©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-408

Implementation of 3D Object Reconstruction Using Multiple Kinect Cameras

Dong-Won Shin and Yo-Sung Ho; Gwangju Institute of Science of Technology (GIST); Gwangju, Republic of Korea

Abstract

Three-dimensional (3D) object reconstruction is to represent objects in the virtual space. It allows viewers to observe the objects at arbitrary viewpoints and feel a realistic sense. Currently, RGB-D camera from Microsoft was released at a reasonable price and it has been exploited for the purpose in various fields such as education, culture, and art.

In this paper, we propose a 3D object reconstruction method using multiple Kinect cameras. First, we acquire raw color and depth images from triple Kinect cameras; the cameras are placed in front of the object as a convergent form. Since raw depth images include hole regions where don't have any depth value, we employ a depth-weighted joint bilateral filter using depth differences between center and neighbor pixels in the filter kernel to fill such hole regions. In addition to that, a color mismatch problem occurs in color images from multi-view data. We exploit a color correction method by means of 3D multi-view geometry to adjust color tones in each image. After matching the correspondences between source and target images by using 3D image warping, we obtain the color correction matrix for the target image via a polynomial regression.

In order to evaluate the proposed depth refinement method, we estimate the bad pixel rate (BPR) of depth images. Our results show that the BPR of the refined depth image is lower than that of the raw depth image. Through the test results, we found that the reconstructed 3D object by the proposed method is more natural than the 3D object using raw images in terms of color and shape.

Introduction

The real world consists of a three-Dimensional (3D) space which has X, Y and Z axes. We can reproduce the real world into the virtual environment by 3D reconstruction technique. Recently, the 3D reconstruction has the important role in various fields, e.g., games, films, advertisement, construction, and art. For example, 3D facial reconstruction is used for medical counseling and 3D human body reconstruction is applied to real-time cloth fitting services [1]. Another important example is a cultural asset reconstruction. When a fire had broken out at Sungnyemun Gate in 2008 which is South Korea's top cultural landmark, a pre-scanned 3D information of Sungnyemun Gate assisted the rebuilding of the cultural asset. After the accident, the Cultural Properties Administration of South Korea acknowledged the importance of 3D virtual reconstruction and proceeded 3D scanning of most cultural assets of South Korea. In addition to that, 3D reconstruction was used for investigation of fire accidents. Unlike the existing ways to record pictures or clips, it record 3D information of the scene which aids the investigation.

In this paper, we propose 3D object reconstruction by low-cost Kinect cameras so that users can fast-access accurate 3D reconstruction at a reasonable price. The contributions of this paper include the implementation of a 3D object reconstruction method using multiple Kinect cameras and the acquisition of the color corrected multi-view color images and the corresponding multi-view depth images. To achieve this goal, we propose a multiple Kinect camera system. It captures multi-view color and depth videos at 30 frame per second. The resolution of the color and depth image is 640×480.

Second, we propose a depth image refinement method for captured multi-view depth images. The proposed algorithms is based on joint bilateral filter [2]. We analyze the existing joint bilateral filter and add a depth weighting factor which is affected by the difference of the depth value between center and neighboring pixel in the filter kernel.

Third, we propose a color correction method using 3D multi-view geometry. In order to perform color correction on the target image, we find the correspondences between source and target images via 3D image warping. Next, we eliminate outliers from occlusion areas among correspondences using the difference between two points. Then we apply polynomial regression to the correspondence set. Finally, we obtain the color correction matrix on the target image.

The overall contribution of this paper is to generate high-quality 3D objects in the virtual space. Specifically, depth map refinement from Kinect and color tone matching among multi-view images are performed.

Microsoft Kinect

Kinect is one of motion sensing input devices developed by Microsoft for Xbox 360 and Xbox One video game consoles and Windows PCs. It allows users to control and interact with their game console or computer just using their own body through a natural user interface using hand & body gestures and spoken orders. The appearance of Kinect is shown in Fig. 1.

Figure 1. Appearance of Kinect

Kinect as a first-generation was introduced in November 2010 in an attempt to enlarge Xbox 360's user base. Kinect for windows was released on February 1, 2012. Microsoft released Kinect

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-408

software development kit for Windows 7 On June 16, 2011 and it meant to allow developers to develop Kinect applications to broaden their own platform.

Kinect sensor has a horizontal bar connecting with a small base including a motorized pivot. The device consist of an RGB camera, depth sensor, and multi-array microphone running proprietary software. Figure 2 shows the structure of Kinect camera [3].

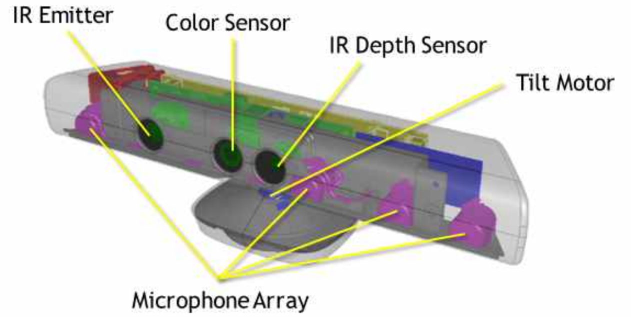


Figure 2. Structure of Kinect camera

Thanks to those devices, Kinect can provide full-body 3D motion capture, facial recognition and voice recognition capabilities. Table 1 shows the device specification of Kinect.

Table 1. Device specification of Kinect

Viewing angle	Vertical: 43°, Horizontal: 57°
Vertical tilt range	±27°
Color image size	640×480
Depth image size	640×480
Frame rate (color and depth stream)	30 frames per second
Audio format	16-kHz, 24-bit mono pulse code modulation (PCM)
Interface	USB 2.0
Power	AC adapter

The sensor has a field of view of 43° vertically and 57° horizontally and the motorized pivot can tilt the sensor up to 27° either up or down. Kinect has various sensors output video at a frame rate of from 9 Hz to 30 Hz depending on resolution. Even though RGB video stream usually uses VGA resolution (640×480), the hardware can capture the video stream up to 1280×1024, and use a colour format as RGB and YUV. The resolution of the depth video stream is in VGA resolution with 11 bit depth (2048 depth levels). Kinect includes Infrared camera in it and can stream the IR view as VGA resolution or 1280x1024 at a lower frame rate. It has a practical ranging limit from 1.2 - 3.5 m when using it as connecting to Xbox console. The microphone array has four microphone capsules and works with each channel processing 24 bit mono pulse code modulation (PCM) at a sampling rate of 16 kHz. Because Kinect's tilting system requires more power than the Xbox 360's USB ports can supply, the device uses USB communication with additional power: power is supplied from the main AC adapter.

Proposed 3D object reconstruction method

In this paper, we propose the 3D object reconstruction method using multiple Kinect cameras. Figure 3 shows the system structure of the proposed method. We placed three Kinect cameras in front of the object as a convergence form and set the position of the 3D object within the field of view of all Kinect cameras. Figure 4 shows the flowchart of the proposed method

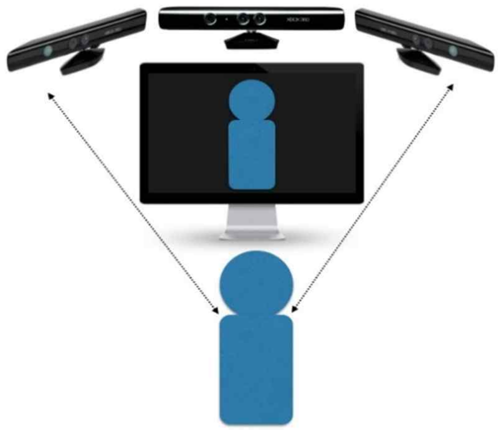


Figure 3. System structure of the proposed method

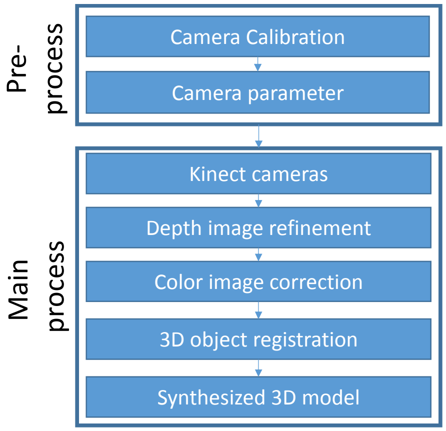


Figure 4. Flowchart of the proposed 3D object reconstruction method

First of all, we perform a camera calibration on Kinect cameras as a preprocessing step. The camera calibration is performed by using a planar chessboard pattern. We capture the images displaying the planar chessboard pattern as various poses and perform the camera calibration step on it by [4]. Then we can obtain intrinsic and extrinsic parameters for each camera.

Next, we refine color and depth images from Kinect cameras as a main process. First, we need to refine the depth image because the raw depth image has a lot of hole regions [5]. In this paper, we refine the depth image by using depth weighted joint bilateral filter. Second, we need to adjust the color image since the raw color image from the multi-view system has a color mismatch problem. When we construct the 3D model in the virtual space, the color could be not seamlessly represented. In order to solve this problem, we explain a color correction method using multi-view geometry information. Next, we make a point cloud model in an integrated 3D space by applying the 3D image warping; the 3D image warping is forwarding a 2D image pixel to a 3D space by using refined color and depth images. In this case, the point cloud model looks sparse because it consists of 3D points. Therefore, we change it to a smooth model through the surface modeling that connects the points. And then finally we can obtain the completed 3D object model.

Depth weighted joint bilateral filter

The raw depth image acquired from Kinect depth camera has many hole regions due to the limitation of Kinect depth camera. These hole regions make us not to smoothly represent the 3D object into the virtual space. Therefore, we need a refinement process to fill the hole regions as proper depth values. In this paper, we employ the depth weighted joint bilateral filter to refine the depth image. Prior to the explanation of the proposed method, we introduce the joint bilateral filter (JBF). The equation (1) represents the joint bilateral filter.

$$D_o(x,y) = \frac{\sum_{u\in\vec{u}_p}\sum_{v\in\vec{v}_p} W(u,v)\cdot D_i(u,v)}{\sum_{u\in\vec{u}_p}\sum_{v\in\vec{v}_p} W(u,v)} \tag{1}$$

D_o is an output depth pixel and D_i is an input depth pixel. *(x,y)* represents a center position of the filter kernel and *(u,v)* represents a neighbor position in the filter kernel. Vector U_p is a vertical position set and V_p is a horizontal position set in the kernel. Lastly, the function W means a weighting function. The symbols are illustrated in Fig. 5 in detail.

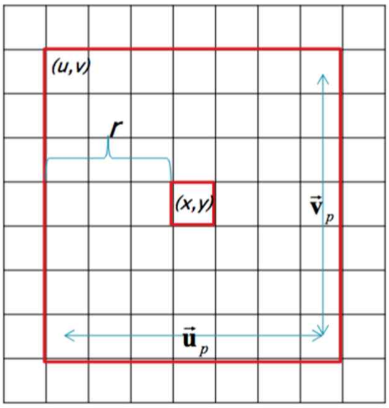


Figure 5. Symbols for the conventional joint bilateral filter

The weighting function *W* is represented as a multiplication of two Gaussian function, which shown in below equations.

$$W(u,v) = \begin{cases} 0 & \text{if } D_i(u,v) = 0 \\ r(u,v)\cdot s(u,v) & \text{otherwise} \end{cases} \tag{2}$$

$$r(u,v) = exp\,(-\frac{|I(x,y) - I(u,v)|^2}{2\sigma_r^2}) \tag{3}$$

$$s(u,v) = exp\,(-\frac{(x-u)^2 + (y-v)^2}{2\sigma_s^2}) \tag{4}$$

where *r(u,v)* is a range filter representing a color difference between center and the neighbor pixels and *s(u,v)* is a spatial filter representing a distance between center and neighbor pixels.

In addition to the conventional JBF, we added a weighting factor for the depth difference. As we can see in Fig. 4, if a neighbor depth pixel has a difference with a center depth pixel, it does not belong to the same region with the center depth pixel [6]. Specifically, there are center (red) and neighbor (blue) pixels in the square kernel and it shows the different depth level and they don't belong to the same region.

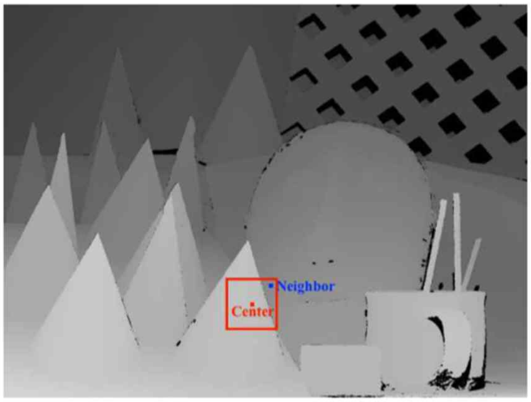


Figure 6. Illustration for the proposed principle

In order to reflect the depth difference between them, we added one more term to the weighing function in joint bilateral filter. So the weighting function is modified to

$$\widehat{W}(u,v) = \begin{cases} 0 & \text{if } D_i(u,v) = 0 \\ r(u,v)\cdot s(u,v)\cdot d(u,v) & \text{otherwise} \end{cases} \tag{5}$$

We added a depth weighting factor *d(u,v)* representing the difference between the center depth pixel and the neighbor depth pixel. It is defined by

$$d(u,v) = exp\,(-\frac{|D_i(x,y) - D_i(u,v)|^2}{2\sigma_d^2}) \tag{6}$$

The depth weighting function *d(u,v)* calculates the weight by reflection the depth difference into Gaussian function.

Color correction using multi-view 3D geometry

In the multi-view camera system, we need to consider the color mismatch problem generated from an inconsistent illumination and a noise on the cameras. The color mismatch problem makes us difficult to construct the seamlessly represented 3D model in terms of the color.

In this paper, we employ the 3D geometry information of the multi-view system to solve this color mismatch problem [7]. Figure 7 shows the flowchart of the proposed color correction method.

DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-408

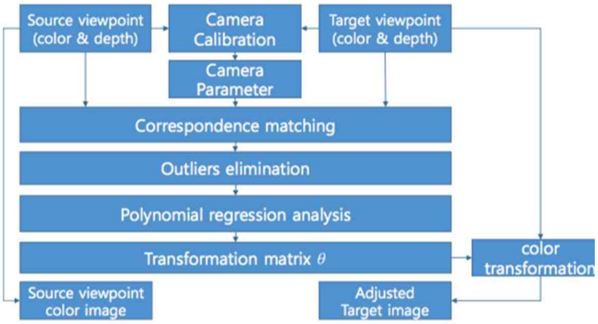


Figure 7. Flowchart of the color correction method

We considered a Kinect camera in the middle as a source viewpoint and Kinect cameras in the both left and right side as a target viewpoint. First, we consider the color correction between source and left target cameras. After obtaining the refined depth images, then we perform the 3D image warping from the left target viewpoint to a middle source viewpoint. Through the 3D image warping, we extract correspondences representing the same 3D position in each 2D view. Because of some correspondences shown in the source viewpoint but not in the target viewpoint, we eliminate the outliers depending on the color difference between correspondences. After that, we perform the color correction step by the polynomial regression for the red color. We can obtain transformation matrix for each color model (R, G, and B) and then apply it to a target view point image. Finally, the adjusted target image is obtained as a result of the proposed color correction method.

3D image warping is the method generating a new view from a base view with its depth information [8]. Figure 8 illustrates the 3D image warping process.

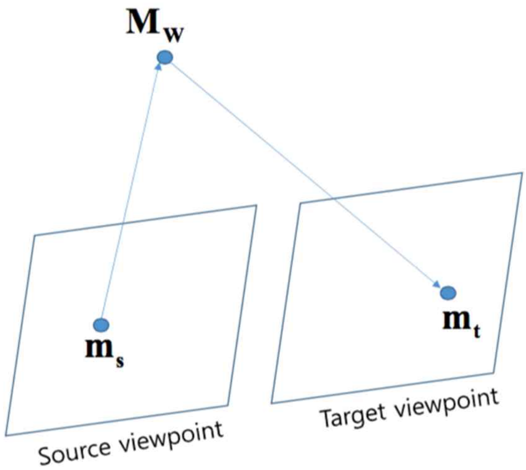


Figure 8. 3D image warping process

The point in the source viewpoint is forwarded to the 3D space and we can get a 3D point $\mathbf{M}_W$. After that, it is sent to the target viewpoint to obtain the point $\mathbf{m}_t$. By using this, we can find which pixel in the target viewpoint is related with the pixel in the source viewpoint. In this paper, we mainly use it in the color correction and the 3D model registration. Both Eq. (7) and Eq. (8) show 3D image warping. Equation (7) represents a translation from the source viewpoint to the 3D space and Eq. (8) represents a translation from the 3D space to the target viewpoint.

$$\mathbf{M}_W = \mathbf{R}_s^{-1}\mathbf{A}_s^{-1}\tilde{\mathbf{m}}_s - \mathbf{R}_s^{-1}\mathbf{t}_s \tag{7}$$

$$\tilde{\mathbf{m}}_t = \mathbf{A}_t\mathbf{R}_t\mathbf{M}_w + \mathbf{A}_t\mathbf{t}_t \tag{8}$$

Symbol **A** stands for an intrinsic parameter through a camera calibration. Symbol **R** and **t** stand for a rotation matrix and translation vector for an extrinsic parameter respectively. Subscript **s** means the source viewpoint and **t** means target viewpoint. By using these equation, we can find correspondence between source and target viewpoints.

Now, we found the correspondences between source and target views by 3D image warping. Fig. 7 shows a part of correspondences between two views (about 10%). We can see that most of correspondences are well matched.

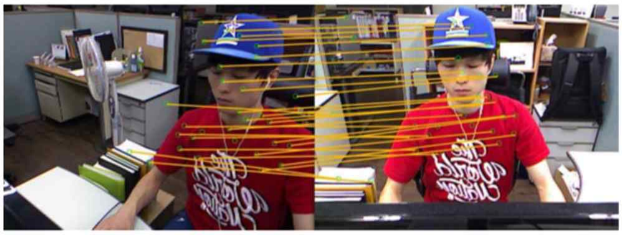


Figure 9. A part of correspondences between two views

When we obtain the correspondence between source and target images by 3D image warping, an occlusion can be occurred [9]. Figure 10 shows an example for the occurrence of the occlusion. Some point on the left image can be misdeemed like as it exists on the right image even it blocked by some object. To eliminate the outliers, we will use the color information. If color values between correspondences exceed some threshold, we can consider the correspondence as an outlier and remove it from the correspondence set.

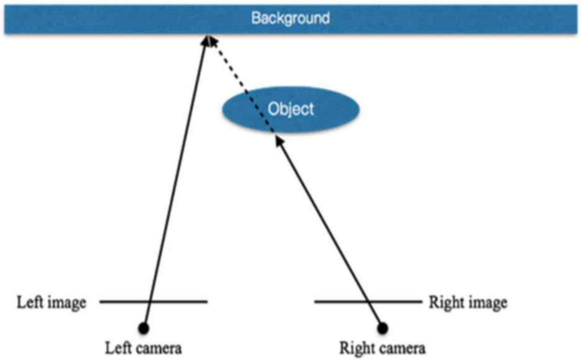


Figure 10. An example for the occurrence of the occlusion

After eliminating the outliers, we can organize the color correction matrix. First, we need to construct a vector X like Eq. (9).

DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-408

$$\mathbf{X} = [1 \quad r_t \quad r_t^2 \quad r_t^3 \quad r_t^4 \quad r_t^5] \tag{9}$$

where r_t means the red value of the pixel in the target viewpoint image. We can get the color correction matrix for the red value by the polynomial regression analysis with fifth-order polynomial equation using vector **X**. Equation (10) shows the calculation of the color correction matrix **θ** between the source viewpoint and target viewpoint images.

$$\boldsymbol{\theta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \tag{10}$$

where **y** means the red values at the source viewpoint. We can adjust the color pixels in the target viewpoint image by using this transformation matrix. Next, we apply the same process on the green and blue color and get the color correction matrix.

Surface Modeling

3D model registration is a stage which represents the object in virtual 3D space by using the corrected color image and the refined depth image. This stage can be performed by 3D image warping, but we just apply forwarding the 2D depth image to 3D virtual space rather than applying all process of 3D image warping.

When we forward 2D depth information from each viewpoint to an integrated 3D space, we can acquire a point cloud model. In case of the point cloud model, it represents a sparse form because point to point connection is not completed. Therefore, we need to do a surface modeling which connects point to point and make a smooth object. Figure 11 shows the surface modeling method.

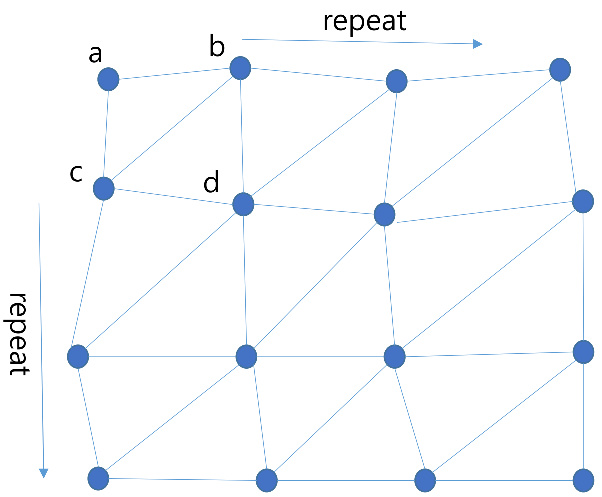


Figure 11. Surface modeling method

First of all, we make a triangle by using point a, b and c: the point b is at the right side and below the point a, there is point c. Next, we make a triangle bcd which shares a leg with the triangle abc by using point d. Because it is hard to determine that which point is located exactly at the right or left side of a basis point, we can check the depth image from each viewpoint to determine left or right position. Although it is decided as a neighbor point, we classify it as outliers if a distance from a basis point is much far, and then eliminate it. When completing this process, we can reconstruct the 3D object model in virtual space.

Test results

Table 2 describes experimental environments for the proposed method.

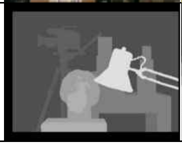
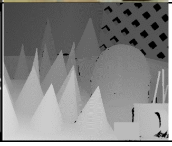
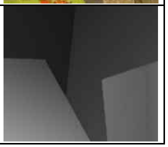
Table 2. Experiment environments

CPU specification	Intel Xeon 2.53GHz
GPU specification	NVidia Geforce GTX Titan X
Color image size	640×480
Depth image size	640×480
Range filter sigma	3
Spatial filter sigma	5
Depth filter sigma	5
Filter kernel radius	3

We employ a graphics processing unit (GPU) in the depth image refinement and 3D image warping process to accelerate the calculating speed.

We used image datasets for the experiment from the middlebury computer vision group, which is famous in the computer vision field [6]. We calculated the bad pixel rate (BPR) representing the rate of the hole region in comparison with the whole pixel in the image. Table 3 shows the sequences used in BPR experiment.

Table 3. Sequences used in the BPR experiment

	Tsukuba	Cones	Venus
Left color image			
Right color image			
Ground truth depth image			

Result of the proposed depth refinement method

Figure 12 and Fig. 13 show the raw depth image and the result of the proposed depth refinement method respectively. The raw depth image shows a lot of hole regions at the object boundary and the planar part of the desk. However, we can see a neatly filled region in the result of proposed depth refinement method.

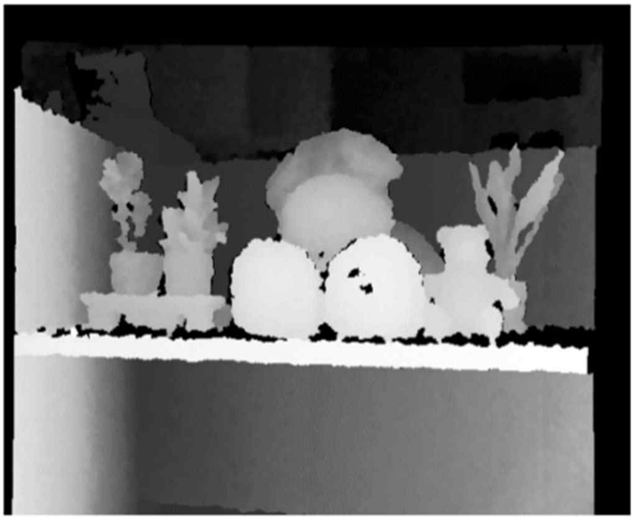


Figure 12. Raw depth image

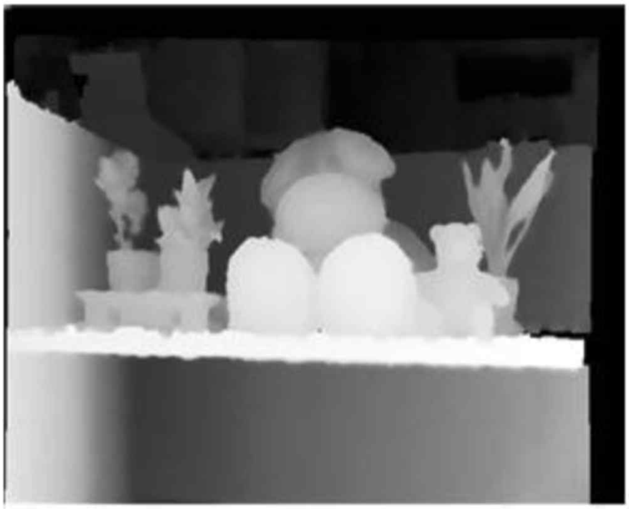


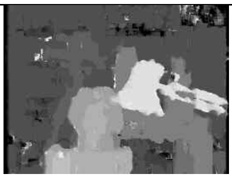
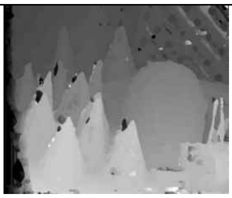
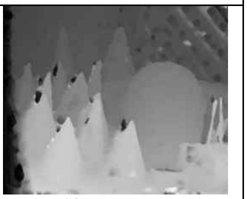
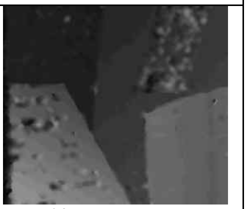
Figure 13. Result of the proposed depth refinement method

For the objective comparison, we compared the bad pixel rate (BPR) which represents a rate for pixels having a difference more than one between a target depth image and the ground truth depth image [10].

First, we perform stereo matching to the left and right images of the middlebury sequences. After obtaining depth images from stereo matching using the simple squared sum of differences (SSD), then we apply the proposed depth refinement method on these depth images to fill the hole regions. After that, we compare the bad pixel rate of the depth images. Table 4 shows the result images and BPR values from the comparison.

In Table 4, the BPR for the result depth images are reduced in comparison with those of the raw depth images. From this result, we can see that the proposed method refines the raw depth image as closely as the ground truth depth image.

Table 4. Result images and BPR values

	Raw depth images from stereo matching (image/ BPR)	Proposed method (image/ BPR)
Tsukuba	 8.90%	 5.68%
Cones	 16.39%	 16.28%
Venus	 12.02%	 6.96%

When we employ the proposed depth refinement method to 3D object reconstruction, we can visually find the difference between two results. Figure 14 shows the result of the 3D reconstruction model from the raw depth and color images and Fig. 15 shows the result using the proposed method respectively. The result 3D model shows that noises which have a large amount of depth difference are much reduced.

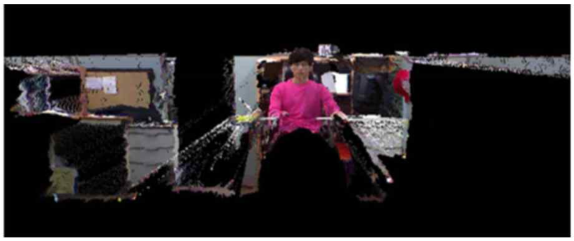


Figure 14. 3D model from the raw depth images

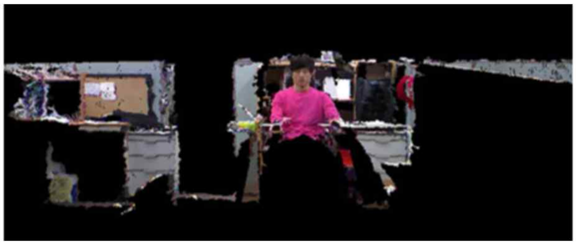


Figure 15. 3D model from the proposed depth refinement method

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-408

Final result of 3D object reconstruction

Figure 16 and Fig. 17 show the 3D object model using the raw images from Kinect cameras and the proposed method respectively. In Fig. 16, there are hole regions at the face and body part since it is constructed by the raw depth image. Moreover, we can find a color inconsistency from the surface of the 3D model made by the raw color image. Whereas, Fig. 17 shows the improved 3D model in terms of the shape and the color.

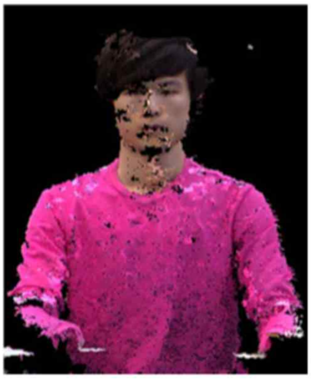


Figure 16. 3D model using the raw depth and color images

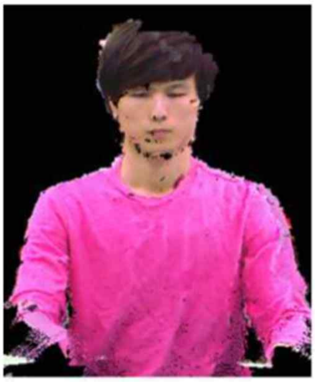


Figure 17. 3D model using the proposed 3D object reconstruction method

Conclusions

As of 2007, the standardization for 3D image technologies has been proceeded by the Moving Picture Experts Group (MPEG). Now, the public can easily access 3D images. To keep pace with the trend, many studies have been taken so that more people can access high-quality of 3D contents at a reasonable price.

In order to achieve this purpose, we proposed the 3D reconstruction method using multiple Kinect cameras. For the proposed system, we placed triple Kinect cameras in front of the object as a convergent form. Next, we proposed the depth weighted joint bilateral filter to refine the raw depth image from Kinect cameras. Next, we used the 3D geometric information of the multi-view system to correct the raw color image. By using 3D image warping, we found the correspondences between source and target images. After that, we applied polynomial regression to obtain transformation matrices for color correction.

We compared the BPR with the depth images from the stereo matching using the simple squared sum of differences method. The depth images from the proposed method showed the lower BPR. We discovered that the 3D object reconstruction results of the proposed method exhibit the better quality than the results based on raw images in terms of color and shape.

We hope this cost-efficient 3D object reconstruction method will be practical in various fields, e.g., game development, film production, and education environment.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(2011-0030079)

References

[1] F. Remondino, “3-D reconstruction of static human body shape from image sequence,” *Computer Vision and Image Understanding*, vol. 93, no. 1, pp. 65–85, 2004.

[2] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, “Joint bilateral upsampling,” *ACM Transactions on Graphics*, vol. 26, no. 3, p. 96, Jul. 2007.

[3] “Kinect for Windows Sensor Components and Specifications.” [Online]. Available: http://msdn.microsoft.com/en-us/library/jj131033.aspx.

[4] D. W. Shin and Y. S. Ho, “Pattern Feature Detection for Camera Calibration Using Circular Sample,” *Advances in Multimedia Information Processing - PCM 2015*, vol. 9315, pp. 608–615, 2015.

[5] D. S. Alexiadis, D. Zarpalas, and P. Daras, “Real-Time, Full 3-D Reconstruction of Moving Foreground Objects From Multiple Consumer Depth Cameras,” *Multimedia, IEEE Transactions on*, vol. 15, no. 2, pp. 339–358, Jan. 2013.

[6] D. Scharstein and R. Szeliski, “High-accuracy stereo depth maps using structured light,” *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1, pp. I–195 – I–202, 2003.

[7] D. W. Shin and Y. S. Ho, “Color correction using 3D multi-view geometry,” in *Proceedings of SPIE - The International Society for Optical Engineering*, 2015, vol. 9395, p. 93950O.

[8] W. R. Mark, L. McMillan, and G. Bishop, “Post-rendering 3D warping,” in *Proceedings of the 1997 symposium on Interactive 3D graphics - SI3D ’97*, 1997, p. 7–ff.

[9] W. S. Jang and Y. S. Ho, “Efficient depth map generation with occlusion handling for various camera arrays,” *Signal, Image and Video Processing*, vol. 8, no. 2, pp. 287–297, Sep. 2013.

[10] Y. S. Kang and Y. S. Ho, “Low-Resolution Depth Map Upsampling Method Using Depth-Discontinuity Information,” *The Journal of Korea Information and Communications Society*, vol. 38C, no. 10, pp. 875–880, Oct. 2013.

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. in electronic engineering from the Seoul National University, Seoul, Korea (1981) and his Ph.D. in electrical and computer engineering from the University of California, Santa Barbara (1990). He worked in Philips Laboratories from 1990 to 1993. Since 1995, he has been with the Gwangju Institute of Science and Technology, Gwangju, Korea, where he is currently a professor. His research interests include image analysis, 3D television, and digital video broadcasting.