https://doi.org/10.2352/ISSN.2470-1173.2017.17.COIMG-444

Multiple View Depth Generation Based on 3D Scene Reconstruction Using Heterogeneous Cameras

Dong-Won Shin and Yo-Sung Ho
Gwangju Institute of Science of Technology (GIST)
123 Cheomdangwagi-ro, Buk-gu, Gwangju, 61005, South Korea
{dongwonshin, hoyo}@gist.ac.kr

Abstract

In this paper, we introduce the multiple view depth generation method using heterogeneous cameras based on 3D reconstruction. The main goal of this research is to generate accurate depth images at each viewpoint of color cameras by using depth cameras placed at different positions. The conventional filter-based framework has critical problems such as truncated depth regions and mixed depth values. It degrades not only the quality of depth images but also synthesized intermediate views. A proposed framework is based on the 3D reconstruction method from the multiple depth cameras. The proposed system setup consists of two camera layers including four color cameras on a lower layer and two depth cameras on an upper layer as a parallel form. First, we estimate correct camera parameters using the camera calibration method on the offline process. In the online process, we capture synchronized color and depth images from the heterogeneous multiple camera system. Next, we generate 3D point clouds from 2D depth images and register them by the iterative closest points method. Then we can obtain an integrated 3D point cloud model. After that, we create the volumetric surface model from the sparse 3D point clouds by the truncated signed distance function. Finally, we can estimate the depth image at each color view by projecting the volumetric 3D model. In the experiment result and discussion section, we will verify not only the proposed framework resolves the aforementioned problems, but also has several advantages over the conventional framework

Introduction

Many realistic 3-Dimensional (3D) contents are available now through various devices such as 3D TV, head mounted display and smartphone. We can readily enjoy a realistic 3D environment by those gadgets. At this point, the 3D depth information is necessary to generate a seamless representation of a model and to give an immersive experience to a user.

These contents can be created by diverse camera systems like stereo cameras, multiple view cameras and depth cameras. We can generally classify the system into passive sensor based method and active sensor based method. In case of the passive sensor based method, it exploits the stereo cameras and calculates the disparity map by using stereo matching technique. In case of the active sensor based method it employs the depth camera and measures the distance between the camera and the object.

When we use the heterogeneous cameras (i.e. color and depth cameras) in a single system to generate the 3D contents, it is critical to make corresponding depth images at each color camera view. In the conventional Multiview depth generation system, it employs the 3D image warping and bilateral filter only even though they provoke some drawbacks such as truncated and mixed depth values.

In this paper, we propose multiple view depth generation framework based on 3D reconstruction. It will reduce some problems in the conventional multiview depth generation framework.

Conventional Multiview Depth Generation Framework

In the conventional multiview depth generation framework, they used a heterogeneous multi-view camera system including eight color cameras and three depth cameras and they captured a single object at a blue-screen studio. The objective of the system is to generate corresponding depth maps at each viewpoint with the same resolution as color cameras. In order to achieve this goal, several computer vision techniques including camera calibration, multi-view image rectification, 3D image warping and bilateral filter were exploited. Figure 1 represents the overall procedure of the presented multi-view depth image acquisition [1].

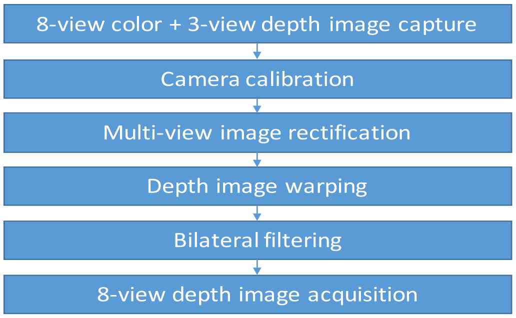

8-view color + 3-view depth image capture
Camera calibration
Multi-view image rectification
Depth image warping
Bilateral filtering
8-view depth image acquisition

FIGURE 1. OVERALL PROCEDURE OF MULTI-VIEW DEPTH IMAGE ACQUISITION

The distance between color cameras is 5.5 cm each. This length represents the approximate distance between human eyes. Since there is not enough space between color cameras to fit in depth cameras, they are placed under the color cameras. Figure 2 exhibits the conventional multiview camera system setup.

Among three depth cameras, they fixed the middle one at the center x-position of color cameras. Then, the other two depth cameras are placed 14 cm each apart from the middle depth camera. From Figure 2, “Depth 2” data corresponds to data for “Color 4” and “Color 5”. Similarly, “Depth 1” and “Depth 3” correspond to the adjacent three color cameras. This configuration is inevitable since too far distance between color and depth cameras provokes erroneous depth results. Therefore, it causes a truncated depth region problem. That is, the depth image is truncated by the field of view(FOV) of each depth camera since the depth values of depth camera are not sent to the all color cameras like Figure 3. They are sent to the view point of color cameras in the designated section which is adjacent to the source depth camera.

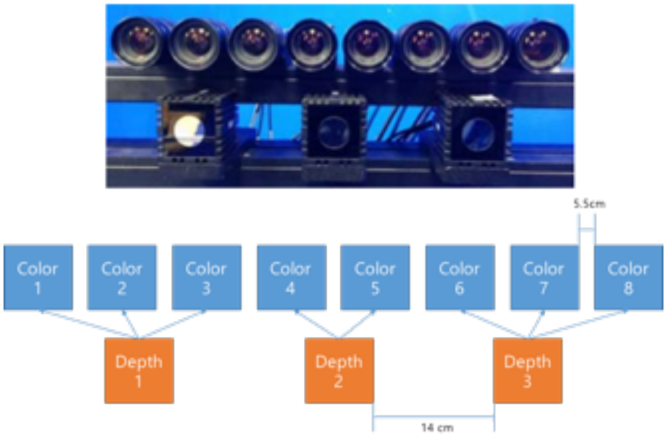


FIGURE 2. MULTIVIEW CAMERA SYSTEM SETUP

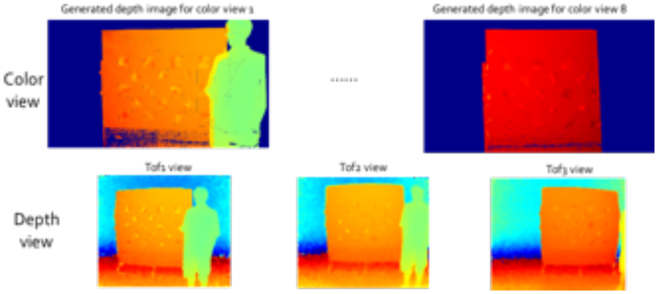

Generated depth image for color view 1
Generated depth image for color view 8
Color view
......
Tof1 view
Tof2 view
Tof3 view
Depth view

FIGURE 3. TRUNCATED DEPTH IMAGES

One of the drawbacks of depth cameras is the low resolution. In the multi-view camera system, depth data at color view positions are generated based on depth values acquired by depth cameras. In order to match the resolution of depth images and color images, they apply 3D warping to the depth image using the original depth camera parameters and rectified color camera parameters [2].

The result of 3D warping of the depth image is shown in Figure 4. The target depth image is inverted in the figure for display purpose. Some boundary errors can be observed. This is due to the limitation of camera calibration accuracy. From depth image warping, eight depth images at each view of color camera are acquired. These images contain empty pixels due to the resolution difference. They exploited the bilateral filter to fill such areas [3].

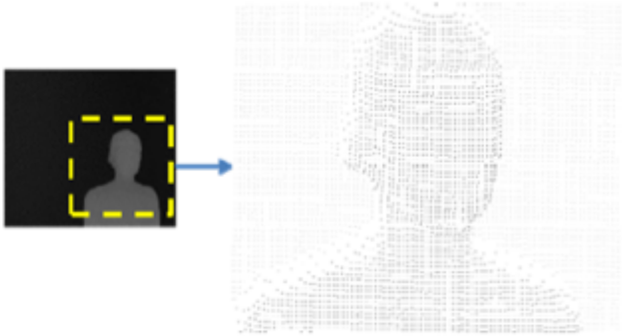


FIGURE 4. 3D WARPING OF DEPTH IMAGE

At this point, the depth values in the warped depth image can be contaminated and blended by the nature of the filter based hole filling methods since the warped depth values to the color view are mixed in a sparse form like Figure 3.

Originally, we concentrated on solving the two major problems (the truncated depth region and mixed depth value). Meanwhile, not only we improved the quality of depth images, but also found out that several advantages beyond the conventional method. We will explain those advantages on the discussion section in detail.

Proposed Multiview Depth Generation Framework

Our heterogeneous multiview camera system has four color cameras and two depth cameras. It consists of two layers having color cameras on the lower layer and the depth cameras on the upper layer as a parallel form as shown in Figure 5.

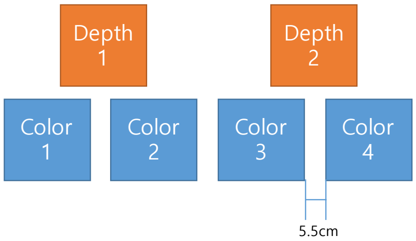


FIGURE 5. SYSTEM SETUP FOR THE PROPOSED FRAMEWORK

The camera calibration step is necessary to obtain the correct camera parameters in the offline process. In the online process, we need to capture the synchronized color and depth images from the multiple camera system. And then, we perform the bilateral filter to reduce noises and build pyramidal image. we send the depth pixels in the 2D image coordinates to the vertices in the 3D world coordinates. After registering the point clouds using an iterative closest points method, we can construct an integrated 3D point cloud model. The dense surface model can be constructed by a truncated signed distance function from the sparse 3D point cloud [4]. Finally, we can obtain the depth images at each color camera by projecting the surface model to the camera viewpoints. Figure 6 shows the flowchart of the proposed method.

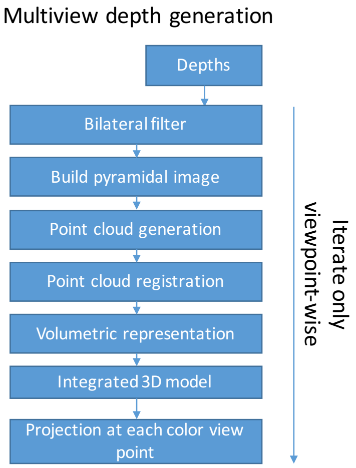

Multiview depth generation
Depths
Bilateral filter
Build pyramidal image
Point cloud generation
Point cloud registration
Volumetric representation
Integrated 3D model
Projection at each color view point
Iterate only viewpoint-wise

FIGURE 6. FLOWCHART OF THE PROPOSED METHOD

Generating 3D point clouds

After capturing the depth images from the ToF cameras, we can generate the 3D point clouds by intrinsic camera parameters. Equation (1) represents the generation of 3D point clouds.

$$\mathbf{M}_{\mathrm{W}} = \mathbf{A}_{\mathrm{i}}{}^{-1}\tilde{\mathbf{m}} \; where\ \boldsymbol{A} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \qquad (1)$$

where A represent the camera intrinsic parameters. $\tilde{m}$ is the 2D point with a homogeneous coordinate system and M_W is the 3D point in the world coordinate system. Figure 7 illustrates the generated 3D point clouds from the raw depth image.

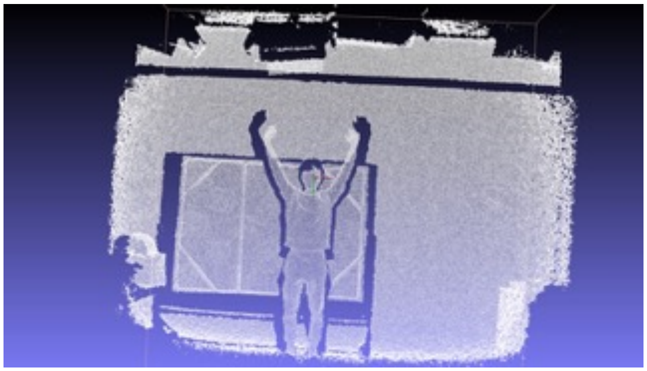


FIGURE 7. GENERATED 3D POINT CLOUDS FROM THE DEPTH IMAGE

Point cloud registration

Once acquiring 3D point clouds from depth image, we need to register the point clouds from each depth camera in an integrated coordinate system. An iterative closest point (ICP) method is a well-known point cloud registration technique by calculating a single six degree of freedom transform between two point clouds [5]. This method consists of following major steps: correspondences establishment, normal vector computation and iterative form optimization.

First, the correspondences between points from a source point cloud to a destination point cloud should be established. We can use a projective data association to achieve reasonable correspondences [6]. Correct correspondences are important since it determines the accuracy of the point cloud registration. Next, the normal vector is an important factor to measure a distance between correspondences. At this point, it is calculated by a simple cross product between two adjacent vectors. For the cost function in the optimization, we will measure the perpendicular distance l_i from source point s_i to tangent plane of a destination point d_i as shown in Figure 8. Here, matrix **M** represents a 3D rigid-body transformation including a rotation matrix and translation vector. Therefore, the normal vector is necessary to measure the distance. The cost function can be optimized by Levenberg-Marquardt method.

Finally, we can get the correct transformation between the point clouds and obtain the integrated point cloud model by applying the estimated transformation and this registration procedure will be the clue to solve the truncated depth region problem.

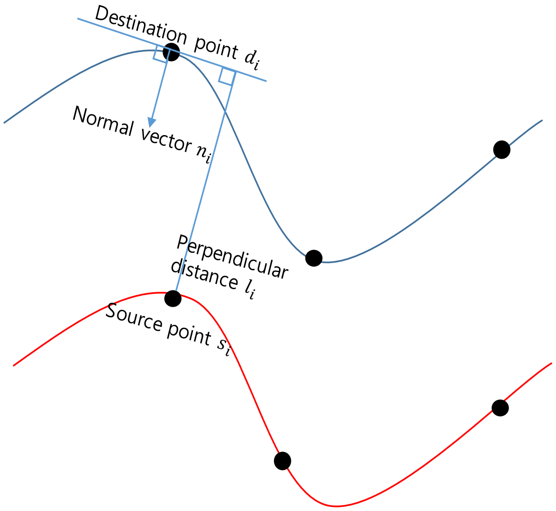


FIGURE 8. POINT-TO-PLANE METRIC

$$\mathbf{M}_{\mathrm{opt}} = \underset{\mathrm{M}}{\operatorname{argmin}} \sum_i ((\mathrm{M} \cdot s_i - d_i) \cdot n_i)^2 \qquad (2)$$

Volumetric representation

The integrated 3D point cloud model will be converted to a surface voxel representation to obtain a smoothed and interpolated representation of the 3D model. At this point, the truncated signed distance function (TSDF) method is used for the conversion [7]. The predefined size of a cube containing the point cloud model in 3D space is divided into a 3D grid of voxels. The main idea of TSDF method is that if the distance from a center of the viewing camera to a specific voxel is greater than a distance of the corresponding vertex, then TSDF value for the voxel will be negative and vice versa.

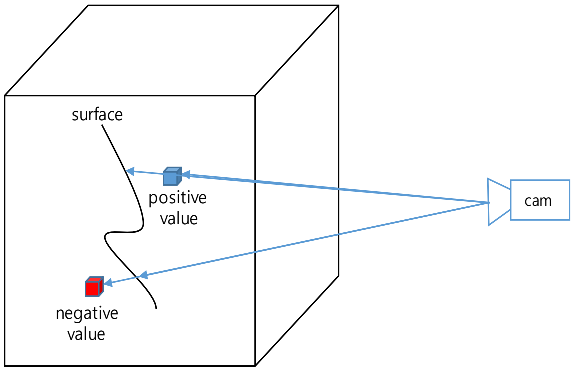


FIGURE 9. TRUNCATED SIGNED DISTANCE FUNCTION

Raycasting method estimates the implicit surface of the volumetric model by finding a zero-crossing point on the TSDF volume as shown in Figure 10. Specifically, the zero-crossing point is estimated by traversing the TSDF volume along the ray connecting the camera center to an image pixel. This volumetric nature will be the breakthrough for the mixed depth value problem.

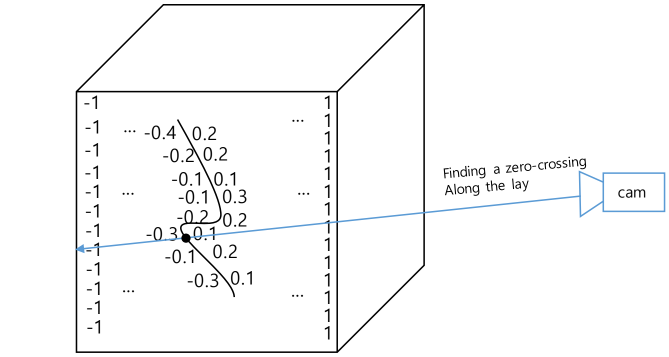


FIGURE 10. RAYCASTING METHOD

Projection onto color camera views

Our ultimate goal is to generate the depth images aligned to the view of the color cameras with the same resolution. By projecting the integrated volumetric model to each color camera view, we can make the depth images. At this point, the camera parameters that we obtained in the offline procedure are exploited.

First, all the extrinsic camera parameters are adjusted by considering the first depth camera as a $[0,0,0]^T$ position in 3D space since the TSDF volume is created at the position depending on the first depth camera. After adjusting the camera parameters, we can project each voxel to the image plane and record the distance from the camera to the corresponding voxel on an image pixel.

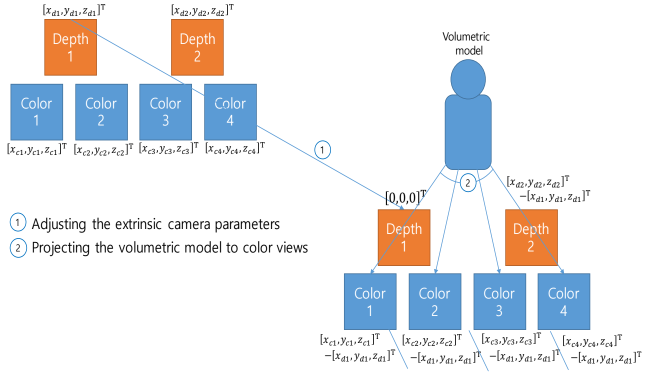

$[x_{d1},y_{d1},z_{d1}]^T$
$[x_{d2},y_{d2},z_{d2}]^T$
Depth 1
Depth 2
Color 1
Color 2
Color 3
Color 4
$[x_{c1},y_{c1},z_{c1}]^T$
$[x_{c2},y_{c2},z_{c2}]^T$
$[x_{c3},y_{c3},z_{c3}]^T$
$[x_{c4},y_{c4},z_{c4}]^T$
Volumetric model
1
2
$[0,0,0]^T$
$[x_{d2},y_{d2},z_{d2}]^T - [x_{d1},y_{d1},z_{d1}]^T$
1 Adjusting the extrinsic camera parameters
2 Projecting the volumetric model to color views
$[x_{c1},y_{c1},z_{c1}]^T - [x_{d1},y_{d1},z_{d1}]^T$
$[x_{c2},y_{c2},z_{c2}]^T - [x_{d1},y_{d1},z_{d1}]^T$
$[x_{c3},y_{c3},z_{c3}]^T - [x_{d1},y_{d1},z_{d1}]^T$
$[x_{c4},y_{c4},z_{c4}]^T - [x_{d1},y_{d1},z_{d1}]^T$

FIGURE 11. CREATION OF DEPTH IMAGES

Experiment Results

In this section, we will introduce the experiment results over the quality of the estimated depth images and elapsed time.

Experimental setup

As we mentioned in Section 3, we employed two depth cameras on an upper layer and four color cameras on a lower layer for the sequence capture system. Specifically, we used Kinect V2 depth cameras with 512×424 depth resolution and Basler color cameras with 1280×720. For the compatibility with the conventional framework, we assigned the depth camera 1 to color camera 1 and 2, the depth camera 2 to color camera 3 and 4 as shown in Figure 12.

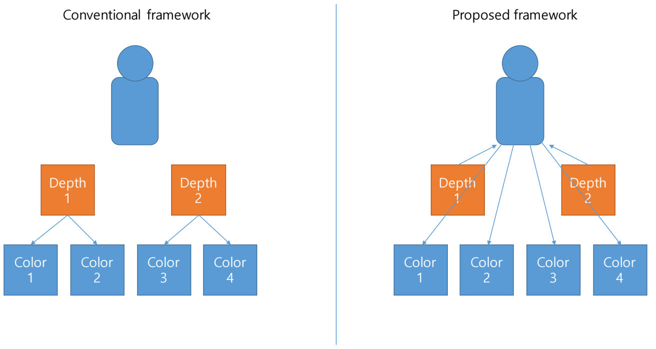


FIGURE 12. EXPERIMENT SETUP

Table 1 shows hardware specifications for the experiments. Since both frameworks exploit GPU parallel programming, also known as CUDA, GPU device is necessary to support them [8].

TABLE 1. HARDWARE SPECIFICATION

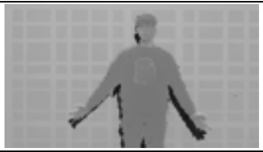
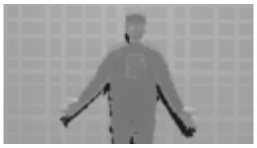
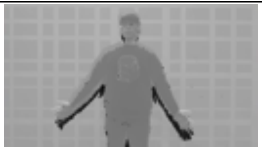
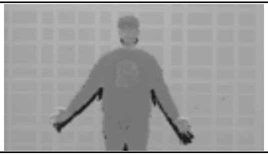
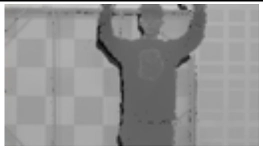
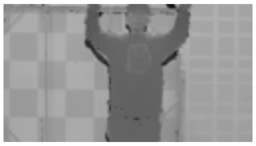
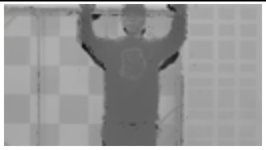
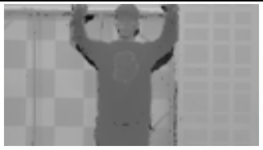
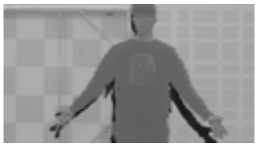
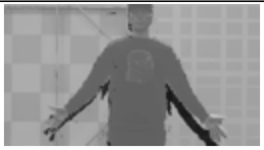
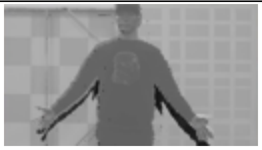
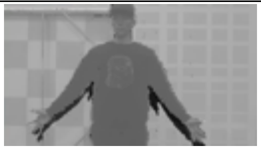
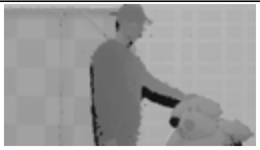
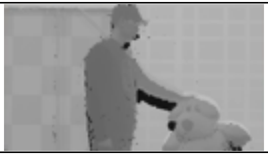
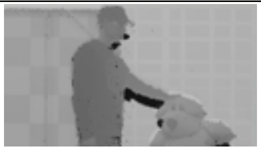
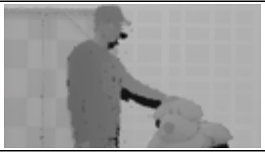
CPU	Intel® Core™ i7-5960X 3.00GHz
GPU	NVIDIA Geforce GTX 1070
RAM	32GB
Color camera	Basler piA1900-32gc GigE
Depth camera	Microsoft Kinect V2

Table 2 describes and illustrates the property of sequences.

TABLE 2. PROPERTY OF SEQUENCES

Name	Description	Still shot
Person	A single person, 175 frames	
Panel 1	Person standing with a panel, 198 frames	
Panel 2	Person standing near an edge of a panel, 197 frames	
Objects	Person, panel, and doll, 109 frames	

TABLE 3. ESTIMATED DEPTH IMAGES WITH CORRESPONDING COLOR IMAGES

Name	Color view 1	Color view 2	Color view 3	Color view 4
Person				
Panel 1				
Panel 2				
Objects				

Estimated depth images

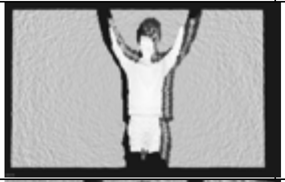
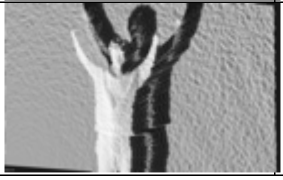
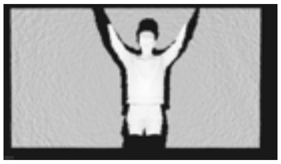
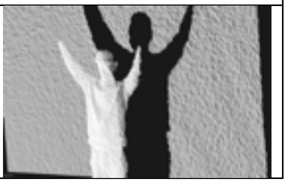
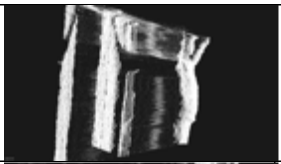
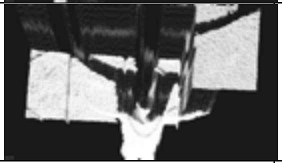
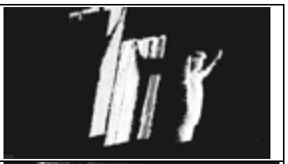
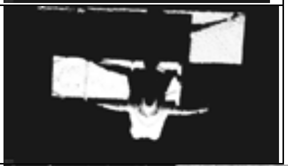
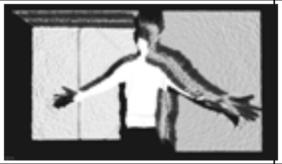
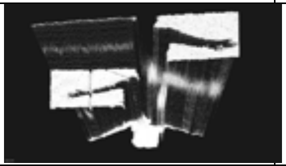
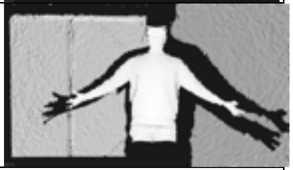
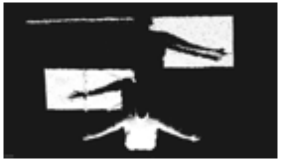
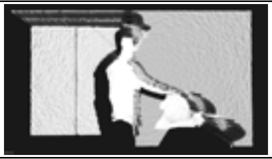
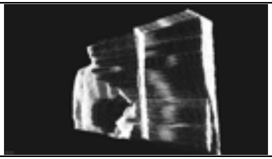
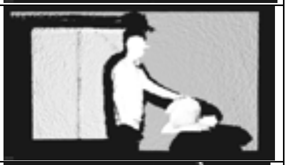
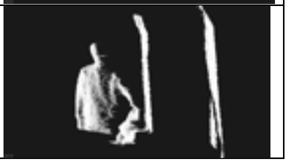
We displayed estimated depth images in Table 3 with corresponding color images in a blending scheme to clearly show whether the color and depth images are well aligned.

As we can find in Table 3, most estimated depth images are well aligned with its corresponding color images. Although it has harsh boundaries and surfaces that we need to improve, we will illustrate the proposed framework is better than the conventional framework in the following section.

Quality of depth images

To validate the quality of depth images, we generated the point clouds by estimated depth images and visualized them in 3D space. Table 4 displays the visualization of point clouds.

TABLE 4. VISUALIZATION OF POINT CLOUDS

Name	Conventional framework	Proposed framework
Person	 	 
Panel 1	 	 
Panel 2	 	 
Object	 	 

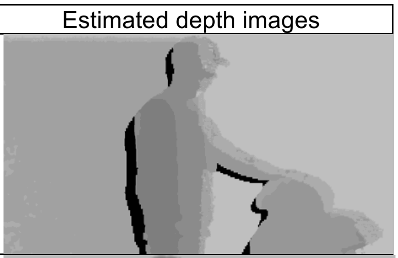
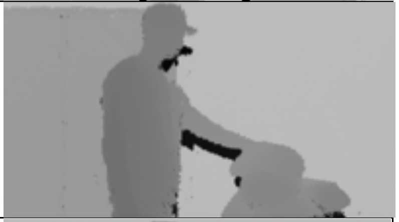
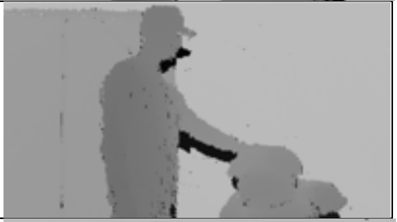
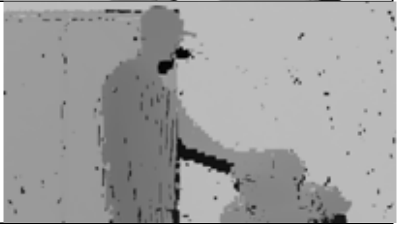
The result of the conventional framework is quite noisy and messy in-between objects because of flying points. This kind of error is caused by the bilateral filter after 3D image warping as we explained in Section 2. Especially in Panel 2 sequence, a depth information on a right half part of a face is distorted since the bilateral filter mixes the depth information from the different objects.

However, the proposed framework shows relatively clear and tidy results than the conventional one since it processes most of the procedure in 3D space with a volumetric fashion.

Processing time

Before comparing the processing time, we would like to mention a scalability of the proposed framework since it affects the processing time. The scalability in the proposed framework means it can control the processing time in proportion to the quality of depth image. In this case, a key factor is the volume dimensions (VD) which mean how many voxels are included in the volume cube. If VD increases, the quality of depth also increases but it takes more processing time. Table 5 shows the comparison of processing time and the quality of depth images.

TABLE 5. COMPARISON OF PROCESSING TIME

	fps	Estimated depth images
Conventional Framework	64.345	
Proposed framework (VD=1024)	24.344	
Proposed framework (VD=512)	99.5365	
Proposed framework (VD=256)	184.757	

As you can see in Table 5, the conventional framework shows a quite fast operation speed but unreliable depth values around the depth boundaries. In the case of the proposed framework, the frame per second varies according to the size of VD. We can properly choose the size of VD depending on applications.

Conclusions

In this paper, we proposed the multiple view depth generation based on 3D reconstruction system with heterogeneous cameras. The truncated depth regions problem and mixed depth value problem from the conventional framework were the primary targets to solve, however, not only we settled the problems but also we found several advantages over the conventional frameworks. It achieved a better depth quality and faster processing time. Moreover, it is beneficial for state-of-the-art 3D displays such as virtual reality (VR) head mounted display, holographic 3D display, and so on. The contribution itself is relatively simple since the projection procedures on the reconstructed 3D model is straightforward but it is powerful for generating the depth images at arbitrary views. we also discussed about how we can improve the quality of depth and processing time. We hope the proposed framework is widely used for acquiring the realistic 3D contents with cutting-edeg technologies such as VR and AR.

Acknowledgment

This research 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], and in part by the 'Brain Korea 21 Plus Project' of the Ministry of Education & Human Resources Development, Republic of Korea (ROK). [F16SN26T2205]

References

[1] Y. Song, D. W. Shin, E. Ko, and Y. S. Ho, "Real-time depth map generation using hybrid multi-view cameras," *APSIPA*, 2014, pp. 1-4.

[2] W. R. Mark, L. McMillan, and G. Bishop, "Post-rendering 3D warping," *Symposium on Interactive 3D Graphics*, 1997, pp. 7-ff.

[3] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint Bilateral Upsampling," *ACM SIGGRAPH*, 2007, pp. 1-5.

[4] S. Izadi, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, S. Davison, and A. Fitzgibbon, "KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera," *ACM symposium on User interface software and technology*, 2011, pp. 559-568.

[5] K. L. Low, "Linear least-squares optimization for point-to-plane icp surface registration," 2004.

[6] S. Rusinkiewicz and M. Levoy, "Efficient Variants of the ICP Algorithm," *International Conference on 3D Digital Imaging and Modeling* 2001, pp. 145-152.

[7] B. Curless and M. Levoy, "A Volumetric Method for Building Complex Models from Range Images," *ACM SIGGRAPH*, 1996, pp. 303-312.

[8] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable Parallel Programming with CUDA," *Queue,* vol. 6, pp. 40-53, 2008.

Author Biography

Dong-Won Shin received his B.S. in computer engineering from the Kumoh National Institute of Technology, Gumi, Korea (2013) and his M.S. in School of Information and Communications from Gwangju Institute of Science and Technology, Gwangju, Korea (2015). He is currently a Ph. D student. His research interests include 3D computer vision and machine learning.

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