GAZE CORRECTION USING 3D VIDEO PROCESSING FOR VIDEOCONFERENCING

Yo-Sung Ho and Woo-Seok Jang

School of Information and Communications
Gwangju Institute of Science and Technology (GIST)
Email: {hoyo, jws}@gist.ac.kr

ABSTRACT

In this paper, we propose a gaze correction method using 3D video processing techniques including depth estimation and virtual view synthesis. We design a set of color and depth cameras to reduce occlusion regions and improve the depth precision in the less-detailed region. The proposed algorithm deals with fully unsolved problems by fusing the depth data from both depth sensors. Furthermore, view synthesis is performed to create the gaze corrected image from the obtained depth information. Experimental results show our contribution is useful for videoconferencing.

***Index Terms*—** depth estimation, gaze correction, videoconferencing, , view synthesis

1. INTRODUCTION

Videoconferencing is the conduct of conference between two or more participants at different locations using a set of telecommunication systems to transmit audio and video data. Although many videoconferencing systems have been developed, the naïve use of cameras in these systems lack eye contact. This creates some kind of disconnected feeling, reducing effectiveness of interaction. Therefore, the gaze correction problem is considered as one of the most important issues in the videoconferencing system [1].

Over the past several decades, a variety of gaze correction methods have been proposed. However, these methods require complex hardware configurations for their performances as well as high cost for system setup [2]. In order to overcome these drawbacks, three-dimensional (3D) video system technology can be used [3].

Depth estimation and view synthesis are core technologies for 3D video systems. These can be used in gaze correction for videoconferencing to design cost-effective system. In this paper, we propose a gaze correction method using depth image based rendering (DIBR). DIBR is one of the most widely used methods which create virtual images in arbitrary view position [4]. In order to obtain accurate depth information, we use two color and one depth cameras. Recently, small and cheap depth camera such as Kinect depth camera, are introduced without high cost burden [5]. Although the Kinect depth data possesses low accuracy, compared to more expensive depth cameras, due to sensor noises and occlusion regions, we can reduce the cost burden for creating of gaze corrected image. Thus, we utilize a Kinect depth camera by making up for its weaknesses. Furthermore, the proposed method generates gaze corrected image via view synthesis.

2. SYSTEM OVERVIEW FOR GAZE CORRECTION

Recently, available display sizes in the market have increased rapidly. Thus, our system setup is designed to target large screen environment. 55-inch display is used and the viewing distance 3.0 meters. Figure 1 illustrates the overall system that includes two color camera and one Kinect depth camera. In order to capture texture information from the center view, we set up the color cameras on the left and right of the display. The left and right cameras verge at each other.

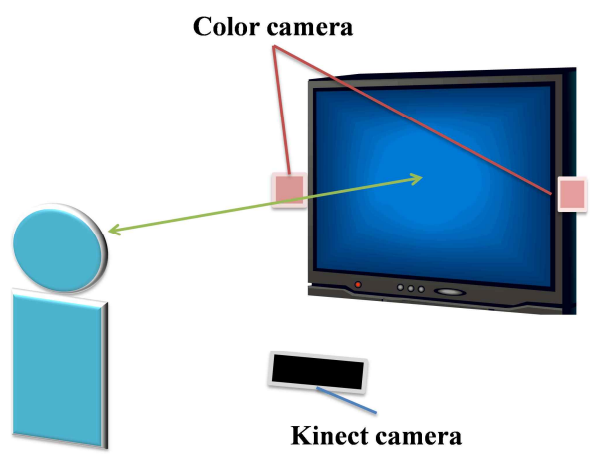


Fig. 1. System setup for gaze correction

When we use two or more cameras, acquisition of the relative camera information is necessary. Camera calibration is the technique of attaining camera parameters [6].Camera parameters have three components: intrinsic matrix including focal length and principal point, rotation matrix, and translation vector. Camera parameters are used for 3D

Microsoft Exhibit 1039 - Page 1
Microsoft Corporation v. RealD Spark, LLC
IPR2023-01046 - U.S. Patent No. 10,740,985

warping of Kinect depth values and view synthesis for gaze corrected image generation.

3. DEPTH ACQUISITION IN COLOR CAMERAS

We use the Kinect depth data to supplement the crude depth map obtained by stereo matching. The initial work for this is that Kinect depth data is mapped to its corresponding position of the color views by 3D warping [7]. 3D warping in the proposed method is composed of two processes. First, the Kinect depth is transformed to the 3D space based on camera parameters. Then the data in the 3D space is projected to the left and right image position. 3D warping is performed as follows.

$$(x,y,z)^T = R_{src}A_{src}^{-1}(u,v,1)^T d_{u,v} + t_{src}, \quad (1)$$

$$(l,m,n)^T = A_{dst}R_{dst}^{-1}(x,y,z)^T - t_{dst}, \quad (2)$$

$$(u',v') = (1/n, m/n), \quad (3)$$

where A_{src}, R_{src}, and t_{src} are intrinsic, rotation, and translation parameters in the structured light pattern depth data, respectively. Similarly A_{dst}, R_{dst}, and t_{dst} are those in the target color view. $d_{u,v}$ is depth value at (u,v) coordinate in the structured light pattern depth camera. The depth data is sent to the 3D space by (1) and projected to the target view by (2). (u', v') in (3) represents the projected coordinate to the target view.

For the next step, we perform upsampling to interpolate the low resolution depth map to the color resolution. We apply joint bilateral upsampling (JBU) for this [8]. The proposed depth estimation is based on the global energy function defined by maximum a posterior Markov random field (MAP-MRF) [9]. The upsampled Kinect depth data is utilized as the additional evidence for the energy function. The Kinect depth data improves the precision and accuracy of the depth map in color images by allowing large depth variations.

The boundary of the acquired disparity map is not matched well with that of the corresponding texture image. The problem degrades the quality of the synthesized image. Thus, we employ a discontinuity preserving filter to solve the unmatched boundary problem [10]. Formally, the depth value D(x, y) at the position (x, y) is computed by this filter as follows:

$$W(u,v) = \exp\left(-\frac{\|I(x,y), I(u,v)\|^2}{2\sigma_R^2}\right) \cdot \exp\left(-\frac{(x-u)^2 + (y-v)^2}{2r^2}\right) \quad (4)$$

$$C(u,v,d) = \min(\lambda, |D(u,v) - d|) \quad (5)$$

$$D(x,y) = \arg\min_{d \in \vec{d}_p} \frac{\sum_u \sum_v W(u,v) \cdot C(u,v)}{\sum_u \sum_v W(u,v)} \quad (6)$$

where W(u, v) consists of intensity and spatial weighting functions that are usually modeled by the Gaussian function, and σ_R and r are smoothing parameters of intensity and spatial weighting functions, respectively. (6) is a truncated linear model to allow depth discontinuities and λ is a constant to reject outliers. In (4), the center pixel is substituted by one of $\vec{d}_p$ = {$D(x-1,y)$, $D(x,y-1)$, $D(x, y)$, $D(x+1,y)$, $D(x,y+1)$} among its neighboring pixels.

4. GAZE CORRECTED IMAGE GENERATION

The depth map is used to create a gaze corrected images through a view synthesis technique. First, we calculate the camera parameters of the gaze corrected position. The camera center and rotation matrix in the gaze corrected position is computed by

$$C_{center} = (1-\lambda)C_{left} + \lambda C_{right}, \quad (7)$$

$$t_{center} = -R_{center} \cdot C_{center}, \quad (8)$$

where $C_{centert}$, C_{left} and C_{right} are the positions of the camera center for the gaze-corrected, left and right views. Generally, λ in (7) is set to 0.5 for the center view. t_{center} and R_{center} are translation and rotation parameters for the intermediate view, respectively [11]. Extracting Euler angles for generating intermediate rotation parameters is carried out [12].

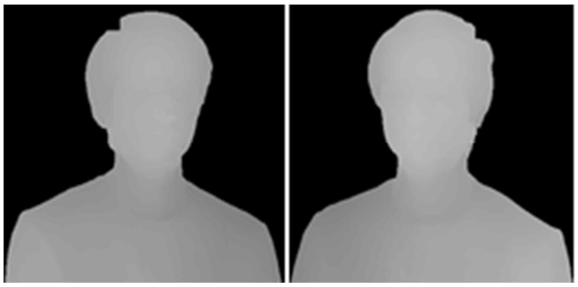
The intermediate rotation parameters are determined by taking middle of angles extracted from the left and right views. Then, in order to create an accurate gaze corrected image, we find homography transformation for the all depth values between reference views and gaze corrected view. Computed homography matrices are used to project the whole texture information of the original image to the target position using the depth information. Consecutively, blending of the texture information from the left and right images and hole filling operations are performed.

5. EXPERIMENTAL RESULTS

Color images are captured at the resolution of 1920×1080 pixels for the proposed method. The Kinect camera employs the resolution of 640×480 pixels. Kinect depth data are represented by 16-bit distance values. Figure 2 to Fig. 4 show the original image and upsampled results of the warped depth in color positions. The results represent that boundary is not good, but we can obtain the precision depth data especially in regards to user's face.

Microsoft Exhibit 1039 - Page 2
Microsoft Corporation v. RealD Spark, LLC
IPR2023-01046 - U.S. Patent No. 10,740,985

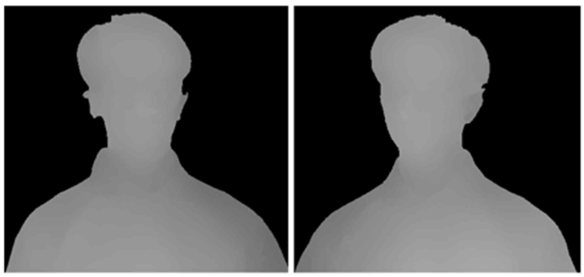
(a) Original color images



(b) Upsampled depth maps

Fig. 2. Upsampling results of warped depth 1

(a) Original color images

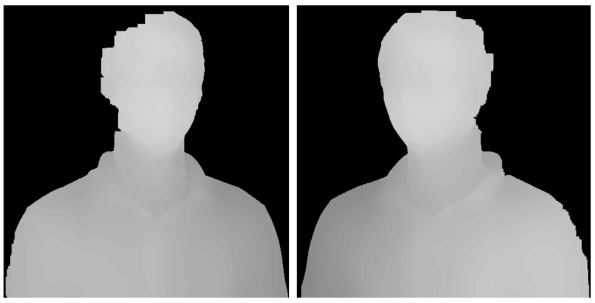


(b) Upsampled depth maps

Fig. 3. Upsampling results of warped depth 2

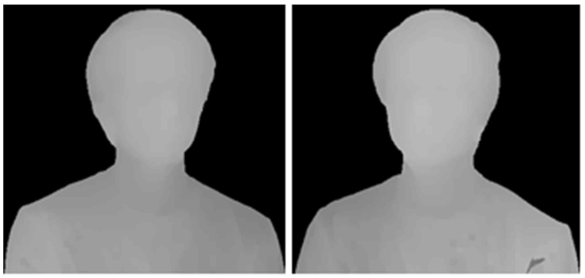
Figure 5 and Fig. 7 show the results of the final depth map including discontinuity preserving postprocessing [9]. From the results, the proposed method can represent the depth details. Furthermore, our method improves the quality in the textureless region such as face. This figure also exhibits gaze correction results using view synthesis. This result shows that the proposed method is highly effective in gaze correction.

(a) Original color images



(b) Upsampled depth maps

Fig. 4. Upsampling results of warped depth 3

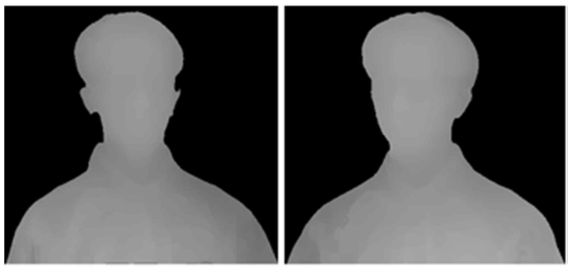


(a) Final depth maps

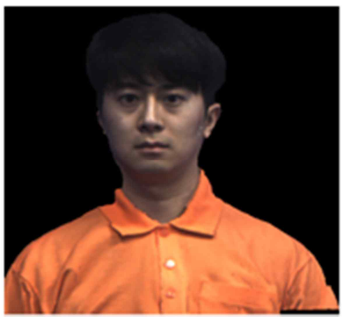
(b) View synthesis result

Fig. 5. Final results of gaze corrected view synthesis 1

Microsoft Exhibit 1039 - Page 3
Microsoft Corporation v. RealD Spark, LLC
IPR2023-01046 - U.S. Patent No. 10,740,985

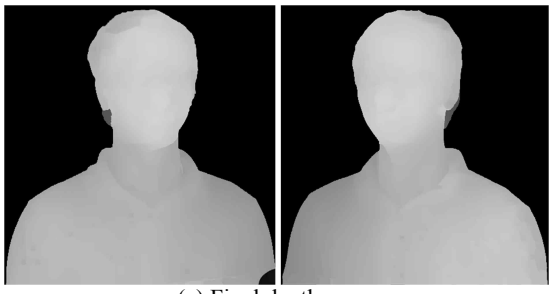


(a) Final depth maps

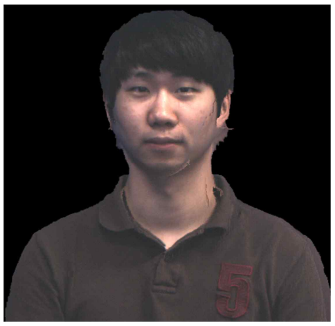


(b) View synthesis result

Fig. 6. Final results of gaze corrected view synthesis 2



(a) Final depth maps



(b) View synthesis result

Fig. 7. Final results of gaze corrected view synthesis 3

6. CONCLUSION

This paper proposes an eye gaze correction method solving inherent problems of depth sensors. Recent development of large display monitors makes it more difficult to find corresponding points between the left and right sides of the display. Furthermore, the less-detailed depth due to stereo setup derails natural eye contact. Kinect depth camera can improve depth precision and reduce occlusion regions. Thus, we design fusion depth system including Kinect depth camera. Experimental results have shown that our method enhances eye contact.

ACKNOWLEDGEMENT

This research was supported by the ICT R&D program of MSIP/IITP (11-921-05-001, Development of Immersive Smart work Core Technology for Collaboration among Multiple Parties), and in part by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2011-0030079).

REFERENCES

[1] C. Kuster, T. Popa, J.C. Bazin, C. Gotsman, and M. Gross, "Gaze correction for home video conferencing," *ACM Transaction on Graphics*, vol. 31, no. 6, pp. 1-6, 2012.

[2] K.-H. Tan, I.N. Robinson, B. Culbertson, and J. Apostolopoulos, "ConnectBoard: Enabling Genuine Eye Contact and Accurate Gaze in Remote Collaboration," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 466-473, 2011.

[3] Video and Requirement Group, "Call for Proposals on 3D Video Coding Technology," N12036, ISO/IEC JTC1/SC29/WG11, 2011.

[4] L. Zhang, and T. Wa James, "Stereoscopic image generation based on depth images for 3DTV," I*EEE Transactions on Broadcasting*, vol. 51, no. 2, pp. 191-199, 2005.

[5] L. Xia, C.C. Chen, and J.K. Aggarwal, "Human detection using depth information by Kinect" Computer Vision and Pattern Recognition Workshops , pp. 15-22, June 2011.

[6] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , vol. 22, no. 11, pp. 1330-1334, 2000.

[7] Y.S. Kang, and Y.S. Ho, "High-quality multi-view depth generation using multiple color and depth cameras," IEEE International Conference on Multimedia and Expo, pp. 1405-1410, 2010.

[8] J. Kopf, M. F. Cohen, D. Lischinski et al., "Joint bilateral upsampling," ACM Trans. Graph., vol. 26, no. 3, pp. 96, 2007.

[9] L. Wang, H. Jin, and R. Yang, Search Space Reduction for MRF Stereo. in: D. Forsyth, P. Torr, and A. Zisserman, (Eds.), Computer Vision – ECCV 2008, Springer Berlin Heidelberg, pp. 576-588, 2008.

[10] Q. Yang, L. Wang, and N. Ahuja, "A constant-space belief propagation algorithm for stereo matching," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1458-1465, 2010.

[11] R. Hartley, and A. Zisserman, Multiple View Geometry in Computer Vision, Second ed.: Cambridge University Press, 2004.

[12] M. Day. "Extracting Euler Angles from a Rotation Matrix," https://d3cw3dd2w32x2b.cloudfront.net/wp-content/uploads/2012/07/euler-angles1.pdf.

Microsoft Exhibit 1039 - Page 4
Microsoft Corporation v. RealD Spark, LLC
IPR2023-01046 - U.S. Patent No. 10,740,985