Real-time Depth Map Generation Using Hybrid Multi-view Cameras

Yunseok Song, Dong-Won Shin, Eunsang Ko, and Yo-Sung Ho
Gwangju Institute of Science and Technology (GIST), Gwangju, Rep. of Korea
E-mail: {ysong, dongwonshin, esko, hoyo}@gist.ac.kr

***Abstract*— In this paper, we present a hybrid multi-view camera system for real-time depth generation. We set up eight color cameras and three depth cameras. For simple test scenarios, we capture a single object at a blue-screen studio. The objective is depth map generation at eight color viewpoints. Due to hardware limitations, depth cameras produce low resolution images, i.e., 176×144. Thus, we warp the depth data to the color cameras views (1280×720) and then execute filtering. Joint bilateral filtering (JBF) is used to exploit range and spatial weights, considering color data as well. Simulation results exhibit depth generation of 13 frames per second (fps) when treating eight images as a single frame. When the proposed method is executed on a computer per depth camera basis, the speed can become three times faster. Thus, we have successfully achieved real-time depth generation using a hybrid multi-view camera system.**

I. Introduction

3D video has gained huge popularity since the success of numerous 3D commercial films. 3D video allows the viewer to experience natural depth perception. In general, depth maps are required to generate 3D data. Depth maps contain camera-to-object distance information. These data are applied to view synthesis [1-3].

Assuming an auto-stereoscopic display, multiple views of color data need to present. For natural 3D video display, the number of views becomes the higher the better. Since the number of color views is limited to the number of cameras at the capturing stage of 3D video system, the receiver side performs view synthesis, i.e., virtual view generation, which requires depth data.

Depth maps can be captured by depth cameras or be estimated by stereo images. The former case allows fast and accurate depth data acquisition via time-of-flight (ToF) sensors [4]. However, depth cameras are not always cost-efficient, and the number of used cameras can be limited due to frequency issues. In addition, depth camera resolutions are smaller than color cameras due to hardware limitations. The latter case does not require depth cameras. Yet, depth estimation can be a lengthy process which is an issue when depth images need to be processed in real-time.

In this paper, we use a hybrid multi-view camera system, including eight color cameras and three depth cameras. The objective is to generate depth data for all eight color camera views in real-time. We capture a single object at a blue-screen studio. The objective is to generate depth maps at each viewpoint of the color cameras with the same resolution as the color camera, 1280×720. Several techniques including camera calibration, multi-view image rectification, image warping and joint bilateral filtering are performed. Fig. 1 represents the overall procedure of the presented multi-view depth image acquisition.

II. Multi-view Camera System Setup

We configure eight color cameras (Basler piA1900-32gc GigE) and three depth cameras (Mesa Imaging SR4000). Due to frequency limitation in such depth cameras, the maximum number that can be used is three. Fig. 2 exhibits the multi-view camera system setup.

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. Among three depth cameras, we fix 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 Fig. 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.

Both color and depth cameras capture images at 30 frames per second (fps). While color images are captured at 1280×720 resolution, depth images are 176×144. Due to this discrepancy, depth image warping and filtering are required, which is explained in Section 4 and Section 5.

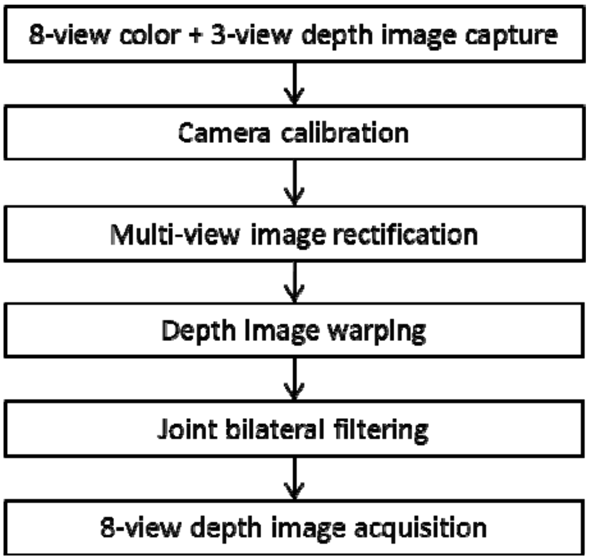

8-view color + 3-view depth image capture
Camera calibration
Multi-view image rectification
Depth Image warping
Joint bilateral filtering
8-view depth image acquisition

Fig. 1 Overall procedure of multi-view depth image acquisition

978-616-361-823-8 © 2014 APSIPA

III. Pre-processing Steps

Pre-processing steps contain camera calibration and multi-view image rectification. Such tools are vital for multi-view image processing.

A. Camera Calibration

Camera calibration is the process of estimating camera parameters. There exist intrinsic and extrinsic camera parameters; intrinsic camera parameters include focal length and principal point; extrinsic parameters contain rotation and translation matrices. These parameters are necessary for not only multi-view image rectification but also image warping. We adopt the widely used calibration method [5]. For each camera, ten distinct images of a square grid pattern are taken for grid-corner based computation. The length of each side of a square is 250 mm. This calibration method is applied to both color and depth cameras.

B. Multi-view Image Rectification

Multi-view image rectification is applied to mitigate alignment discrepancies in color cameras. Notable errors include intrinsic parameters, non-uniform distances between adjacent two cameras, inconsistent horizontal disparities and vertical mismatches. These affect the accuracy of 3D image processing.

After acquiring the camera parameters, we choose the fourth camera as the reference camera since this is closest to the middle. Other color cameras use the intrinsic parameters and rotation matrices of the fourth camera. Camera centers are adjusted as well. For the non-reference cameras, since the extrinsic matrices have changed, pixels of certain areas, i.e., top/bottom, left/right sides may not exist. We add the same focal offset to each camera in order to make the frames contain only captured data. Due to the offset, rectified images become slightly enlarged compared to the original images. Fig. 3 shows the results of multi-view image rectification. The improvement of camera alignment can be confirmed. Rectified camera parameters are stored and used for depth image warping consequently.

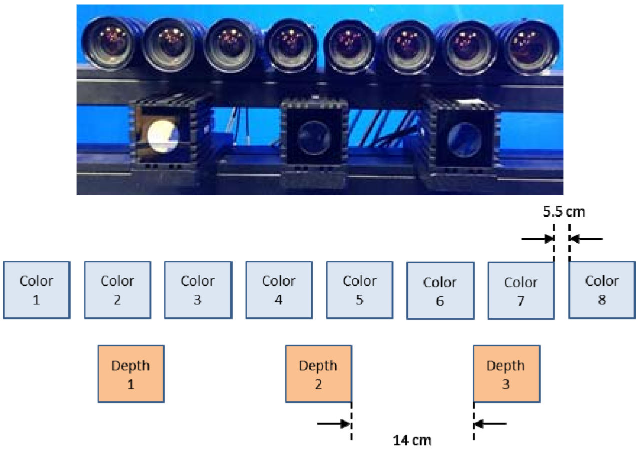
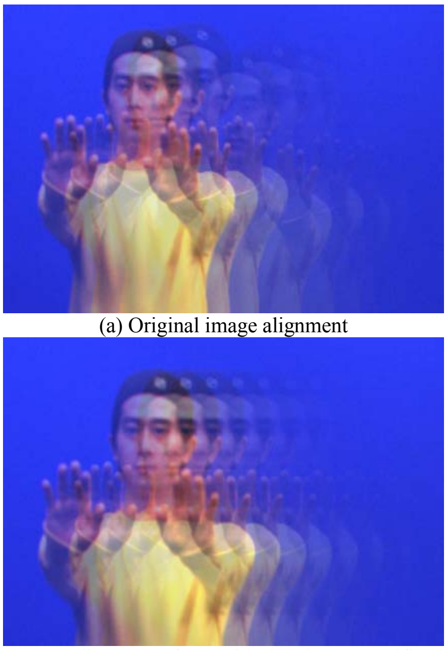


Fig. 2 Multi-view camera system setup


(a) Original image alignment

(b) Image alignment after multi-view image rectification

Fig. 3 Multi-view image rectification results

IV. 3D Warping of the Depth Image

One of the drawbacks of depth cameras is the low resolution. In our 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, we apply 3D warping to the depth image using the original depth camera parameters and rectified color camera parameters.

In the source image, each pixel is projected to a 3D point, i.e., world coordinate, then this 3D point is projected to the target image. The source image represents the original depth image at 176×144 resolution and the target image is 1280×720 resolution depth image. Equation (1)-(5) represents the 3D warping process; this is depicted in Fig. 4. Subscripts *l* and *r* denote left and right, respectively. *m*, *A*, *R*, *t* and $\mathrm{M_w}$ are 2D image point, intrinsic matrix, rotation matrix, translation matrix and 3D image point, respectively. The 2D image coordinate at the target image can be acquired by the 2D image point at the source image and its intrinsic and extrinsic parameters.

$$m_l = A_l \cdot R_l \cdot \mathrm{M_W} + A_l \cdot t_l \tag{1}$$

$$m_l - A_l \cdot t_l = A_l \cdot R_l \cdot \mathrm{M_W} \tag{2}$$

$$A_l^{-1} m_l - t_l = R_l \cdot \mathrm{M_w} \tag{3}$$

$$R_l^{-1} \cdot A_l^{-1} \cdot m_l - R_l^{-1} \cdot t_l = \mathrm{M_W} \tag{4}$$

$$\begin{aligned} m_r &= A_r \cdot R_r \cdot \mathrm{M_W} + A_r \cdot t_r \\ &= A_r \cdot R_r \cdot R_l^{-1} \cdot A_l^{-1} \cdot m_l - A_r \cdot R_r \cdot R_l^{-1} \cdot t_l + A_r \cdot t_r \end{aligned} \tag{5}$$

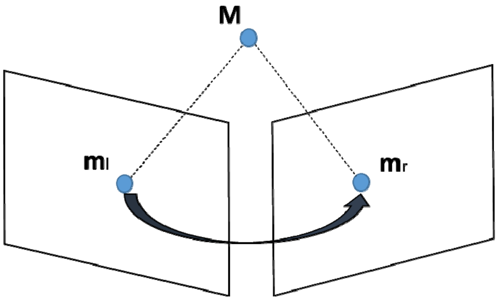


Fig. 4 3D warping of source image to target image

The result of 3D warping of the depth image is shown in Fig. 5. 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.

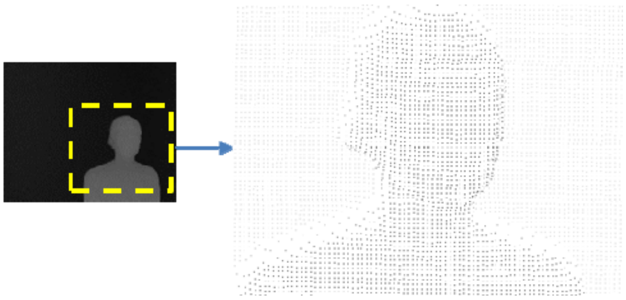


Fig. 5 3D warping of depth image

V. JOINT BILATERAL FILTERING

From depth image warping, eight depth images are acquired. These images contain empty pixels due to the resolution difference. We use the joint bilateral filter (JBF) to fill such areas [6]. The filter is applied to the object region only, assuming the approximate depth value range of the object is known.

JBF is an extension of the bilateral filter, which is generally applied for edge preserving [7]. Equation (6)-(11) describe how JBF generates depth values.

$$D(x,y) = \frac{\sum_u \sum_v W(u,v) D_i(x,y)}{\sum_u \sum_v W(u,v)} \tag{6}$$

$$W(u,v) = \begin{cases} 0 & \text{, if } D_i(x,y) = 0 \\ f(u,v) \cdot g(u,v) & \text{, otherwise} \end{cases} \tag{7}$$

$$f(u,v) = \exp\left\{ -\frac{|I(x,y) - I(u,v)|^2}{2\sigma_f^{\,2}} \right\} \tag{8}$$

$$g(u,v) = \exp\left\{ -\frac{(x-u)^2 + (y-v)^2}{2\sigma_g^{\,2}} \right\} \tag{9}$$

$$u = \{x - r, \cdots, x + r\} \tag{10}$$

$$v = \{y - r, \cdots, y + r\} \tag{11}$$

(x, y) and (u, v) are image coordinates. $D_i(x, y)$ and $D(x, y)$ denote the depth value at (x, y) in the warped depth image and the final depth image, respectively. W represents the weight, which is zero if the pixel value in the warped depth image is zero. Otherwise, the weight is a multiple of spatial weight and range weight.

The spatial weight is based on the intensity difference. When computing the intensity difference, JBF uses the color image while the bilateral filter uses the depth image itself. JBF produces more reliable spatial weights since color data difference can be more specific. The range weight is the same in both JBF and the bilateral filter. This is based on the differences in x- and y-coordinates. These weights depend on kernel size r, spatial parameter σ_f and range parameter σ_g.

VI. SIMULATION RESULTS

We used a test sequence containing a human object with a small movement of arms in front of blue-screen. The distance between the camera and object is about 3.5 m. This type of scenario can be applied to tele-conferencing where simple background and a single object are enough.

Depth generation speed results are presented in Table 2. These results do not include the computation time of preprocessing steps such as camera calibration and multi-view image rectification. Since there exists no ground truth depth maps, the objective quality we can measure is the depth generation speed. Fig. 6 and Fig. 7 show the rectified color images depth images captured from depth cameras. Fig. 8 displays the results of generated depth maps after warping and JBF. Since background information is available in rectified color images, background data of depth maps are set to black; 3D warping and JBF are performed only at the object region. Although some boundary mismatches exist, depth data are consistent. As a future work, we intend to evaluate the quality of depth maps by generating virtual views.

Implementation wise, 3D warping and JBF were executed on graphics processing unit (GPU). Since such tools are independent operation, i.e., processing on a pixel-by-pixel basis, parallel processing is much more effective than serial processing in terms of speed.

In our simulation, a single computer was used for the entire computation. Hence, eight images were counted as a single frame. Depth maps were generated at 13 fps. In a more practical situation, three cameras would be used, one for each depth camera, handling warping and JBF simultaneously since the processes are independent for each view. Thus, depth generation speed can be up to three times faster, i.e., 39 fps. If a single computer is used, the speed can still be enhanced by employing multiple GPUs.

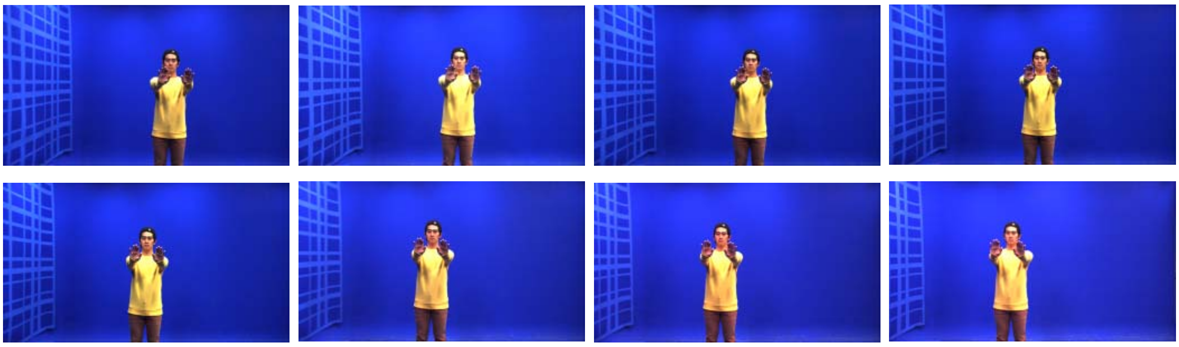


Fig. 6 Rectified color images (1280×720)

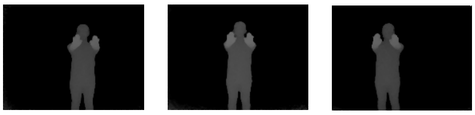


Fig. 7 Depth camera images (176×144)

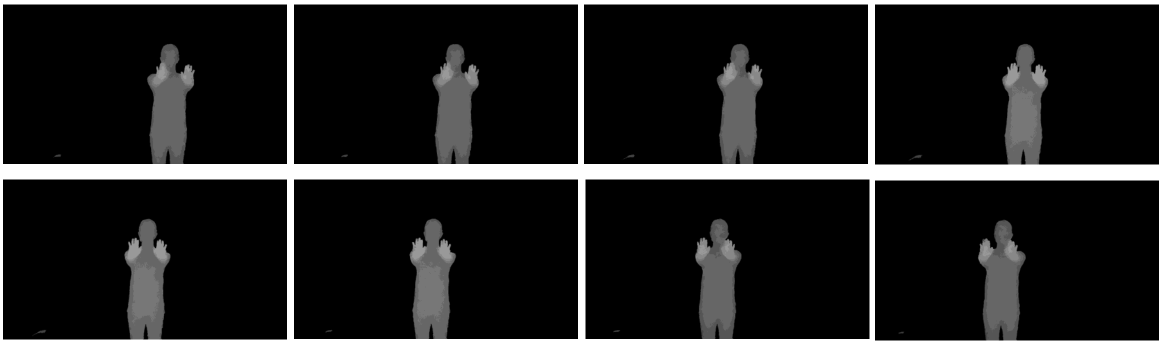


Fig. 8 Generated depth images (1280×720)

VII. CONCLUSION

In this paper, we presented a hybrid multi-view camera system with the objective of generating high-resolution depth maps at viewpoints of color cameras. After capturing images from eight color cameras and three depth cameras, we perform camera calibration and multi-view image rectification. Using the rectified camera parameters, low-resolution depth images are warped to high-resolution images, the same resolution as color images. In the warped images, refining and filling of empty areas are achieved by JBF. Spatial and range weights are considered in this process. GPU processing is exploited for implementation of 3D warping and JBF. For evaluation, the presented camera system captured a single object with a small movement at a blue-screen studio. Simulation results indicate that depth maps of eight color views were successfully generated at 13 fps.

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). [GK13C0100, Development of Interactive and Realistic Massive Giga-Content Technology]

REFERENCES

[1] C. Fehn, "Depth-image-based rendering (DIBR), Compresion and Transmision for a New Aproach on 3-D TV," *Proc. of SPIE Conference Stereoscopic Displays and Virtual Reality Systems*, vol. 5291, pp. 93-104, Jan. 2004.

[2] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View Generation with 3D Warping Using Depth Information for FTV," *Signal Processing: Image Communication*, vol. 24, no. 1, pp. 65-72, Jan. 2009.

[3] Y. Song, C. Lee, and Y.S. Ho, "Adaptive depth boundary sharpening for effective view synthesis," *Picture Coding Symposium (PCS)*, pp. 73-76, May 2012.

[4] Y.S. Kang, Y.S. Ho, "Disparity Map Generation for Color Image Using a TOF Depth Camera," *3DTV Conference*, pp. 1-4, May 2011.

[5] Camera Calibration Toolbox for MATLAB, http://www.vision.caltech.edu/bouguetj/calib_doc/.

[6] J. Kopf, M.F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint Bilateral Upsampling," *ACM Transactions on Graphics*, vol. 26, no.3, pp. 1-5, July 2007.

[7] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images", *IEEE International Conference on Computer Vision*, pp. 839-846, Jan. 1998.