Immersive Stereoscopic 3D System with Hand Tracking in Depth Sensor

Sunjin Yu¹, Eungyeol Song², and Changyong Yoon³

¹School of Digital Media Engineering, Tongmyong University, Busan, Korea

²Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea

³Department of Electrical Engineering, Suwon Science College, Hwaseong, Korea

Ijfis

Abstract

Recently, interest in immersive tele-conference is increasing. Conventional 2D images cannot provide a stereoscopic effect, but stereoscopic 3D (S3D) images can provide a more realistic image to people. Also, hand tracking technology is an interesting research area, in the field of human computer interaction. Hand tracking technology that tracks and recognizes human hand movements has made it possible to provide a new concept of natural user interface. A depth sensor such as Kinect provides color information and depth information at the same time. In this paper, we propose an immersive S3D system by generating stereoscopic images using depth image based rendering technology and applying 3D hand tracking and augmented reality technology.

Keywords: 3D hand tracking, Stereoscopic 3D, Depth sensor, Augmented reality

1. Introduction

Recently, with the development of ICT technology, interest in communication methods through real-time video call or video conference is increasing. Technological advances have solved many problems with video calls, but there are still limitations for immersive video calls. For a realistic video call, calibration between specific equipment or camera rigs and equipment for stereoscopic 3D image generation is required. In order to solve this problem, stereoscopic image production technology using a depth sensor is being studied to apply stereoscopic 3D technology and real time viewing in a video call application.

Depth sensors [1, 2] such as Microsoft Kinect and ToF (time of flight) generate depth map in real time, which is expected to enable real-time implementation, which was a problem in conventional stereoscopic image generation technology. The use of a depth sensor (PrimeSensor, Kinect, ToF, etc.) that can acquire real-time 3D information is applicable to real-time applications such as video calls and has the advantage of being lower in cost than existing 3D acquisition devices. Due to the commercialization of ToF camera and PrimeSensor, which can acquire 3D information in real time, various applications based on existing 2D images are extended to 3D applications, and 2D image-based hand recognition technology is also able to recognize the natural movement of the user by enabling recognition of the 3D motion of the hand. Through the development of 3D hand tracking technology, it became possible to recognize various motions of human being.

The conventional 2D image based natural user interface [3–6] is sensitive to ambient

Received: May 22, 2018

Revised : Jun. 1, 2018

Accepted: Jun. 12, 2018

Correspondence to: Changyong Yoon
 (cyoon@ssc.ac.kr)

©The Korean Institute of Intelligent Systems

©This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

illumination changes, and it is not enough to perceive performance degradation and the movement of human body moving in 3D space through 2D image. Therefore, fusion of 2D and 3D information is required [7–9]. Depth sensor based 3D hand tracking technology solves these problems and provides information to recognize human 3D hand movements in real time. In this paper, we apply augmented reality technology to 3D hand tracking based on stereoscopic 3D system. The proposed system obtains depth map from Microsoft Kinect to generate stereoscopic 3D image and uses depth based 3D tracking technology [10, 11]. Next, 3D objects are augmented on the hand that is tracked in 3D space, and augmented reality is applied. The proposed system is an augmented reality system that generates stereoscopic 3D images from depth images, tracks human hand movements in 3D space, creates 3D virtual objects on hands, and displays stereoscopic 3D images to users.

We use depth image based rendering (DIBR) technology to generate 3D stereoscopic images from depth information [12–14]. 3D hand tracking technology based on depth image is applied for hand motion estimation in 3D space. After placing a 3D virtual object on the traced 3D hand and acquiring a depth map from the 3D information of the 3D object, the virtual object is also displayed as a 3D stereoscopic image by applying the DIBR technique. In this paper, we explain depth image based technologies in Section 2 and describe the proposed stereoscopic 3D system with hand tracking in Section 3. We discuss the experimental results in Section 4 and finally conclude.

2. Background

2.1 Depth Image Based Rendering

A depth sensor such as Kinect provides both a color map and a depth map. Because the depth image contains 3D information, it can be converted into 3D information. In addition, 3D color information can be generated in 3D space by including information obtained from color image. The DIBR technique creates a virtual viewpoint from the depth image and generates a stereoscopic image by separating it into left and right binocular images.

Stereoscopic camera setup can be divided into the toed-in camera setup and shift-sensor camera setup [15]. Toed-in camera setup is similar to human eye structure, but distortion occurs such as depth non-linearity, depth plane curvature, and so on. In the shift-sensor camera setup, the CCD is set to the left and right in parallel and has convergence points. In the shift-sensor

Figure 1. Example of ALBP.

camera setup, there is no depth plane curvature distortion such as a toed-in method. In the DIBR, assuming a shift-sensor camera setup, a simplified parallax shift calculation formula is applied to reduce computational complexity in the real implementation [15, 16]. Non-linear formulation is as follows:

$$p = x_B \left(\frac{1}{\frac{W}{D} \left(\frac{m}{2^N} (k_{near} + k_{far}) - k_{far} \right) - 1} + 1 \right), \quad (1)$$

where w is screen width, and m is depth.

The parallax p_{pix} in units of pixels is expressed as follows:

$$p_{pix} = \frac{p N_{pix}}{W}, \quad (2)$$

where N_{pix} is the number of horizontal pixels of the display.

The linear approximation is expressed as follows:

$$p_{pix} \approx -x_B \frac{N_{pix}}{D} \left(\frac{m}{2^N} (k_{near} + k_{far}) - k_{far} \right). \quad (3)$$

2.2 3D Hand Tracking

An adaptive local binary pattern (ALBP) method is applied for hand tracking in depth images [10]. Conventional local binary pattern (LBP) is robust to rotation, but ALBP is robust to rotation and distance. In feature extraction for hand tracking, ALBP uses only depth information without color information. A texture T is first defined with respect to a pixel with local neighborhoods with radius size r as the joint distribution of the gray levels of I ($I > 0$) image pixels as shown below:

$$T_r(g_c) = t(g_0, g_1, g_2, \dots, g_{t-1}), \quad (4)$$

where g_i ($i = 0, 1, \dots, I-1$) are the pixel intensities of circular neighborhoods around the center pixel. I is the number of circular neighborhood points of ALBP. Figure 1 shows example of ALBP [10].

In this paper, the fast version of ALBP [10] is applied in

Figure 2. Used morphology operation and original and filtered images in depth map.

order to perform real-time processing as shown below:

$$\sum_{i=0}^{I-2} p(s_i + s_{i+1}) + p(s_0 - s_{I-1}), \quad (5)$$

where s_i denotes $s(g_i + g_c)$ and $p(x) = \begin{cases} 0, & x = 0, \\ 1, & \text{otherwise.} \end{cases}$

3. System Overview

3.1 Pre-processing

In the proposed system, smoothing filtering is applied to remove noise from the depth image. As the size of the mask increases, the filtering operation time becomes longer and the information distortion becomes larger. Through the experiment of various morphology operations, erode, median filtering, and dilate are applied in order to remove the noise of the image. Figure 2 used morphology operation and original and filtered images in depth map.

3.2 Object Segmentation

Depth information obtained from Kinect is used to separate foreground and background. The foreground is assumed to be human and other objects are assumed to be background. In the proposed system, the original depth map size is 640×480 pixels. The original depth map consists of x and y coordinates and brightness values according to depth information. We first change the depth map consisting of x and y coordinates to represent x and z coordinates, where the z coordinate normalizes the depth from 0 to 255 so that the size of the transformed image is 640×255 pixels. Through this process, we can convert the view point of the acquired image into a bird view as seen

Figure 3. Converted depth map by x and z coordinates.

Figure 4. Result of object segmentation by connected component labeling.

Figure 5. Original depth map and segmented result in color map.

from the top. Figure 3 shows converted depth map by x and z coordinates.

In the next step, a morphological operation is applied to remove noise and group the pixel masses. Finally, position information between objects is obtained by labeling each component using the connected component labeling method in the image represented by x - z coordinates. Figure 4 result of object segmentation by connected component labeling.

Since the labeling number obtained as a result of connected component labeling changes every frame, it keeps track of the selected object in the first frame. In the first frame, a face detection algorithm [17] is used to designate a human object as foreground in order to specify a component corresponding to a foreground object among the respective components. Figure 5 shows original depth map and segmented result in color map. Figure 6 shows procedure of object segmentation.

Figure 6. Procedure of object segmentation.

Figure 7. Example of depth images and histogram equalization results by original depth, OpenCV and MATLAB, respectively.

4. Experimental Results

4.1 Depth Image Histogram Equalization

In order to maximize the 3D image viewing effect of the object and to clarify the depth difference within the object, the histogram equalization experiment of the depth image is performed. In order to convert 16 bit source depth information into 8 bit depth information, we normalize the range of 50 - 3,500 cm from 0 - 255 from Kinect and then perform histogram equalization using 8 bit depth information using OpenCV function and MATLAB. Since the histogram of the depth image depends on the object shown on the Kinect (to create a normal video call situation), we add an object and apply depth histogram equalization to each of the three different samples. Figure 7 shows example of depth images and histogram equalization results by original depth, OpenCV and MATLAB, respectively. Figure 8 shows change of pixel value according to distance by original depth, OpenCV and MATLAB, respectively.

Figure 8. Change of pixel value according to distance by original depth, OpenCV and MATLAB, respectively.

Figure 9. Generated stereoscopic images with maximum stereoscopic 3D effect.

Histogram equalization showed that both the OpenCV and MATLAB methods showed homogeneous histogram results, but with 0 values in the hole region in the original depth image, the histogram had many frequency distributions to zero values, and the resulting histogram equalization results are not obtained. Therefore, histogram equalization using the remaining intensity values is applied except for the 0 value and the range not actually output in the experiment.

4.2 Stereoscopic 3D with 3D Hand Tracking

We assume a shift-sensor camera setup environment and reduce the computational complexity by applying the parallax shift equation for real-time display. In order to confirm real-time 3D stereoscopic images, LR images are synthesized by interlaced method. In the proposed system, the depth range is limited to 50 - 200 cm in order to enhance stereoscopic 3D effect of the object closer to the camera than the background. By restricting the depth range, it is possible to adjust the change of the values within the range in the depth image, thereby providing a more accurate 3D viewing effect on the object. However, the stereoscopic 3D effect disappears outside the depth range. Figure 9 shows generated stereoscopic images in color map.

When DIBR technique is applied to the depth image ob-

Figure 10. Example of generated stereoscopic image by interlaced method.

tained from the Kinect sensor, occlusion holes are generated similar to the depth map when the image is rendered at the left and right virtual viewpoints. To solve this problem, apply the exemplar-based image in-painting algorithm [18] to the depth map. Exemplar-based image in-painting computes the priority to fill a region in a lost region with image in-painting using the assumption that there will be enough texture information to fill the lost region around the lost region. Then, the lost area is recovered by finding an area similar to the lost area through the template matching of the peripheral area from the priority area and then copying the area.

In the DIBR algorithm, a track bar is created to control the distance of both virtual cameras (or the distance between both eyes), and a depth control experiment is performed. In order to maximize the 3D viewing effect of the object part only, the range of the depth value to be processed is limited to 50 - 180 cm. By limiting the range of the depth value, it is possible to increase the variation of the values within the range of the depth image, so that a more clear view of the object can be obtained. However, there is a disadvantage that the 3D effect disappears and it may become awkward in the outside of the range. Figure 10 shows example of generated stereoscopic image by interlaced method.

4.3 3D Hand Tracking System with AR

When the user takes a focus gesture to detect a hand, the position of the hand is detected in the depth image. The focus gesture can be defined in various ways, indicating the intention of the

Figure 11. Example of focus gesture.

user to start the gesture. In the experiment, the focus gesture is defined as the action of reaching the depth camera. Figure 11 shows example of focus gesture.

First, the searching window is set using the position of the hand detected in the previous frame depth image, and the position of the hand is tracked in the window. Next, the unscented Kalman filter [19–21] is applied to consecutive tracking points. There is noise in the tracking point due to the noise of the depth image. To eliminate this noise, we use the unscented Kalman filter to minimize the effect of noise on the tracking point path. It is necessary to deal with cases where the hand does not exist in the searching window and hand is not detected by the other area. In order to do this, it is necessary to check whether the hand enters the searching window while operating in the tracking mode in the current searching window for a predetermined time, and initialize the hand position information in another area by operating the detection mode simultaneously.

To load a virtual 3D object on the user's hand, we use OpenGL to create a cube-shaped object and create a depth map and texture for that object. Since the object to be raised on the hand is a virtual object, the depth of the object according to its size and position is arbitrarily set and the 3D image is generated according to the depth change by setting the cube to be rotated at an arbitrary angle every frame [22]. Figure 12 shows example of hand tracking with AR.

5. Conclusions

In this paper, we propose an immersive S3D system based on a depth camera. The proposed system generates stereoscopic image based on depth image and applies hand tracking technology at the same time. DIBR technology is applied to generate stereoscopic images from depth images. Applying morphology operations to depth images as pre-processing, and object segmentation using X - Z transformation and connected component labeling are applied. In order to provide stereoscopic images efficiently, face detection is applied to color images to separate

Figure 12. Example of hand tracking with AR.

foreground and background. The proposed system is designed to control stereoscopic effect. The ALBP method is applied for hand tracking. The depth-based hand tracker starts tracking the position of the hand after the focus gesture. When the position of the hand is tracked, the virtual object is augmented on the hand, and the depth of the virtual object is designed to be changed according to the position of the hand. Experimental results show that stereoscopic image generation and hand tracking are possible at the same time.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This research was supported by the Tongmyong University Research Grants 2018 (No. 2018F017). It was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2017R1C1B5017751).

References

- [1] A. Kolb, E. Barth, R. Koch, and R. Larsen, "Time-of-flight sensors in computer graphics," in *Proceedings of Eurographics 2009: State of the Art Report*, Munich, Germany, 2009.
- [2] Microsoft Kinect, Available <http://www.xbox.com/en-us/kinect>
- [3] Q. Y. Zhang, M. Y. Zhang, and J. Q. Hu, "A method of hand gesture segmentation and tracking with appearance based on probability model," in *Proceedings of the 2nd International Symposium on Intelligent Information Technology Application*, Shanghai, China, 2008, pp. 380-383. <https://doi.org/10.1109/IITA.2008.379>
- [4] H. S. Kim, G. Kurillo, and R. Bajcsy, "Hand tracking and motion detection from the sequence of stereo color image frames," in *Proceedings of IEEE International Conference on Industrial Technology*, Chengdu, China, 2008, pp. 1-6. <https://doi.org/10.1109/ICIT.2008.4608702>
- [5] S. Zhong and F. Hao, "Hand tracking by particle filtering with elite particles mean shift," in *Proceedings of Japan-China Joint Workshop on Frontier of Computer Science and Technology*, Nagasaki, Japan, 2008, pp. 163-167. <https://doi.org/10.1109/FCST.2008.9>
- [6] Q. Y. Zhang, M. Y. Zhang, and J. Q. Hu, "Hand gesture contour tracking based on skin color probability and state estimation model," *Journal of Multimedia*, vol. 4, no. 6, pp. 349-355, 2009.
- [7] Y. H. Kim, H. Lee, and S. Kim, "3D radar objects tracking and reflectivity profiling," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 12, no. 4, pp. 263-269, 2012. <https://doi.org/10.5391/IJFIS.2012.12.4.263>
- [8] H. Choi, "CNN output optimization for more balanced classification," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 17, no. 2, pp. 98-106, 2017. <https://doi.org/10.5391/IJFIS.2017.17.2.98>
- [9] B. Kim, S. Park, and E. Kim, "Hough transform-based road boundary localization," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 17, no. 3, pp. 162-169, 2017. <https://doi.org/10.5391/IJFIS.2017.17.3.162>

- [10] J. Kim, S. Yu, D. Kim, K. A. Toh, and S. Lee, "An adaptive local binary pattern for 3D hand tracking," *Pattern Recognition*, vol. 61, pp. 139-152, 2017. <https://doi.org/10.1016/j.patcog.2016.07.039>
- [11] J. Kim and C. Yoon, "Three-dimensional head tracking using adaptive local binary pattern in depth images," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 16, no. 2, pp. 131-139, 2016. <https://doi.org/10.5391/IJFIS.2016.16.2.131>
- [12] C. Fhen, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," *Proceedings of SPIE: Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291, pp. 93-104, 2004. <https://doi.org/10.1117/12.524762>
- [13] W. Chen, M. L. Giger, and U. Bick, "A fuzzy C-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images," *Academic Radiology*, vol. 13, no. 1, pp. 63-72, 2006. <https://doi.org/10.1016/j.acra.2005.08.035>
- [14] T. Chan and J. Shen, "Mathematical models for local nontexture inpaintings," *SIAM Journal on Applied Mathematics*, vol. 62, no. 3, pp. 1019-1043, 2002. <https://doi.org/10.1137/S0036139900368844>
- [15] C. Fehn, "A 3D-TV approach using depth-image-based rendering (DIBR)," in *Proceedings of the IASTED International Conference on Visualization, Imaging and Image Processing (VIIP)*, Benalmadena, Spain, 2003, pp. 482-487.
- [16] ISO/IEC JTC1/SC29/WG11, "Text of ISO/IEC FDIS 23002-3 Representation of Auxiliary Video and Supplemental Information," Document N8768, 2007.
- [17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Kauai, HI, 2001. <https://doi.org/10.1109/CVPR.2001.990517>
- [18] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," in *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, 2003. <https://doi.org/10.1109/CVPR.2003.1211538>
- [19] S. Park, S. Yu, J. Kim, S. Kim, and S. Lee, "3D hand tracking using Kalman filter in depth space," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, article no. 36, 2012. <https://doi.org/10.1186/1687-6180-2012-36>
- [20] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proceedings of the IEEE Symposium 2000 on Adaptive Systems for Signal Processing, Communication and Control*, Lake Louise, Canada, 2000, pp. 153-158. <https://doi.org/10.1109/ASSPCC.2000.882463>
- [21] S. J. Julier and J. K. Uhlmann, "New extension of the Kalman filter to nonlinear systems," *Proceedings of SPIE: Signal Processing, Sensor Fusion, and Target Recognition VI*, vol. 3068, pp. 182-193, 1997. <https://doi.org/10.1117/12.280797>
- [22] Y. B. Kim, H. C. Lee, and S. Y. Rhee, "Augmented reality of robust tracking with realistic illumination," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 10, no. 3, pp. 178-183, 2010. <https://doi.org/10.5391/IJFIS.2010.10.3.178>

Sunjin Yu received a B.S. degree from the Department of Electronics and Information Engineering of Korea University in 2003 and received his M.S. degree in Graduate Program in Biometrics and Ph.D. degree in Electrical and Electronic Engineering from Yonsei University in 2003 and 2011, respectively. He was a senior research engineer in LGE Advanced Research Institute from 2011 to 2012. From 2012 to 2013, he was a research professor in the Department of Electrical Engineering in Yonsei University and from 2013 to 2016, he was a professor in the Department of Broadcasting and Image, Cheju Halla University. He is currently a professor in the School of Digital Media Engineering, Tongmyong University. His research interests include 3D computer vision, human computer interaction, and augmented/virtual reality. E-mail: sjyu@tu.ac.kr

Eungyeol Song is a Ph.D. student in the Electrical and Electronic Engineering at the Yonsei University, Seoul, Korea. He received his B.S. degree in Computer Science at Dankook University in 2010 and his M.S. degree from the department of Electronic and Electrical Engineering at the

Dankook University in 2012. His current research interest includes hand tracking, 3D reconstruction and machine learning. E-mail: wp2001@yonsei.ac.kr

Changyong Yoon received B.S., M.S., and Ph.D. degrees in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea, in 1997, 1999, and 2010, respectively. He was a senior research engineer in LG Electronics, Inc., and LG-Nortel, and he developed system software

for the DVR and WCDMA from 1999 to 2006. From 2010 to February 2012, he was a chief research engineer in LG Display and developed the circuit and algorithms in touch systems. Since 2012, he has been a Professor in the Department of Electrical Engineering, Suwon Science College, Gyeonggi-do, Korea. His main research interests include intelligent transportation systems, pattern recognition, robot vision, and fuzzy application systems.

E-mail: cyyoon@ssc.ac.kr