavy to perform the stereo matching for all those pairs. Consequently, it is not appropriate for broadcast applications like FTV requiring a real-time functionality. Other feasible method is to use the depth camera system together with multi-view cameras. Then, we can have benefits from both systems: acquiring accurate depth in real-time for one viewpoint and multi-view information.

In this paper, we try to combine these two advantages to generate multiple depth maps in an effective way. In order to avoid the complex stereo matching for all pairs, we use one depth map obtained from the depth camera system as an accurate reference. Then, depth images at other viewpoints are constructed using this reference depth and geometric relationships between multiple cameras in real-time. Therefore, the proposed algorithm is faster than other methods based on stereo matching. Moreover, it provides more accurate multiple depth images.

2. RELATED WORKS

2.1 Free Viewpoint TV (FTV)

Television realized the human dream of seeing a distant world in real-time and has served as the most important visual information technology to date. However, users can get only a single view of a 3-D world with conventional TV. The view is determined not by users but by a camera placed in the 3-D world. This is quite different from what we experience in the real world.

FTV is an innovative visual media that enables us to view a distant 3-D world by freely changing our viewpoints as if we were there. FTV has much potential since such a function has not yet been achieved by conventional TV technology [1].

ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) has recognized the importance of FTV technologies. FTV was proposed to MPEG in 2002. It got a strong support from industry in response to “Call for Comments on 3DAV” [3]. Then, MPEG has started the standardization of Multi-view Video Coding (MVC) for multiple videos that are the input of the FTV system.

However, MVC technologies deal with compression issues only. Therefore, the standardization is needed for other parts of the FTV system and for practical uses. Requirements on FTV were investigated and an Ad Hoc Group (AHG) on FTV has been established in April 2007 [4].

FTV realizes free navigation of users due to its new function of view generation. FTV can generate views at infinite number of viewpoints from finite number of camera images. FTV is implemented as the real-time complete chain from multi-view capturing to free viewpoint display. The user can freely control the viewpoint for a real dynamic 3D scene. It should be noted that FTV provides us not only free viewpoints but also 3D scene reproduction. In order to achieve free navigation functionality, depth information is required in addition to video signals. Figure 1 shows a system that transmits multi-view video with depth information (MVD) [1]. The content may be produced in a number of ways, e.g., with multi-camera setup, depth cameras, or 2-D/3-D conversion. At the receiver, depth image based rendering can be performed to project the signal to various types of displays.

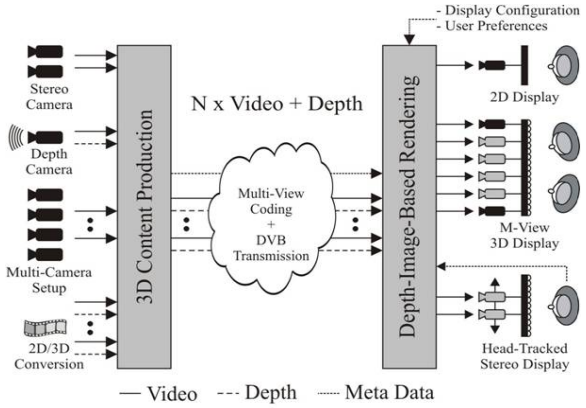

Fig. 1: FTV system and FTV/MVD data format.

2.2 Multi-baseline Stereo

Stereo is a useful technique for obtaining 3-D information from 2-D images in computer vision. In stereo matching, we measure the *disparity d*, which is the difference between the corresponding points of left and right images. The *disparity d* is related to the distance *Z* by

$$d = B \cdot f \cdot 1/Z \qquad (1)$$

where *B* and *f* are baseline and focal length, respectively. This equation indicates that for the same distance, the disparity is proportional to the baseline or that the baseline length *B* acts as a magnification factor in measuring *d* in order to obtain *Z*, that is, the estimated distance is more precise if we set the two cameras farther apart from each other, which means a longer baseline. A longer baseline, however, poses its own problem. Because a longer disparity range must be searched, matching is more difficult, and thus, there is a greater possibility of a false match. Therefore, there is a tradeoff between precision and accuracy (correctness) in matching.

The multi-baseline stereo presented use of multiple images with different baselines obtained by a lateral displacement of a camera. The matching technique, however, is based on the idea that global mismatches can be reduced by adding the sum of squared difference (SSD) values from multiple stereo pairs, that is, the SSD values are computed first for each pair of stereo images.

2.3 Realistic Broadcasting System

As the rapid development of telecommunication techniques and high-speed networks, we live in an age of the information revolution and the digital epoch. Humans acquire useful knowledge and information through the Internet since high-speed networks are connected with high-performance personal computers. We cannot only feel deep impressions by a high definition television with a large screen and a high power speaker, but also call to someone using a cellular phone with moving pictures. In addition, banking services and product purchases are possible at home. The digital technologies make a human life more convenient and livelier.

It is not too much to say that the essence of the digital age is the multimedia technologies for digital broadcasting. The digital broadcasting system converts analog multimedia data into digital multimedia data and then transmits the digitized data to the end users. The digital broadcasting is suitable for a high-quality and multi-channel broadcasting in comparison with an analog broadcasting. Furthermore, a digital broadcasting system can make use of the frequency bandwidth effectively and have great advantages for the data broadcasting [5].

Especially, realistic broadcasting makes an appearance for the next generation broadcasting. Realistic broadcasting provides not only high quality visual services, but also a variety of user-friendly interactions. Unlike other digital broadcasting systems, we can experience the realism through our five senses. Several countries have already served 3-D broadcasting experimentally using their satellites. Furthermore, a large number of researchers are interested in developing 3-D displays and 3-D data processing techniques since the scale for 3-D data markets is supposed to be three billion dollars in 2010. In this paper, we will survey the technologies for a realistic broadcasting system, described by Realistic Broadcasting Research Center (RBRC) in Korea, so as to keep pace with this trend. Figure 2 shows the conceptual illustration for realistic broadcasting [6].

Fig. 2: Realistic broadcasting system.

3. Generation of Multiple Depth Images

3.1 The Concept of Depth from Stereo

Stereoscopic imaging and display is motivated by how humans perceive depth through two separate eyes. Although there are several monocular cues that the human brain uses to differentiate objects at different depths, the most important and effective mechanism is though a binocular cue known as stereoscopic system [7].

Simple geometry and algebra is needed to understand how 3D points can be located in space using a stereo sensor as shown in Fig. 3. Assume that two cameras are carefully aligned so that their *X*-axes are collinear and their Y and Z-axis are parallel. The origin or center of projection of the right camera is offset by *B*, which is *baseline* of the stereo system. The system observes some object point *X* in the left image point x_l and the right image x_r. Geometrically, we know that the point *X* must be located at the intersection of ray $C_l\, x_l$ and the ray $C_r\, x_r$.

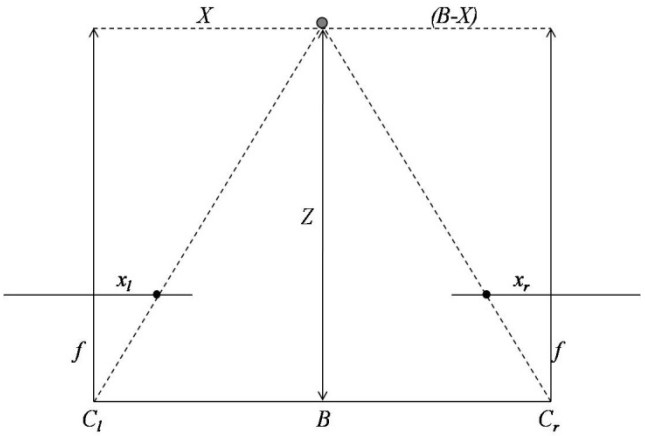

Fig. 3: Geometric model for a simple stereo system.

From similar triangles,

$$Z/f = X/x_l \quad (2)$$

$$z/f = (x - B)/x_r \quad (3)$$

In solving for the depth of point *X*, we have introduced the notion of disparity *d* in Eq. (4), which is the difference between the image coordinate x_l and x_r in the left and right. Solution of these equations yields all three coordinates completely locating point *X* in 3-D space. Eq. (4) clearly shows that the distance to point *X* increases as the disparity decreases and decreases as the disparity increases. By construction of the simple stereo imaging system, there is no disparity in the two *y* coordinates.

$$Z = fB/(x_l - x_r) = fB/d \quad (4)$$

Figure 3 shows a single point *X* being located in 3-D space so there is no problem identifying the matching image point x_l and x_r. Determining these corresponding points can be very difficult for a real 3-D scene containing many occlusions and discontinuous areas because it is often unclear which point in the left image corresponds to which point in the right image.

3.2 Multi-view Depth from a Single Depth Using Multi-baseline Information

The most popular configuration for stereo imaging is to use two cameras with parallel imaging planes that are located on the same X−Y plane of the world coordinate. In this paper, we extend stereoscopic system to multi-baseline stereo to get the multiple depth images.

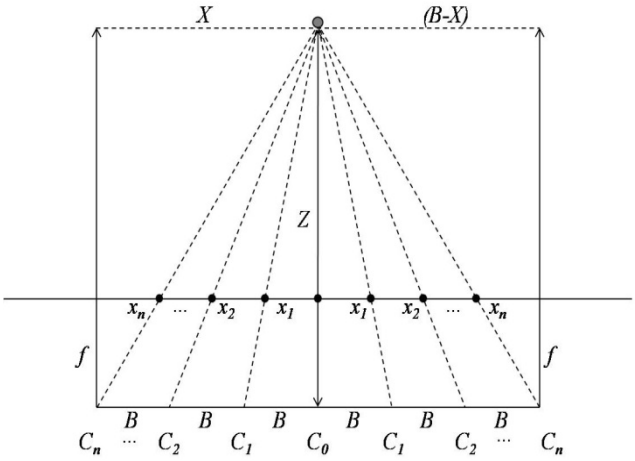

Fig. 4: Geometric relationship of multiple cameras.

Figure 4 depicts the geometry structure for the multi-view camera system when the reference camera is camera C_0. We assume that all cameras C_n for i from 0 to n are equally spaced and the pin-hole model. The cameras C_n for i from 1 to n will be ordered by their distance from the reference camera, i.e., C_1 is the closest camera to C_0.

$$Z/f = X/nx_n \quad (5)$$

$$z/f = (x - nB)/x_n \quad (6)$$

The disparity value has only a horizontal component in this case and can be derived by the preceding relation based on Fig. 4.

$$d_n = f\, nB/Z \quad (7)$$

Eq. (7) plays an important role in generating multi-view depth. Since we already have depth information *Z*, focal length f, and baseline *nB*, the disparity value *d* can be calculated directly. It increases linearly as the distance from the reference camera C_0 to other cameras. This relation forms the basis for deriving depth values at the each camera position. Since we assume that there are no vertical displacements, we only have horizontal disparities in this case. Thus, the corresponding points in the left and right images are on the same horizontal line.

Each pixel of the depth map at C_0 is moved to other camera locations by d_n. Since the d_n is varying based on *Z*, pixels having the same depth value are transferred by the same distance. Therefore, we can get multiple depth maps for each camera position after moving pixels of the single depth map at the reference camera position. However, generated depth maps have holes caused by the occlusion. Moreover, some pixels have incorrect depth values compared with the ground truth.

3.3 Hole Filling and Depth Correction

In order to fill holes by occlusion, we use the neighboring images generated in the first step. Figure 5 shows the generated depth images. The multi-depth generation is performed from the reference camera position to other camera locations. When the process is carried out from the reference camera to the left cameras (camera 0, 1, 2, and 3), major holes are created along the left side of objects. On the other hand, holes are mainly distributed in the right side of objects as the generation is done from the reference view to the right cameras (camera 5, 6, 7, and 8).

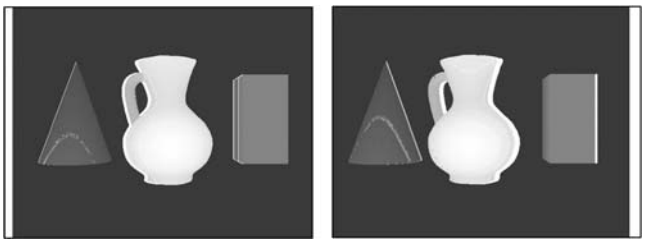

Fig. 5: Generated depth map with holes.

Our approach is to use neighboring pixels for interpolating empty pixels of the current depth maps. The pixel interpolation is performed by

$$I_s(x,y) = \frac{1}{k}\sum_{i=0}^{w}\sum_{j=0}^{w} I(R_{(i,j)}) \quad (8)$$

where $I_S(x, y)$ is the interpolated pixel value at the (x, y) position using neighboring pixels of the current depth map, k is the valid number of pixels within a W x W window, and R means the neighboring pixels. We use these equations to interpolate the empty pixels of the depth map.

Figure 6 shows the difference image with the generated depth map and the ground truth. When the camera distance is far from the center position C_0, disparity errors are increased. Therefore, we correct the disparity value using a color matching method. First, we find the same depth map group because when the depth value is same, the disparity is also same. Then, we calculate the shift value of the smallest absolute intensity differences (AD) to correct the disparity error of each depth value [2]. In order to generate the depth map accurately, we add this correction value to the previous disparity.

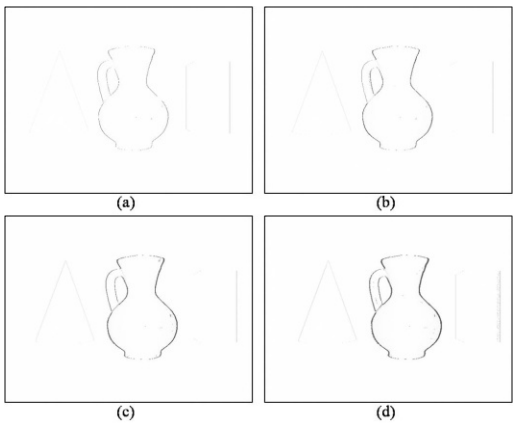

Fig. 6: Difference images: ground truth and generated depth: (a)~(d) for camera 0 ~ 3.

4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed method, we need a quantitative way to estimate the quality of the computed depth maps. Two general approaches to this are to calculate error statistics with respect to some ground truth data and to evaluate the synthetic images obtained by warping the reference or unseen images by the computed depth map. In this paper, we compute the following three quality measures based on known ground truth data.

$$B_O = \sum_{s\in O}(|d(s) - d_T(s)| > \delta_d) \quad (9)$$

$$B_T = \sum_{s\in T}(|d(s) - d_T(s)| > \delta_d) \quad (10)$$

$$B_D = \sum_{s\notin D}(|d(s) - d_T(s)| > \delta_d) \quad (11)$$

For quantitative evaluation, Scharstein and Szeliski measure the percentage of wrong pixels, i.e. pixels whose absolute disparity error is larger than one [8]. Error percentages are computed in four different image regions. First, they determine the percentage of wrong pixels over the non-occluded image. Then they estimate the percentages of bad pixels in untextured regions and close to disparity discontinuities, since those image areas are specifically challenging in stereo computation. In all three cases, only unoccluded pixels are considered. This is due to the reason that most stereo methods are not able to produce meaningful depth estimates for points affected by occlusion. Recently, the all region including half-occluded region is also considered for quantitative statistics. The evaluated results are then listed according to these errors.

The main part of generation multiple depth images from a single depth map is the multi-baseline stereo construction. Figure 7 depicts input images at multiple camera locations to measure the performance of our depth generation algorithm. The reference camera is the camera 4. In order to evaluate the performance of the algorithm, experiments were carried out using the depth images from the computer graphics (CG) data. These test sequences also provide ground truth depth maps. Each image sequence consists of nine images, taken at regular intervals with the camera pointing perpendicularly to the direction of motion.

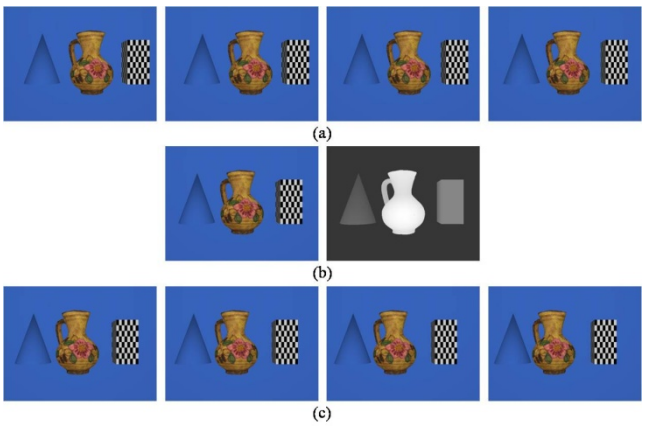

Fig. 7: Multi-view color images with a depth map: (a) data set for camera 0~3, (b) reference camera 4 with a depth map, (c) data set for camera 5~8.

The performance of our depth generation algorithm was also confirmed by the quality measures proposed by Scharstein and Szeliski based on known ground truth data listed in Equation (9) ~ (11). B_T reveals the performance in textureless regions and B_D shows the performance in depth discontinuity regions. In particular, B_O represents the overall performance of the proposed algorithm. Table 1 shows the results of applying our algorithm. Our results with multi-depth generation using a depth map are shown in the high level of performance, respectively. We also adopted the peak signal-to-noise ratio (PSNR) to evaluate the quality of the generated depth map image. In general, a processed image is acceptable to the human eyes if its PSNR is greater than 30 dB.

Table 1: Quantitative statistics based on the ground truth.

	B_O	B_T	B_D	*PSNR*
Camera 0	1.25	1.62	6.68	32.51
Camera 1	1.21	1.75	6.39	32.92
Camera 2	1.11	1.37	5.79	33.16
Camera 3	0.88	1.29	4.76	33.81
Camera 5	0.97	1.75	5.09	33.25
Camera 6	1.29	1.71	6.83	32.92
Camera 7	1.30	1.57	6.92	32.37
Camera 8	1.39	1.64	6.85	32.12
Average	1.18	1.59	6.16	32.88

Figure 8 shows the generated multiple depth maps with holes. We can observe that objects are moving as the camera number changes. In order to identify the generating results clearly, we did not interpolate holes. White pixels in each image represent the holes generated by the initial depth map.

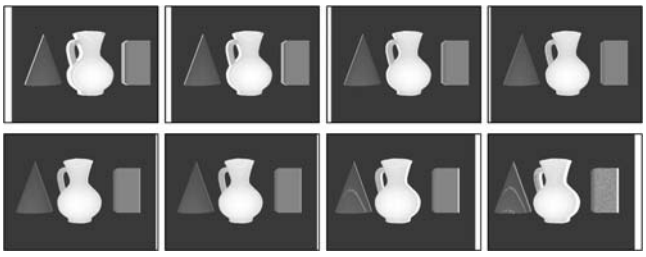

Fig. 8: Generated multiple depth maps with holes: camera 0 ~ 8 except for camera 4.

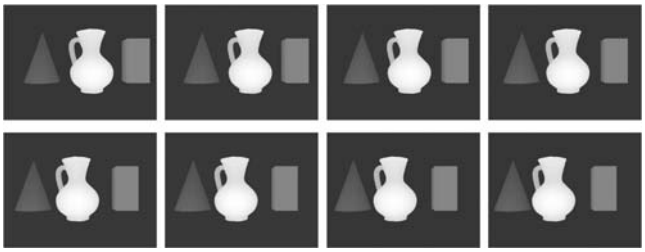

Fig. 9: Ground truth for camera 0 ~ 8 except for camera 4.

Figure 9 is the ground truth for each camera from CG data and Fig. 10 represents the final reconstruction results using the interpolation method. Although there are some artifacts in the final depth map results, it could be improved by applying more accurate methods of selecting proper information to fill the holes.

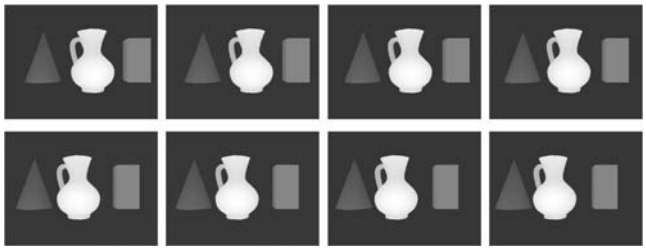

Fig. 10: Multiple depth maps with interpolation: generated depth maps for camera 0 ~ 8 except for camera 4.

The resulting depth maps are then compared with the corresponding ground truth. In addition, in order to show that the proposed algorithm is applicable to real camera systems, we add one test sequence “Home-shopping”, which is captured from our stereoscopic camera system. Since the data set does not include the ground truth data, we compare the results from proposed method in subjective aspects. To show the performance of the proposed depth generation algorithm, we add the results from our own data set which is captured using our high definition (HD) stereo camera system. The image size is 1920 x 1080.

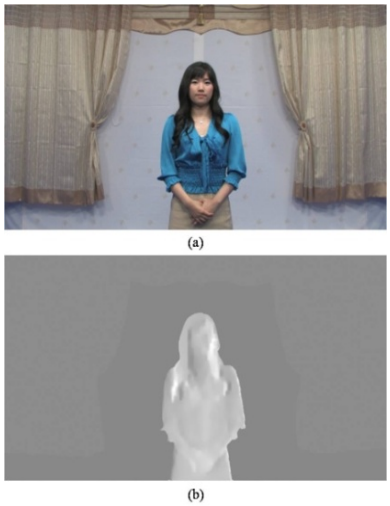

Fig. 11: Home-shopping: (a) the original color image for the left camera, (b) generated depth map for the left camera.

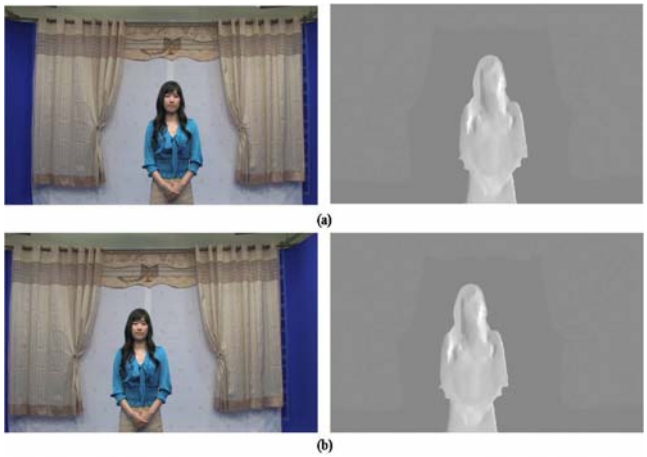

Fig. 12: Generated depth maps for Home-shopping dataset: (a) color and depth for the left camera, (b) color and depth for the right camera.

As shown in the Fig. 12, we generate high-quality depth map for the real camera system when the image is the high dimensional quality. Since there is no ground truth depth map for the test scene, quantitative statistics based on the known ground truth data set is not achievable.

Fig. 13: Generated reference disparity map

In order to generate the multi-view depth maps with a single depth map, we have to get one depth map. However, in real system it is very difficult to get the real depth map without active sensor depth camera system. The depth camera system has disadvantages: the high cost and limited viewing range. Therefore, we need a disparity map to generate multi-view depth maps without a real depth map. Figure 13 shows the result of the disparity map estimated by Zitnick's stereo matching algorithm [9]. Since the depth value can be calculated from the disparity, we can generate the approximated depth map using the disparity map.

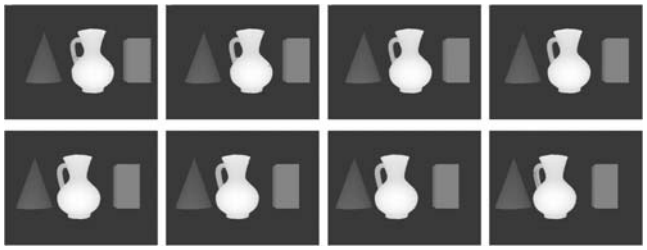

Fig. 14: Generated multiple depth maps based on the disparity map: camera 0 ~ 8 except for camera 4.

Figure 14 shows the results of the generated depth map with a high-quality disparity map and Table 2 shows the results of the quantitative statistics based on the disparity map. Since the quality of the generated depth map depends on the disparity map, most results are commonly decreased compare to previous quantitative results

Table 2: Quantitative statistics based on the disparity map.

	B_O	B_T	B_D	*PSNR*
Camera 0	1.53	1.74	6.72	30.31
Camera 1	1.25	1.62	6.58	31.22
Camera 2	1.21	1.75	6.42	32.56
Camera 3	1.13	1.37	6.39	32.81
Camera 5	1.29	1.71	6.43	32.65
Camera 6	1.35	1.57	6.52	32.22
Camera 7	1.49	1.64	6.74	31.57
Camera 8	1.62	1.81	6.85	30.32
Average	1.36	1.65	6.58	31.70

4. CONCLUSION

In this paper, we have proposed an algorithm for generation of multiple depth images from one depth map acquired from a depth camera and geometric relationships between multiple cameras. We have moved the reference depth map with different disparity values based on the multi-baseline information. Experimental results show that the generated multiple depth images are more accurate than those from other state-of-the art stereo matching algorithms. In addition, since our depth maps have been constructed in real-time, we believe that our algorithm is suitable for applications requiring multiple depth information like FTV.

ACKNOWLEDGMENT

This work was supported in part by the Ministry of Information and Communication (MIC) through the Realistic Broadcasting Research Center (RBRC), and in part by the Ministry of Education (MOE) through the Brain Korea 21 (BK21) project.

REFERENCES

[1] M. Tanimoto, "Overview of Free Viewpoint Televis ion," Signal Processing: Image Communication, Vol. 21, No. 6, pp. 454-461, July 2006.

[2] M. Okutomi and T. Kanade, "A Multiple-baseline Stereo," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 4, pp. 353–363, April 1993.

[3] ISO/IEC JTC1/SC29/WG11, "Call for Comments on 3DAV", N6051, October 2003.

[4] ISO/IEC JTC1/SC29/WG11, "Preliminary FTV Model and Requirements," N8944, April 2007.

[5] C. Fehn, R. de la Barre, and S. Pastoor, "Interactive 3DTV Concepts and Key Technologies," Proceedings of the IEEE, Vol. 94, No. 3, pp. 524-538, March 2006.

[6] S. Kim, S. Yoon, and Y. Ho, "Realistic Broadcasting Using Multi-Modal Immersive Media," Lecture Notes in Computer Science, Vol. 3768, pp. 164-175, November 2005.

[7] Y. Wang, J. Ostermann, and Y. Zhang, *Video Processing and Communications*, Prentice Hall, 2002.

[8] 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, April-June 2002.

[9] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, " High-Quality Video View Interpolation Using a Layered Representation," ACM Transactions on Graphics, Vol. 23, No. 3, pp. 600–608, August 2004.