J. Vis. Commun. Image R. 22 (2011) 73–84

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

# Generation of high-quality depth maps using hybrid camera system for 3-D video

Eun-Kyung Lee *, Yo-Sung Ho

*Gwangju Institute of Science and Technology (GIST), 1 Oryong-dong, Buk-gu, Gwangju 500-712, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 1 August 2009
Accepted 13 October 2010
Available online 25 October 2010

*Keywords:*
Depth map generation
Depth camera
Multi-view camera system
3-D video
3-D TV
Multi-view video
Stereo matching
View synthesis

ABSTRACT

In this paper, we present a hybrid camera system combining one time-of-flight depth camera and multiple video cameras to generate multi-view video sequences and their corresponding depth maps. In order to obtain the multi-view video-plus-depth data using the hybrid camera system, we capture multi-view videos using multiple video cameras and a single view depth video with the depth camera. After performing a three-dimensional (3-D) warping operation to obtain an initial depth map at each viewpoint, we refine the initial depth map using segment-based stereo matching. To reduce mismatched depth values along object boundaries, we detect the moving objects using color difference between frames and extract occlusion and disocclusion areas with the initial depth information. Finally, we recompute the depth value of each pixel in each segment using pairwise stereo matching with a proposed cost function. Experimental results show that the proposed hybrid camera system produces multi-view video sequences with more accurate depth maps, especially along the boundary of objects. In addition, it is suitable for generating more natural 3-D views for 3-D TV than previous works..



## 1. Introduction

As three-dimensional (3-D) video becomes attractive in a variety of 3-D multimedia applications, it is essential to obtain multi-view video sequences with corresponding depth maps, which are often called as multi-view video-plus-depth data [1]. In near future, consumers will be able to experience 3-D depth impression and choose their own viewpoints in the immersive visual scenes created by 3-D videos. Recently, the ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) has recognized the importance of the multi-view video-plus-depth data for free-viewpoint TV (FTV) or 3-D TV [2], and has investigated the needs for standardization on 3-D video coding [3,4].

With respect to the current 3-D TV and FTV research activities, it is important to estimate accurate depth information from real natural scenes. Although various depth estimation methods have been developed in the field of computer vision, accurate measurement of depth information from natural scenes still remains problematic.

In general, depth estimation methods can be classified into two categories: passive depth sensing and active depth sensing. The former calculates depth information indirectly from 2-D images captured by two or more video cameras. Typical examples include shape from focus [5] and stereo matching [6]. The advantage of indirect depth estimation is a low price because we can create depth maps using cheap off-the-shelf video cameras. However, accuracy of the depth maps is relatively lower than those produced from active approaches in occlusion and textureless regions.

On the other hand, active depth sensing methods usually employ physical sensors, such as laser, infrared ray (IR), or light pattern, to obtain depth information from natural scenes directly. Structured light patterns [7] and depth cameras [8–11] are major examples of these approaches. Although currently available direct depth estimation tools are quite expensive and support low-resolution depth maps only, they can produce more accurate depth maps in a short time.

For instance, we can obtain depth maps of natural scenes in real time using active range depth cameras. They capture color images and their associated per-pixel depth information simultaneously by integrating a high-speed pulsed IR light source into a conventional broadcast TV camera [8]. However, even though they can capture depth values directly in real time, there are crucial disadvantages in the currently available depth camera systems. They only produce low-quality depth maps with optical noises.

To solve those problems, fusion camera systems which consisted of multiple video cameras and one or more time-of-flight (TOF) camera have been introduced [12,13]. Zhu et al. [14] presented a calibration method to improve depth quality using a TOF depth sensor. They used the probability distribution function of the depth information measured by the TOF depth sensor and provided a more reliable depth map. Lee et al. [15] enhanced the depth resolution and accuracy by combining the actual distance information measured by the depth camera with the disparity map estimated by the passive depth sensing method. However,

* Corresponding author. Fax: +82 62 970 3164.
*E-mail addresses:* eklee78@gist.ac.kr (E.-K. Lee), hoyo@gist.ac.kr (Y.-S. Ho).

1047-3203/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2010.10.006

the previous fusion systems have produced only low-resolution depth maps and focused on generating depth maps of static 3-D scenes.

Since forthcoming 3-D multimedia applications are expected to use high-quality and high-resolution 3-D videos, we need to create multi-view video-plus-depth data with high quality. In this paper, we devise a hybrid camera system with one depth camera and multiple high-definition (HD) video cameras. The proposed camera system can produce high-resolution multi-view depth maps for dynamic 3-D scenes by enhancing the low-resolution depth information measured by the depth camera. The main contribution of our work is to propose a practical solution to generate high-quality depth maps for dynamic 3-D scenes.

The remainder of this paper is organized as follows. In Section 2, we present the overall architecture of the proposed hybrid camera system. Section 3 describes preprocessing steps for enhancing depth maps and Section 4 presents how to generate the multi-view video sequences with their corresponding depth maps using the proposed camera system. After showing experimental results in Section 5, we make conclusions in Section 6.

## 2. Hybrid camera system

The proposed hybrid camera system is composed of one standard-definition (SD) depth camera and five high-definition (HD) video. Those multiple video cameras are arranged in a one-dimensional array to construct a multi-view camera system. A clock generator sends synchronization signals constantly to each camera and its corresponding personal computer equipped with a video capture board. Basically, the proposed hybrid camera system captures multi-view images by the multiple video cameras and a depth map from the depth camera at each sampling time.

Fig. 1 illustrates the overall framework to generate multi-view video sequences with their corresponding depth maps using the hybrid camera system. After calibrating each camera independently, we perform an image rectification to adjust vertical mismatches in multi-view images. Then, we apply a color correction operation to maintain color consistency among multi-view images. Fig. 2 shows the proposed hybrid camera system and it's configuration.

To obtain depth maps for multi-view images, we perform a 3-D warping operation onto each multi-view camera using the depth map measured by the depth camera. The warped depth data is used as an initial depth at each camera position. After we segment each multi-view image, we assign the depth value of the warped depth data in each segment as the initial depth of the segment. In order to improve the depth accuracy of object boundaries, we separate the moving objects and detect occlusion and disocclusion regions. Then, the depth of each segment is refined by a color segmentation-based stereo matching method. Finally, we obtain multi-view depth maps by conducting a pixel-based depth map refinement using a proposed cost function in each segment.

## 3. Preprocessing for depth enhancement

### 3.1. Relative camera calibration

Since the proposed hybrid camera system consists of two different types of cameras, a depth camera and multiple HD video cameras, it is essential to find out relative camera information through camera calibration [16]. For that, we apply a camera calibration algorithm [17] to each camera in our camera system and obtain projection matrices for the depth camera and each video camera.

$$P_s = K_s[R_s|t_s] \quad (1)$$

$$P_k = K_k[R_k|t_k] \quad (2)$$

where $P_s$ is the projection matrix of the depth camera represented by its intrinsic matrix $K_s$, rotation matrix $R_s$, and translation vector $t_s$. $P_k$ indicates the projection matrices of the $k$th video camera which consisted of its intrinsic matrix $K_k$, rotation matrices $R_k$, and translation vector $t_k$. We then employ a multi-view rectification operation [18]. The multi-camera array have geometric errors because they are set manually by hand. In order to minimize the geometric errors, we find the common baseline, and then apply

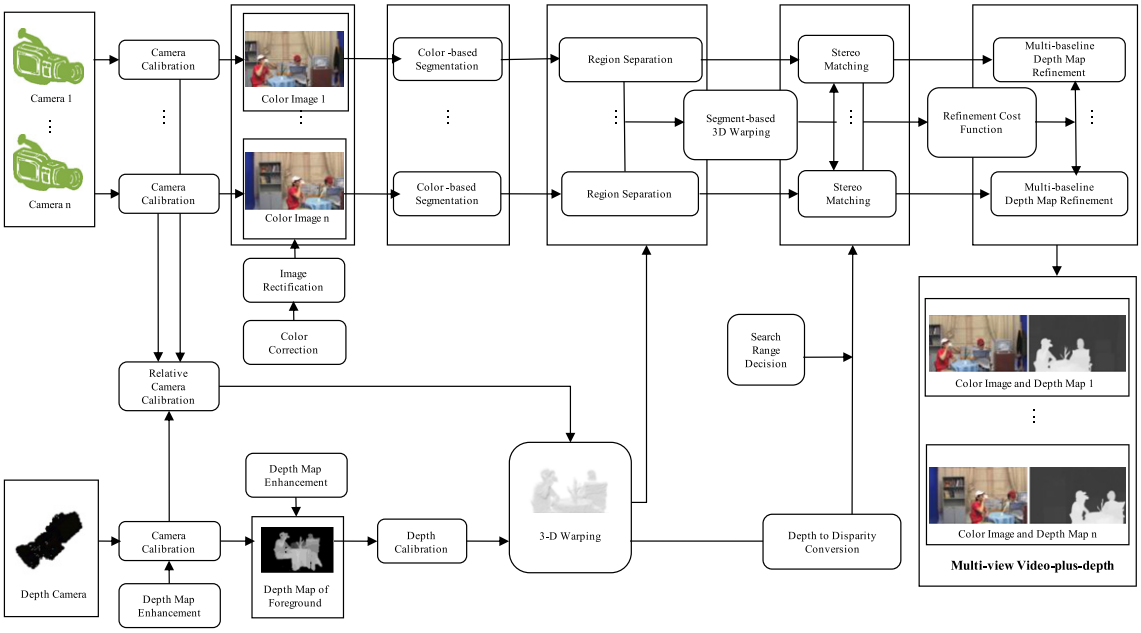



**Fig. 1.** Overall architecture of the proposed 3-D video generation system.

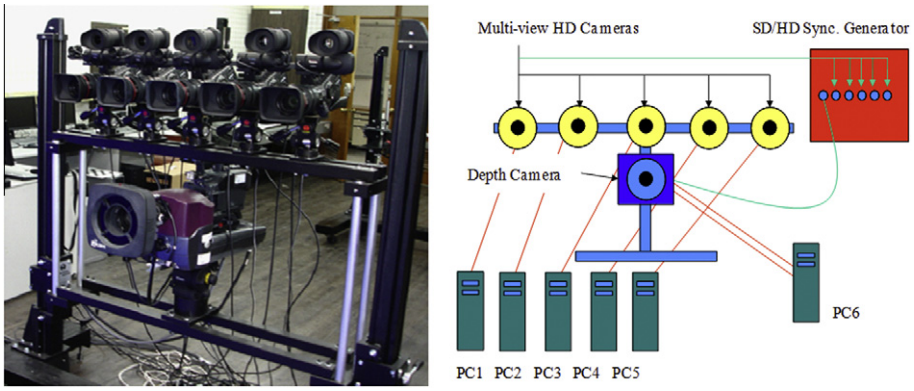



**Fig. 2.** The proposed hybrid camera system.

the rectifying transformation to the multi-view image. Consequently, the projection matrix of video cameras is changed as

$$P'_k = K'_k[R'_k|t'_k] \tag{3}$$

where $K'_k$ and $R'_k$ are the modified camera intrinsic matrix and rotation matrix of the $k$th video camera, respectively. Thereafter, we convert the rotation matrix $R_s$ of the depth camera into the identity matrix $I$ by multiplying inverse rotation matrix $R_s^{-1}$. The translation vector $t_s$ of the depth camera is also changed into the zero matrix $O$ by subtracting the translation vector $t_s$. Hence, we can define new relative projection matrices for the multi-view cameras on the basis of the depth camera as

$$P'_s = K_s[I|O] \tag{4}$$

$$\tilde{P}'_k = K'_k\left[R'_k R_s^{-1}|t_k - t_s\right] \tag{5}$$

where $P'_s$ and $\tilde{P}'_k$ are final projection matrices of the depth camera and the $k$th video camera, respectively. After relative camera calibration, we resolve the color mismatch problem of multi-view images using a color calibration method [19]. The color characteristics of captured images are usually inconsistent due to different camera properties and lighting conditions even the hardware type and specification of the multiple cameras are the same. Thereafter, we perform bilateral filtering to reduce optical noises included in the depth map acquired from the depth camera [20].

### 3.2. Depth calibration

The depth values measured by the depth camera are very sensitive to noises. Their sources are diverse including physical limitation of hardware and specific object properties, etc. Therefore, depth data are noticeably contaminated with random and systematic measurement errors dependent on reflectance, angle of incidence, and environmental factors like temperature and lighting [21]. To reduce those errors, we employ a depth calibration method [14].

For depth calibration in indoor environments, we compute the depth of the planar checker pattern within the limited space by increasing the distance from the image pattern to the depth camera using our system as shown in Fig. 3. To extract the corresponding feature points in two different types of cameras efficiently, we use the color checker pattern. The pattern image is captured in every 10 cm distance. The plane pattern is orthogonal to the image plane.

Thereafter, we make a four dimensional look-up table (LUT) mapping 3-D positions of the multiple video cameras and the depth value from the depth camera. 3-D position is constructed by $x$, $y$ position of the feature point and the real depth value $z$ calculated from the multi-view image by pairwise stereo matching. Since we have already obtained camera parameters, the real depth value is calculated by

$$D_n(p_x, p_y) = \frac{K \cdot B}{d_n(p_x, p_y)} \tag{6}$$

where $K$ is the focal length of the center camera, *Camera 3*, and $B$ is the baseline distance between three neighboring video cameras, *Camera 2*, *Camera 3*, and *Camera 4*. Since we have rectified the multi-view image, the baseline $B$ between neighboring cameras is the same [18]. $D_n(p_x, p_y)$ is the real depth value corresponding to the measured disparity value $d_n(p_x, p_y)$ at the pixel position $(p_x, p_y)$ in the checker pattern. To reduce the depth error, we use mean disparity value between disparity from *Camera 2* and *Camera 3* and that from *Camera 3* and *Camera 4*.

To check the accuracy of the calibrated depth value, we perform 3-D warping to the HD camera. Fig. 4(a) is the 3D warping result using the acquired depth map and Fig. 4(b) shows that of the calibrated depth map using the LUT. While there are many mismatched depth values in Fig. 4(a), most of them are correctly matched in the boundaries of the rectangular box in Fig. 4(b). The other problem is that even though the distance from the depth camera to the object is constant, depth information from the depth camera can be different depending on the object color and lighting conditions.

To analyze the depth sensitivity of a static object in the dynamic scene, we check the depth values of a black cup, as shown in Fig. 5. We can notice the inconsistent depth value changes of the static object caused by object movement and material properties. Especially, the depth value of the dark color region measured by the depth camera is very unstable and unreliable. The black cup has to sustain a near-constant depth in the scene; however, the acquired depth values are unpredictable and random. The reason is that dark or black colors absorb light of all frequencies and the depth camera uses near IR rays.

Although we perform the depth calibration to correct the acquired depth map, there are still limitations in the depth values acquired from the depth camera. To obtain the high-quality multi-view depth maps, we need to refine the acquired depth value using an efficient stereo matching algorithm.

## 4. 3-D video generation

### 4.1. Initial depth computation of the multi-view image

We generate initial depth of the multi-view image by performing 3-D warping of the depth values obtained from the depth camera. First, we project pixels of the depth map into the 3-D world

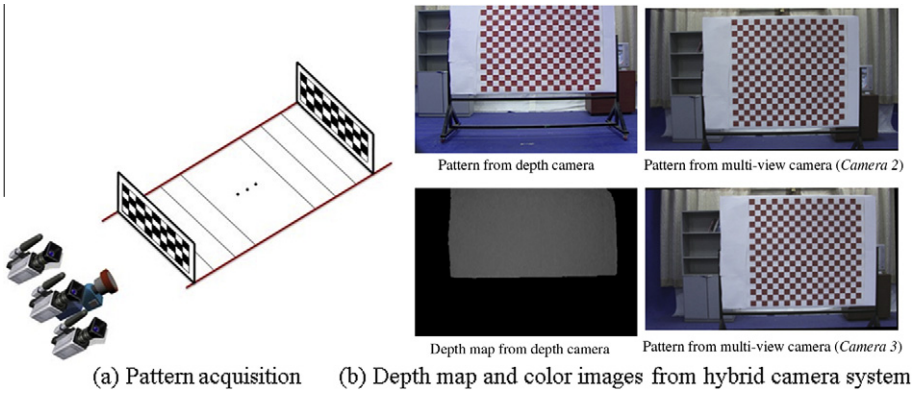



**Fig. 3.** Acquisition of the planar check pattern for depth calibration.

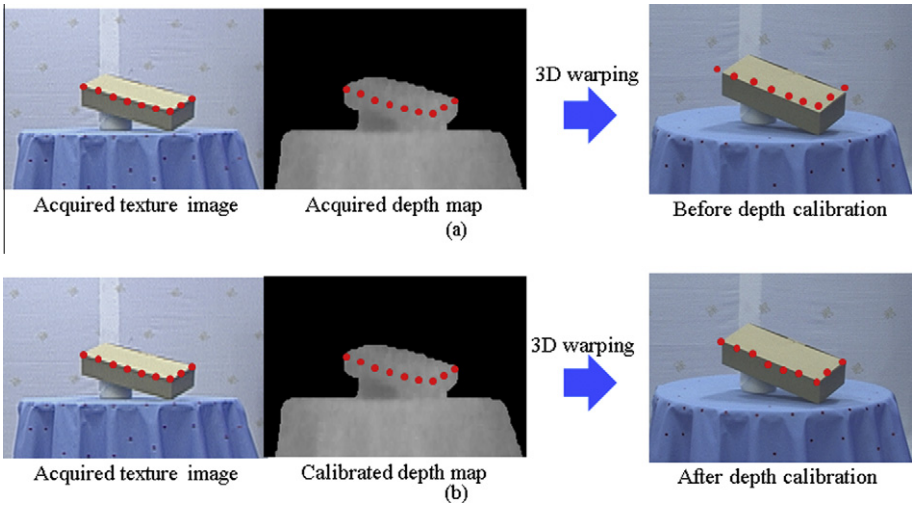



**Fig. 4.** Depth accuracy test using the acquired depth map and calibrated depth map.

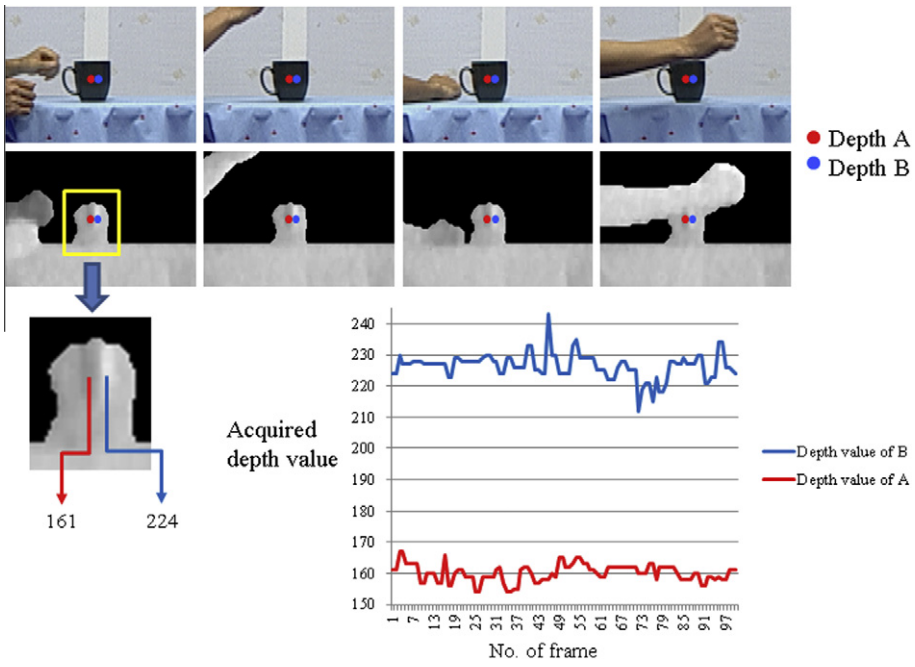



**Fig. 5.** Depth inconsistency of a static scene.

**Fig. 6.** 3-D warped depth map.

coordinate using the depth values. We then reproject the 3-D points into each view. Let us assume that $D_s(p_{sx}, p_{sy})$ is the depth intensity at the pixel position $(p_{sx}, p_{sy})$ in the depth map. $P_s((p_{sx}, p_{sy}, D_s(p_{sx}, p_{sy}))$ is a 3-D point corresponding to $D_s$. The backward projection for moving $D_s$ to the world coordinate is carried out by

$$P_s = K_s^{-1} \cdot p_s \tag{7}$$

where $K_s^{-1}$ indicates the intrinsic matrix of the depth camera. In the backward 3-D warping, since rotation and translation matrices of the depth camera are the identity matrix $I$ and zero matrix $O$ in Eq. (4), respectively, we only consider its intrinsic matrix. Thereafter, we project the 3-D points $P_s$ into the each view to get its corresponding pixel position $p'_k(u_k, v_k)$ of the $k$th-view image by

$$p'_k = \tilde{P}'_k \cdot P_s \tag{8}$$

where $P'_k$ indicates the projection matrix of the $k$th-view video camera. Fig. 6 shows the result of 3-D warping using the acquired depth maps.

### 4.2. Region separation

To estimate depth maps of multiple video cameras using the warped depth information, we segment the multi-view image by a mean-shift color segmentation algorithm [22]. However, we cannot control the maximum segment size because there is no parameter to control the maximum segment size.

When we perform the segment-based stereo matching, one segment has one depth value. If the size of segment is too large, we cannot get a smooth depth map. The other way, if the size of segment is too small, it is hard to overcome textureless problem during the stereo matching. To solve this problem, we split one image into $16 \times 16$ block segments, so that we can limit the maximum segment size.

Fig. 7 shows the procedure of the segment merging. A block can have two or more color segments. Before merging the segment, we split the segmented image into block-based segment again. If each segment is smaller than half size of the block, we merge it into one segment by searching adjoined blocks to find the same indexed segment. If the size of the merged block is larger than threshold, the merging procedure is finished; otherwise we repeat the same process until merging condition is satisfied.

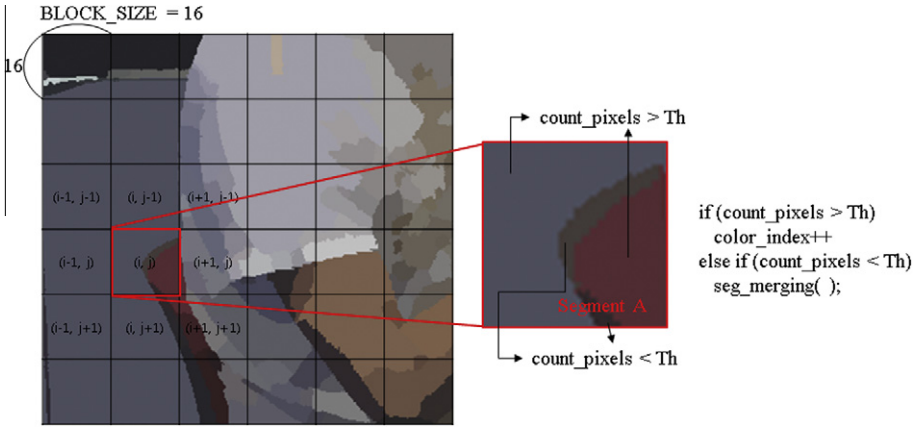



**Fig. 7.** Block-based segment merging.


**Fig. 8.** Segmentation results in the temporal domain.

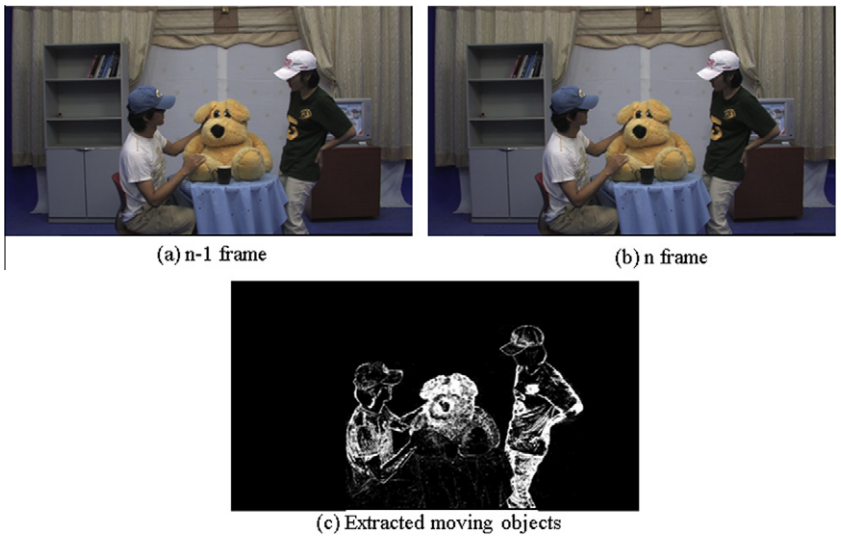



**Fig. 9.** Moving object detection using color difference between frames.

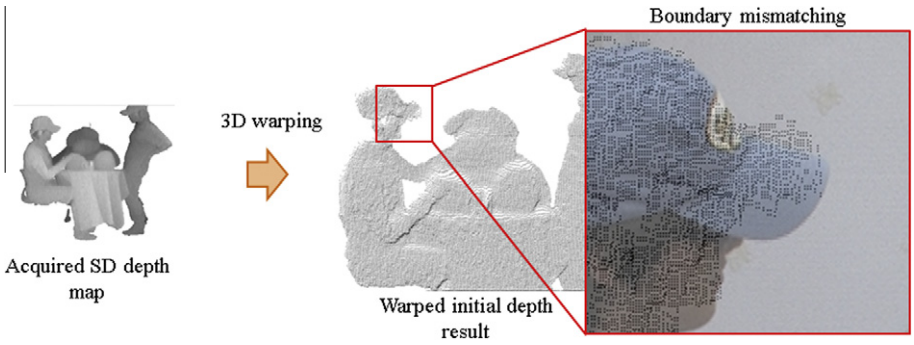



**Fig. 10.** Boundary mismatching problem.

The searching order of connected blocks is right, bottom, left, and top including the diagonal directions because left and top blocks are merged before and right and bottom blocks will be merged. For example, *Segment A* is divided into many block-based segments and *Block* $(i, j)$ have four segments. Since the size of *Segment A* in *Block* $(i, j)$ is smaller than the predefined threshold value in Fig. 7, the same indexed segment of *Segment A* is the block in $(i, j + 1)$ by the searching order. We merge the current *Segment A* and the same indexed segment in $(i, j + 1)$.

Before we estimate depth maps, we separate moving object using color difference between frames. To extract the moving object in the current frame, we calculate color differences between the previous frame $n - 1$ and the current frame $n$ by using the threshold which indicates the current position is foreground or not. We cannot directly use the segment-based moving object detection because shape of each segment can be varied in the temporal domain as shown in Fig. 8.

Since color segmentation is performed frame by frame, it is hard to find the same segment in the temporal domain. Therefore, we use the Euclidean distance between frames to extract the moving objects as

$$E_n(x,y) = \sqrt{(R_{n-1}(x,y) - R_n(x,y))^2 + (G_{n-1}(x,y) - G_n(x,y))^2 + (B_{n-1}(x,y) - B_n(x,y))^2} \quad (9)$$

where $R$, $G$, and $B$ indicate the pixel values in RGB color domain. To find the moving object, we compute the $E_n(x, y)$ at each pixel location for all pixels. If we subtract the RGB value between frames, camera noises can be mixed up. To remove them, we calculate the average RGB value for $3 \times 3$ block. If the average is larger than

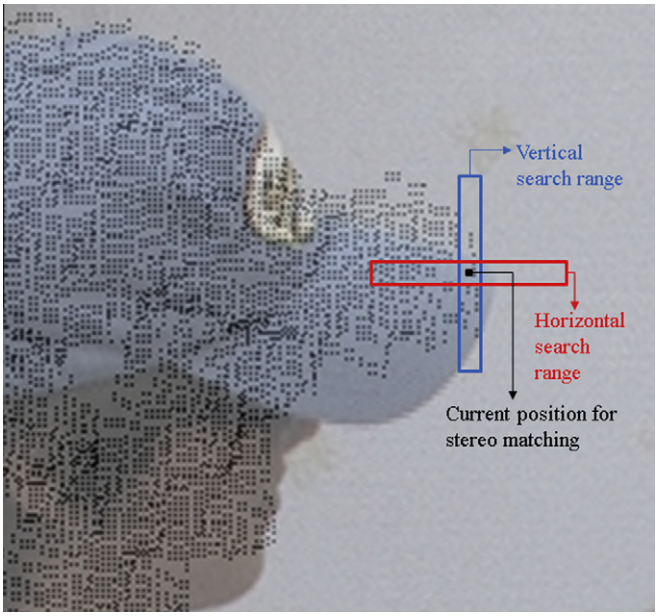



**Fig. 11.** Set of multiple initial depth values.


**Fig. 12.** Results of occlusion and disocclusion detection in *Camera 3*.

the threshold value, we set the center pixel of each $3 \times 3$ block as the foreground pixel. From our experiments, the threshold value of Euclidean distance, 10 is used. Fig. 9 (a) and (b) present 78th frame and 79th frame images in *Camera 3* and Fig. 9(c) shows the result of the extracted moving objects.

### 4.3. Segment-based multi-view depth estimation

We define the initial depth of each segment as 3-D warped depths in the segment; the assumption is that each segment has one depth value [6]. However, there is one problem to set the initial depth using warped depth value. The 3-D warping is performed from the small resolution depth map to the HD image in our system. Since there are many errors such as camera calibration error and depth error acquired from the depth camera, the warped result is not exactly matched with the HD image as shown in Fig. 10.

To obtain the accurate initial depth value, we use the warped results as multiple initial depth values for stereo matching. If we start the stereo matching with the initial depth, we can reduce the search range for finding the matched region. In addition, depending on search range reduction, we can overcome the mismatched problem in the textureless regions. However, if the given initial depth is the error value, we could find wrong areas which has local minimum. Therefore, the assignment of the correct initial depth is crucial in using the depth camera. Because there are correct initial depths around the currently warped position, which are not exactly matched with the original image, we increase the candidates of the initial depth value to resolve this problem. Fig. 11 shows the position of the initial depth in two directional regions, horizontal and vertical regions. One or more initial depth values usually exist in a $3 \times 3$ area because of the difference of the resolution. In this case, we set the horizontal search region as $30 \times 5$ and the vertical search region as $5 \times 30$. By using the multiple initial depths, we can set initial depth for the depthless regions in the boundary of objects as shown in Fig. 11.

To increase the depth accuracy for stereo matching, we utilize a pairwise stereo matching method. When the current view is *Camera 1*, there is no left image. Therefore, we use the input images of *Camera 2* and *Camera 3* for current view. For input image of *Camera 3*, we use the value of the initial depth multiply by 2. When we perform the stereo matching operation twice with the left and right images for one depth, we can find the occlusion and disocclusion regions. If some regions are not observed in one view while they are visible in the other views, those areas are occluded in one view and disocclued in the other views. From the fact, we can determine the reliable and unreliable regions. After getting corresponding values with the multiple initial depth information in small search range, we also determine the occlusion and disocclusion regions using the calculated depth value.

Since stereo matching measures the difference between the corresponding points of two or more images, called as the disparity, we convert the initial depth into its disparity for stereo matching by

$$InitDisp(x,y) = \frac{K \cdot B}{InitDepth(x,y)} \tag{10}$$

where $InitiDisp(x,y)$ is the converted disparity at the pixel position $(x,y)$ from the corresponding initial depth $InitDepth(x,y)$. $B$ and $K$ are the distance between neighboring video cameras and the focal length of the current video camera, respectively. After performing stereo matching with the initial disparity, we convert again the calculated disparity into its depth value to produce the depth map. Before performing bi-directional stereo matching, we need to set the candidate of the initial depth value. For determining the disparity of each segment, we calculate the mean of absolute difference (MAD) values between the segment in the current view image and its matched region in the left and right view images by

$$FG\_d_i(InitDisp) = min\left(min\left(\sum_{j=0}^{a} MAD(j)\right), min\left(\sum_{k=0}^{b} MAD(k)\right)\right) \tag{11}$$

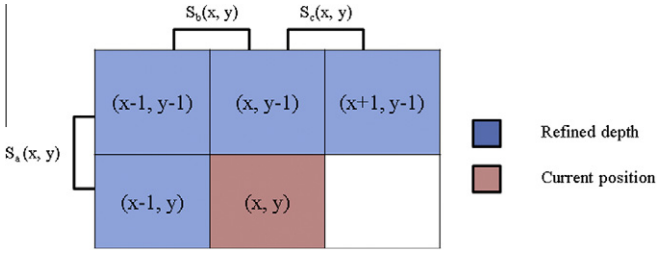
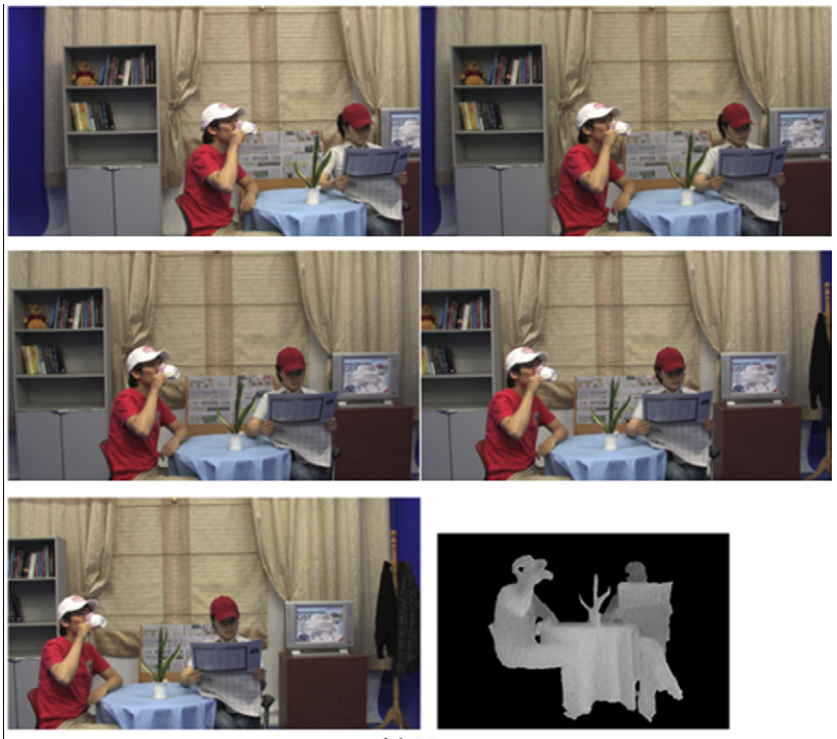


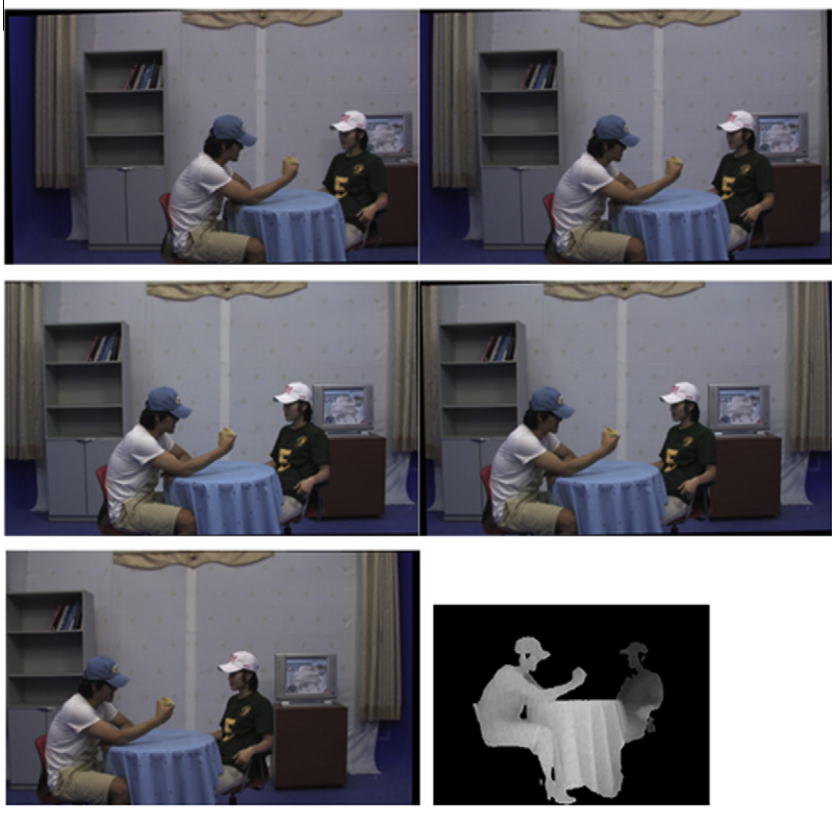

**Fig. 13.** Smoothness with gradient of the refined depth values.

**Table 1**
Specification of the hybrid camera system.

| Device | Specifications | Details |
|---|---|---|
| Stereo camera (Cannon XL-HI) | Output format | NTSC or PAL (16:9 ratio, high definition) |
| Depth camera (Zcam) | Depth range<br>Field of view<br>Output format | 0.5–7.0 m<br>40°<br>NTSC or PAL (4:3 ratio, standard definition) |
| Sync. generator (LT443D) | Output signal | SD/HD video generation |



(a) Newspaper



(b) Delivery

**Fig. 14.** Test sequences: multi-view image and its depth map.

where $i$ is the index of the segment, $j$ and $k$ means index of the multiple initial depth. $a$ and $b$ are the number of the initial depth in the horizontal and vertical regions, respectively. $FG\_d_i(InitDisp)$ is the refined initial depth value from pairwise stereo matching. Search range to estimate disparities of the current view image is from $InitDisp - 5$ to $InitDisp + 5$. The disparity with the minimum MAD in the

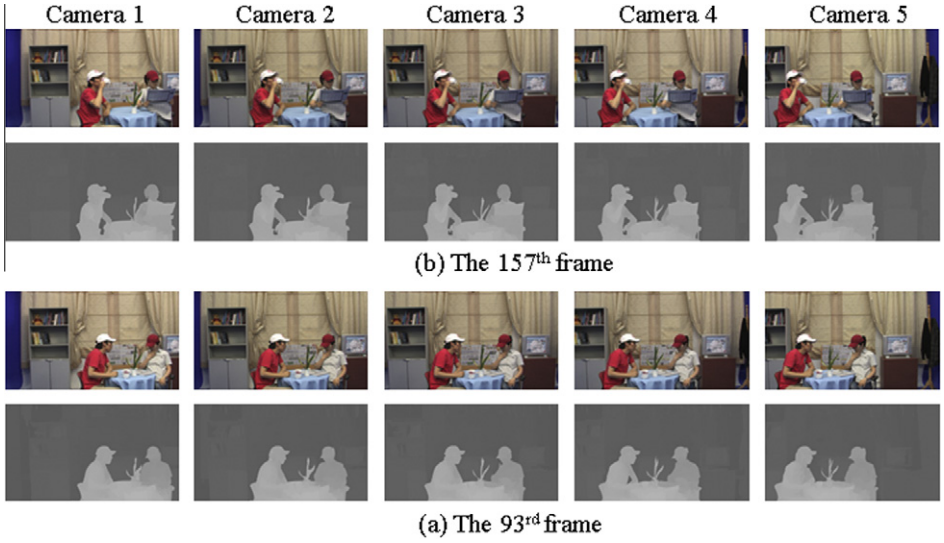



**Fig. 15.** Results of multi-view disparity map generation for *Newspaper*.

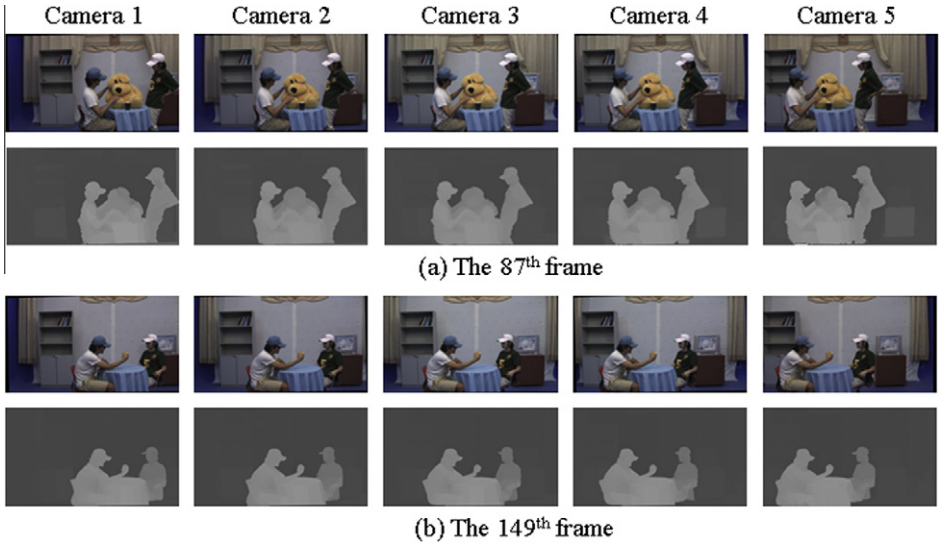



**Fig. 16.** Results of multi-view disparity map generation for *Delivery*.

search range is chosen as the refined initial disparity of the segment in the current view image. Since the acquired depth map is only for foreground regions, there is no depth information for background areas. We define that the background has no initial depth or the number of the included initial depth in the segment is less than 10% of the size of the segment. In estimating depth of background, we set the minimum and maximum depth/disparity value. We then find the minimum MAD as the initial disparity of the current segment in the background by

$$BG\_d_i(InitDisp) = min\left(\sum_{i=minDisp}^{maxDisp} MAD(i)\right) \quad (12)$$

where $BG\_d_i(InitDisp)$ is the disparity for background, *minDisp* and *maxDisp* mean minimum and maximum disparity search range for background. The disparity with the minimum MAD is chosen as the initial disparity $d_i(Initdisp)$ of the segment $i$ in the target view image $n$ by

$$d_i(InitDisp) = min(FG\_d_i(InitDisp), BG\_d_i(InitDisp)) \quad (13)$$

To detect the occlusion and disocclusion regions, we set the threshold value for reliable and unreliable areas. The threshold value of MAD is 20 in our experiments. Fig. 12 is the detection of the occlusion and disocclusion regions: the red circles mean the unreliable regions from the occlusion and disocclusion area.

### 4.4. Multi-view depth map refinement

In stereo matching, depth refinement usually enhances depth accuracy through iteration at the cost of long processing time, lots of memory requirement, and heavy computation. However, it has challenges when our target is to generate high-resolution 3-D video based on multi-view depth maps. We therefore propose a simplified depth refinement approach using the proposed cost function for the depth map refinement, which has the following features: low memory consumption, fast processing time, and no iteration steps.

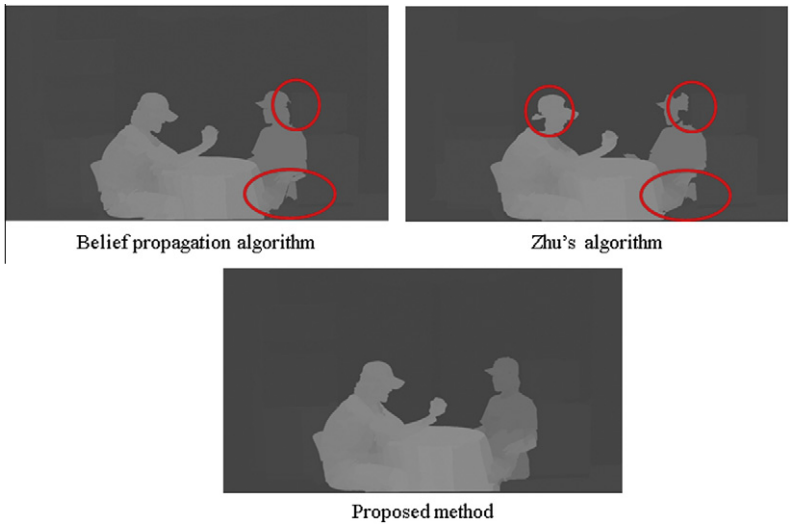



**Fig. 17.** Results comparison with the previous works.

**Fig. 18.** Results depth maps from 8th to 48th frame.

**Table 2**
Objective quality comparison of synthesized intermediate views.

| Sequence | Average PSNR | | |
|---|---|---|---|
| | Belief propagation | Zhu's algorithm | Proposed method |
| Delivery | 29.892 | 26.373 | 31.961 |
| Newspaper | 29.911 | 26.449 | 31.707 |

**Table 3**
Comparison of the processing time.

| Sequence | Processing time (s) | | |
|---|---|---|---|
| | Belief propagation | Zhu's algorithm | Proposed method |
| Delivery | 836.26 | 528.74 | 337.21 |
| Newspaper | 845.68 | 541.07 | 350.59 |

In order to enhance the multi-view depth map along the boundary of the objects, we refine it for two regions: moving region and static region. We have already defined the moving regions using color difference between frames as shown in Fig. 9. If there is no variance of a pixel in the time domain, we assume that pixel is static. In that case, we can refer the previous depth value for the static pixel. Otherwise, we just use the refined disparity value without referring the previous one.

$$E(x,y,d)=\begin{cases} w_s f_s(x,y,d_s(x,y))+w_d f_d(x,y,d_d(x,y)) \\ \quad if\ \ mov\_obj=1 \\ w_s f_s(x,y,d_s(x,y))+w_d f_d(x,y,d_d(x,y)) \\ \quad +w_t f_t(x,y,d_t(x,y)) \quad if\ \ mov\_obj=0 \end{cases} \tag{14}$$

where $w_s$, $w_d$, $w_t$ are the weighting factors for depth refinement. $f_s(x, y, d_s(x, y))$ is the smoothness term with gradient of the refined depth value in this refinement step. $f_d(x, y, d_d(x, y))$ is the data term for the refined initial depth value in the segment-based stereo matching step and $f_t(x, y, d_t(x, y))$ is the temporal term for depth value of the previous frame for the static pixel. From our experiments, the weighting factors of the cost function, $w_s$, $w_d$, and $w_t$ are 0.3, 0.5, and 0.2. $mov\_obj(x, y)$ indicates the result of the moving object detection. If $mov\_obj(x, y)$ is 0, this pixel is not moved. Then, we can refer the depth value of the previous frame.

$f_d(x, y, d_d(x, y))$ means the minimum MAD with the refined initial depth value in the search range from $InitDisp - 5$ to $InitDisp + 5$. $f_s(x, y, d_s(x, y))$ is the depth difference with neighborhood depth in the same segment and calculated by

$$f_s(x,y,d_s(x,y))=med(s_a(x,y),s_b(x,y),s_c(x,y)) \tag{15}$$

We can calculate the smoothness value as shown in Fig. 13. $s_a(x, y)$ is the refined depth difference at positions between $(x - 1, y - 1)$ and $(x - 1, y)$. $s_b(x, y)$ is the refined depth difference at positions between $(x - 1, y - 1)$ and $(x, y - 1)$. $s_c(x, y)$ is the refined depth difference at positions between $(x, y - 1)$ and $(x + 1, y - 1)$. The function $med()$ takes the median value among arguments to avoid the wrong

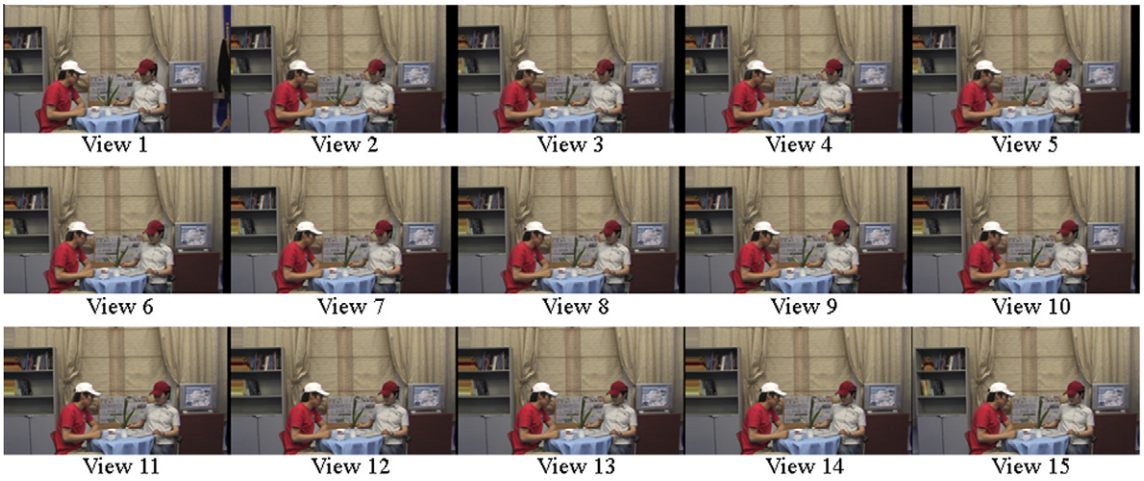



**Fig. 19.** Intermediate views using generated depth maps for *Newpaper*.

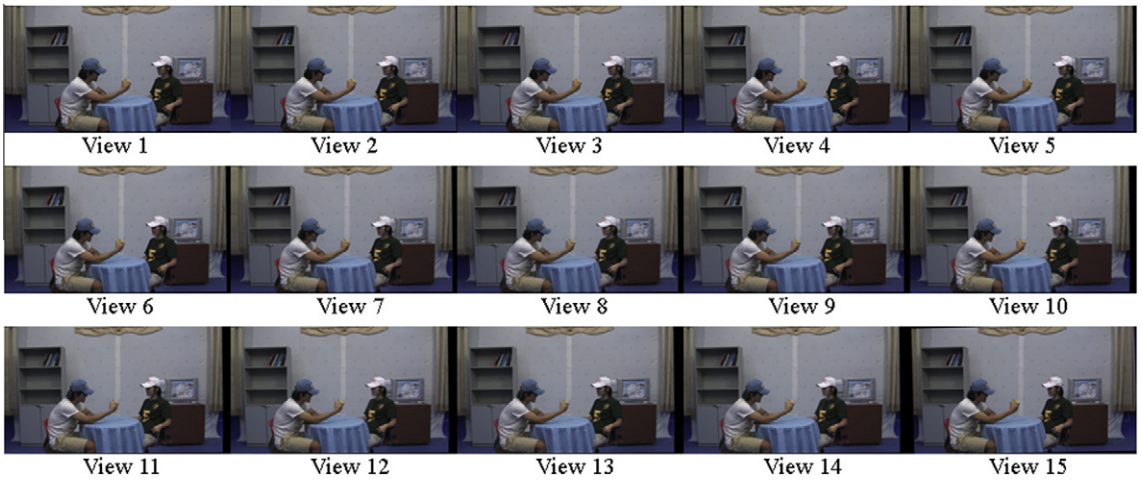



**Fig. 20.** Intermediate views using generated depth maps for *Delivery*.

depth selection, so that it maintains depth continuity along the vertical and horizontal direction. If the selected smoothness gradient is a vertical direction, this depth difference is calculated from $(x, y-1)$. Otherwise, the depth difference is computed from $(x-1, y)$.

## 5. Experimental results and analysis

In order to generate the high-quality multi-view depth maps, we have constructed a hybrid camera system with one depth camera and five HD video cameras. The specification of the hybrid camera system is shown in Table 1. The measuring distance of the depth camera was from 0.5 m to 7.0 m. The baseline distances among multi-view HD cameras are 20 cm. The proposed camera system's baseline distance depends on the physical volume of each HD video camera as shown in Fig. 1. Therefore, it is hard to reduce the baseline between HD cameras at current configuration. Fig. 14 shows the multi-view test sequences, *Newspaper* and *Delivery*, captured by the hybrid camera system.

In this paper, we use two types of sequences for generation of multi-view depth maps using the proposed hybrid camera system. In order to obtain the high-resolution depth maps, we capture the multi-view images from multiple video cameras and one depth map with a depth camera. Our experiments are performed for two types of sequences: one has small motion including complex objects and textureless regions in natural scenes; the other has fast motion changes in dynamic scenes. Since 3-D video contents have the depth impression in the scenes, we also have configured various depth differences from the white wall to the table. The resolution of the multi-view video sequences is full HD of $1920 \times 1080$, and the resolution of the depth maps is SD of $720 \times 486$.

Figs. 15 and 16 show the finally generated multi-view depth maps for the 93rd, 157th frames of *Newspaper* and 87th, 149th frames of *Delivery*. As shown in Fig. 15, we can observe that depths for the orchid in the flowerpot in the scene of *Newspaper* were generated successfully, although the boundary of the orchid is sharp. In addition, as shown in Fig. 16, the depth quality of the yellow bear doll was good, although the color of the bear was monotonous.

From the shown results, we have overcome the main problems of passive depth sensing: poor depth estimation on the occluded and textureless regions, based on the proposed hybrid camera system.

To compare the quality of depth map generated by the proposed method with previous works, we depicted the depth map generated by the BP algorithm [23] and Zhu's method [14] with the acquired initial depth information for the 3rd view image of the 93th frame in *Newspaper*. The generated depth maps using previous methods and the proposed one are shown in Fig. 17. We can check that some regions of the depth maps generated by the previous approaches had mismeasured depths, which are marked as red circles

in Fig. 17. This is because of the boundary mismatching problem as described in Section 4.3. However, the proposed method has overcome the problem using the two directional multiple initial depth values. From the result, the proposed method have outperformed the previous ones. Fig. 18 shows the result of the depth maps in *Camera* 3 from the 8th to 48th frame in every eight frames.

In order to measure the performance of our scheme objectively, we generated the synthesized views using the depth maps generated for *Camera 2* and *Camera 4*. Table 2 presents that the average peak signal-to-noise ratio (PSNR) of the synthesized images generated by the BP algorithm, Zhu's method, and the proposed method, respectively. The synthesized views produced by our hybrid camera system shows higher PSNR values than those obtained by the previous methods.

Table 3 shows the comparison of the processing time in the depth refinement step. Since each algorithm have different processing step to generate the depth map, it is hard to measure the exact processing time in the same condition. Therefore, we compare the processing time for the depth map refinement step. As shown in Table 3, the proposed method is faster than others without the accuracy reduction for depth map generation. From the result, it is useful for the high-resolution multi-view depth map generation.

We have also produced intermediate views using the finally depth maps and multi-view images using a view synthesis algorithm [24] for *Delivery* and *Newspaper*. In this experiment, we have generated 15 intermediate views between *Camera 3* and *Camera 4* by moving a virtual camera with one degree interval. As shown in Figs. 19 and 20, we could generate intermediate views successfully without noticeable artifacts subjectively.

## 6. Conclusions

In this paper, we have presented a new approach to generate multi-view HD depth maps corresponding to HD color images using the proposed hybrid camera system. We have used depth information acquired by a depth camera to generate the initial depth maps of multi-view images. We then have generated the final depth maps using a segmentation-based pairwise stereo matching and the proposed cost functions. Experimental results have shown that our scheme produced more reliable depth maps compared with previous methods. With the proposed hybrid camera system, we could solve the two main problems in the current passive depth sensing, which is depth estimation on occluded and textureless regions. Finally, we have generated high-resolution and high-quality multi-view depth maps from our system. Therefore, our proposed system could be useful for various 3-D multimedia applications.

## Acknowledgments

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-(C1090-1011-0003)).

## References

[1] P. Kauff, N. Atzpadin, C. Fehn, M. Muller, O. Schreer, A. Smolic, R. Tanger, Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability, Signal Processing: Image Communication 22 (2) (2007) 217–234.
[2] C. Fehn, R. de la Barre, S. Pastoor, Interactive 3DTV concepts and key technologies, Proceedings of the IEEE, Special Issue on 3-D Technologies for Imaging and Display 94(3) (2006) 524–538.
[3] ISO/IECJTC1/SC29/WG11 N8944, Preliminary FTV Model and Requirements, July 2007.
[4] A. Smolic, D. McCutchen, 3DAV exploration of video-based rendering technology in MPEG, IEEE Transactions on Circuits and Systems for Video Technology 14 (3) (2004) 348–356.
[5] S.K. Nayar, Y. Nakagawa, Shape from focus, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (8) (1994) 824–831.
[6] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, R. Szeliski, High-quality video view interpolation using a layered representation, Proceedings of ACM SIGGRAPH 23 (3) (2004) 600–608.
[7] D. Scharstein, R. Szeliski, High-accuracy stereo depth maps using structured light, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2003, pp. 195–202.
[8] G. Iddan, G. Yahav, 3D imaging in the studio and elsewhere, in: Proceedings of SPIE Vidometrics and Optical Methods for 3D Shape Measurements, vol. 4298, 2001, pp. 48–55.
[9] M. Kawakita, T. Kurita, H. Hiroshi, S. Inoue, HDTV axi-vision camera, in: Proceedings of International Broadcasting Conference, 2002, pp. 397–404.
[10] CanestavisionTM Electronic Perception Development Kit, Canesta Inc. <http://www.canesta.com/html/developmentkits.htm>.
[11] Swiss Ranger SR-2, The Swiss Center for Electronics and Microtechnology. <http://www.csem.ch/fs/imaging.htm>.
[12] G. Um, K. Kim, C. Ahn, K. Lee, Three-dimensional scene reconstruction using multi-view images and depth camera, in: Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems XII, vol. 5664, 2005, pp. 271–280.
[13] J. Diebel, S. Thrun, An application of Markov random fields to range sensing, in: Proc. of Advanced Neural Information Processing Systems, MIT Press, Cambridge, MA, 2005.
[14] J. Zhu, L. Wang, R. Yang, J. Davis, Fusion of time-of-flight depth and stereo for high accuracy depth maps, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 231–236.
[15] E. Lee, Y. Kang, Y. Jung, Y. Ho, 3-D video generation using hybrid camera system, in: International Conference on Immersive Telecommunications (IMMERSCOM), 2009, pp. T5(1–6).
[16] Z. Zhang, A flexible new technique for camera calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 1330–1334.
[17] Camera Calibration Toolbox Program for Matlab provided by Caltech. <http://www.vision.caltech.edu/bouguetj/calibdoc/>.
[18] Y. Kang, Y. Ho, Geometrical compensation for multi-view video in multiple camera array, in: Proceedings of International Symposium ELMAR, 2008, pp. 83–86.
[19] N. Joshi, B. Wilburn, V. Vaish, M. Levoy, M. Horowitz, Automatic Color Calibration for Large Camera Arrays, UCSD CSE Technical Report, 2005, CS2005-0821.
[20] J. Cho, I. Chang, S. Kim, K. Lee, Depth image processing technique for representing human actors in 3DTV using single depth camera, in: Proceedings of 3DTV Conference, 2007, Paper No. 15.
[21] S. Schuon, C. Theobalt, J. Davis, S. Thrun, High-quality scanning using time-of-flight depth superresolution, in: CVPR Workshop on Time-of-Flight Computer Vision, 2008, pp. 1–8.
[22] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 603–619.
[23] F. Pedro, Felzenszwalb, P. Daniel, Efficient belief propagation for early vision, International Journal of Computer Vision 70 (1) (2006) 41–54.
[24] ISO/IEC JTC1/SC29/WG11 M15377, Reference Softwares for Depth Estimation and View Synthesis, 2008.