3D Virtual Object Manipulation Interface for Virtual Surgery using Gesture Recognition

Ho-chul Shin*, Jae-chan Jeong and Jae-il Cho

[a]Department of Cognitive System Research,
Electronics and Telecommunications Research Institute, Rep. of Korea

**Corresponding author's email: creatrix [AT] etri.re.kr*

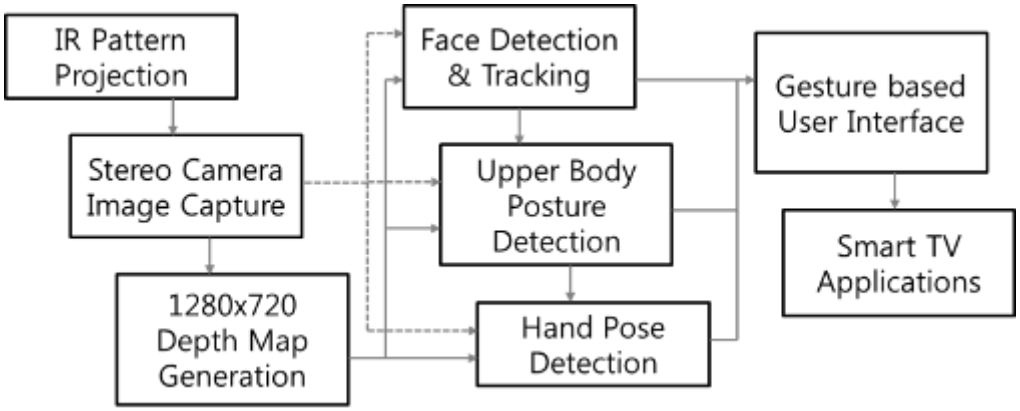
ABSTRACT---- ***In this paper we introduce long range smart TV user interface using real-time HD quality depth map and user gesture recognition. For a stable depth map generation, we designed random dot patterns, and developed DOE (Diffractive Optical Element) and an infrared laser projection module. By capturing stereo image and processing stereo matching algorithm, we developed a real-time 1280x720, 30 Hz depth map generation hardware. Using stereo image and generated depth map, we developed and integrated user detection and gesture recognition including upper body posture detection and hand pose recognition. We also introduce 3d gesture user interface with this system and virtual object 3d manipulation application. This interface will be applied to virtual surgery.***

Keyword---- Gesture Recognition, Depth Map Generation, User Interface

1. INTRODUCTION

As appearing new IT devices and services such as iPhone, iPad, smart TV, there are large demands of various intuitive UI/UX (User Interface/User eXperience). Beyond touch interface such as iPad, various non-touchable interfaces such as Microsoft Kinect and Leap Motion are appearing. In case of Microsoft Kinect I/II [1], they are showing a long range game interface through full body gesture recognition using structured light and ToF (Time of Flight) depth map, and Leap Motion [2] provides a short distance finger gesture interface. To make long range elaborate gesture interface, high quality depth map generation is required. In this study, we introduce real-time high quality depth map generation module, user detection, upper body posture and hand pose detection, intuitive 3d gesture interface and its application.

The overall structure of our system is shown in figure 1. For a stable depth map generation, an intra-red pattern is projected and stereo images are captured. Through the stereo image matching process, 1280x720 30Hz depth map is generated. Using these stereo image and depth map, we developed face detection and tracking algorithm for user detection. During the user face tracking, upper body posture detection is processed. Using the upper body posture, we can extract both hand position and recognize both hand pose. We developed 3d gesture user interface with this information, and 3d virtual object manipulation application.



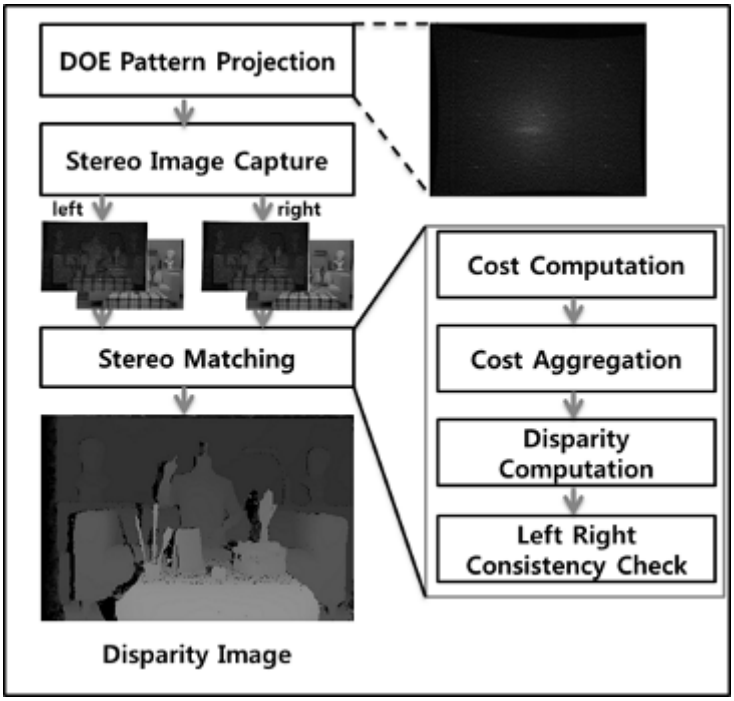
< Figure 1 System Overview >

2. HIGH QUALITY DEPTH MAP GENERATION

2.1 Depth Map Generation Overview

For a depth map generation there are several methods such as ToF, laser based scanner, and stereo visions. The ToF methods show high distance accuracy, but picture resolution is relatively low (QVGA~VGA) [1~3]. The laser based scanners show high accuracy and reliability but very high price [4]. The stereo visions have advantages such as high picture resolution, low price, but they are sensitive for environmental conditions [5]. To overcome this sensitivity, active pattern projection stereo vision has been studied [6]. By projecting artificial patterns, stereo matching result can be enhanced drastically.

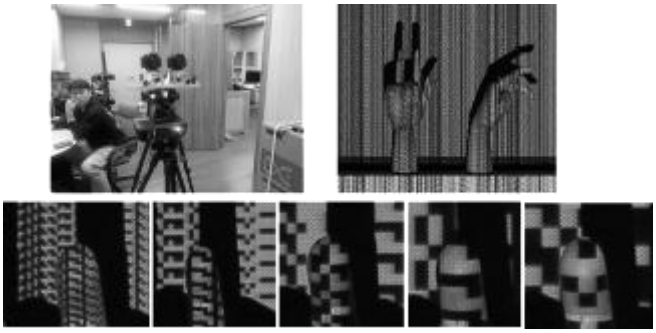
In this study, we developed intra-red pattern projection active stereo vision as shown in figure 2. Using DOE (Diffractive Optical Element), we can make a compact intra-red pattern projection module. By capturing stereo image pares with stereo camera, stereo matching process which consisted of cost computation, cost aggregation, disparity computation, and left-right consistency check is calculated. After that process disparity image can be obtained and converted to depth map.



< Figure 2 Depth Map Generation Overview using Pattern Projection >

2.2 Pattern Design and Projection

As mentioned above, to obtain stable stereo depth image, we need well designed pattern and suitable projection. As shown in Figure 3, we tested various patterns with beam projector and optimized random dot patterns [7].



< Figure 3 Experimental Setup for Pattern Design >

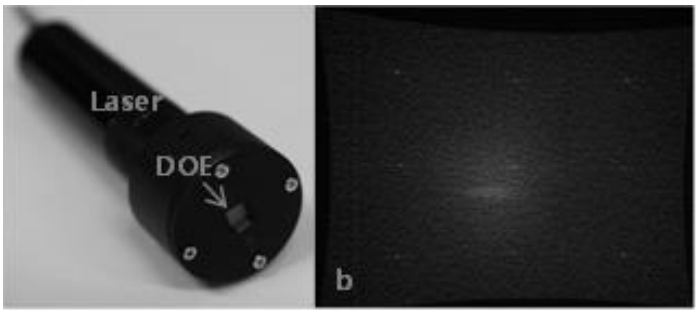
< Figure 4 Designed Patterns >

Figure 4 shows a section of the optimized final pattern, the uniqueness check result is shown in Table 1. To improve matching quality, pattern uniqueness has to be guaranteed, but it depends on the image processing window size. This window size is determined in stereo matching process which explained in next chapter.

< Table 1 Designed Pattern Uniqueness Check >

Window Size(Pixel)	3x3	5x5	7x7	9x9
Non-unique Pixel	95%	38%	0.50%	0%
Non-unique Position (Pixel)	16	81	103	212

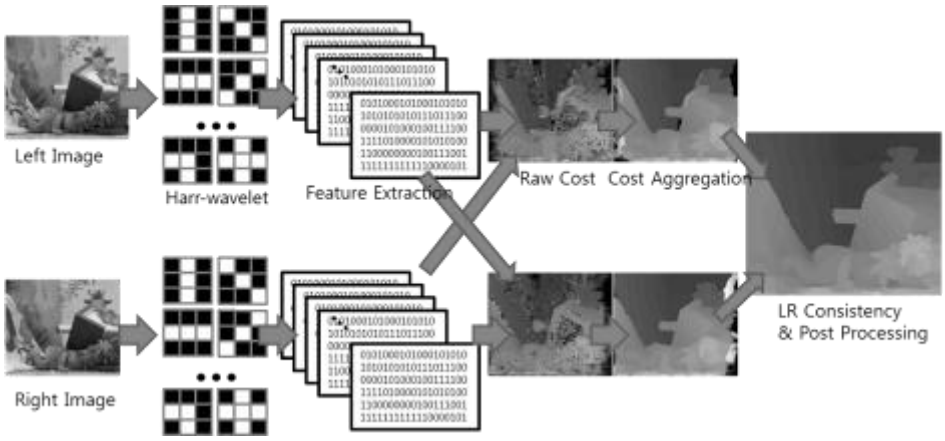
The DOE (Diffraction Optical Element) is a thin etching glass processed by semiconductor device, and we can make a compact laser pattern projector with it. As shown in figure 5, we developed a DOE and laser projector. The designed pattern is projected through 3x3 copy layer, so the final projected pattern has 633x495 pixels.



< Figure 5 Projection Module with DOE and Projected Pattern >

2.3. Stereo Matching Algorithm

After capturing pattern projected stereo images, a disparity map can be calculated through stereo matching process. The disparity map is converted to depth map with stereo camera focal length and baseline information. We developed stereo matching process [8] as shown in figure 6. Using wavelet transform and census transform, we extracted features in each pixel and calculated left and right raw costs respectively. The calculated raw costs are filtered by cost aggregation process and disparity images are calculated. By comparing left and right disparity image, noises are filtered out thorough left-right consistency check and post processing.



< Figure 6 Developed Stereo Matching Processes>

3. USER GESTURE DETECTION

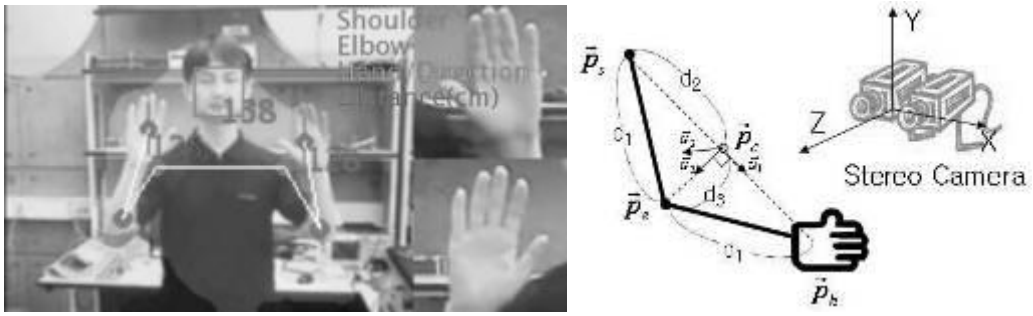
3.1 Face Detection and Tracking

For the final target of this study is smart TV user gesture interface, a user detection is required. Because user looks camera attached TV screen during the gesture interaction, a face detection and tracking is useful for user detection. We developed face detector based on census transform and Adaboost training [9]. A face tracker based on mean-shift algorithm also developed and integrated. By combining face detection and tracking, effective user detection is possible as shown in figure 7. The developed face detector can detect 20 faces simultaneously, and 3 faces that most close to camera are tracked using depth map information.

< Figure 7 Face Detection and Tracking >

3.2 User Upper Body Gesture Detection

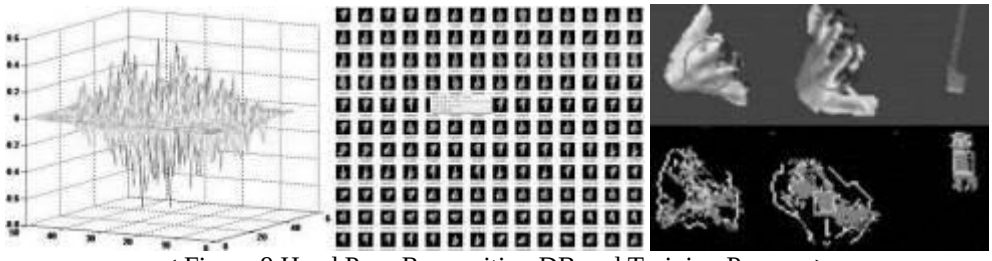
After user detection, user upper body posture is calculated. Using the user face position and distance, user shoulder is searched downward and shoulder position can be extracted using depth map. Because during the interaction, user hands are located in front of his shoulder, the hands position candidates can be suggested and identified. Once the shoulder and hand positions are identified, the elbow position can be calculated using inverse a kinematic solver [10] as shown in Figure 8. Through this process, user face, shoulder, elbow, hand positions can be obtained.



< Figure 8 Upper Body Posture Detection >

3.3 User Hand Pose Detection

Using the user hand positions, hand pose recognition is processed. First, the segmented hand region with depth map and stereo image is obtained and classified into fist, palm and others. We build sixty thousand hand image DB, and developed Adaboost based trainer and classifier as shown in figure 9.



< Figure 9 Hand Pose Recognition DB and Training Process >

3.4 User Interface Design and Algorithm Integration

Generally most intuitive and natural gesture interactions for human are hand grabbing and releasing. We defined pose change of palm to fist as grabbing, and fist to palm as releasing. The recognition algorithms which user detection, upper

body posture detection, and hand pose recognition as mentioned above, were integrated and 3d gesture interface was developed. We defined grabbing as mouse button down commend and releasing as mouse button up commend.

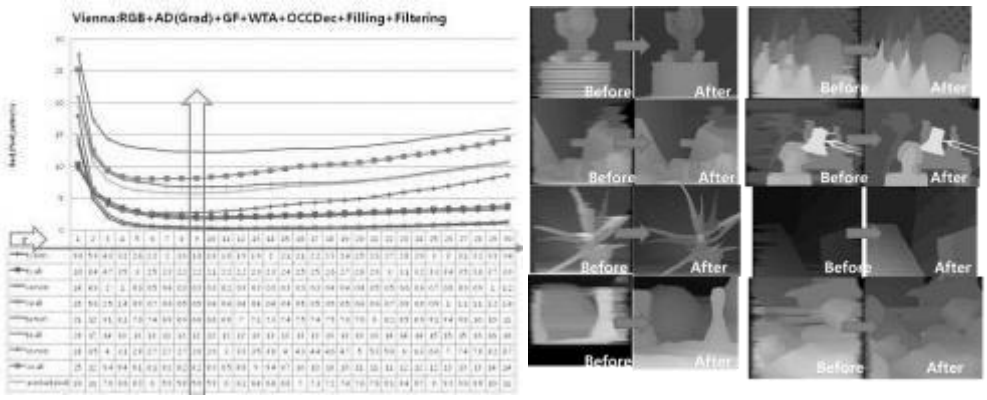
< Figure 10 Algorithm Integration and User Interface >

4. REAL-TIME HARDWARE IMPLEMENTATION

Both depth map generation and gesture recognition require high computational cost. As developed gesture recognition is composed of various image processing algorithms, CPU calculation through algorithm optimization is appropriate. But as our depth map generation requires simple and fast large data calculation, GPU (graphic processing unit) or FPGA (Field Programmable Gate Array) processing is suitable. By optimizing gesture recognition algorithms, real-time processing above 30Hz was possible on Intel i5 2GHz general PC.

4.1 Depth Map Generation Algorithm Optimization

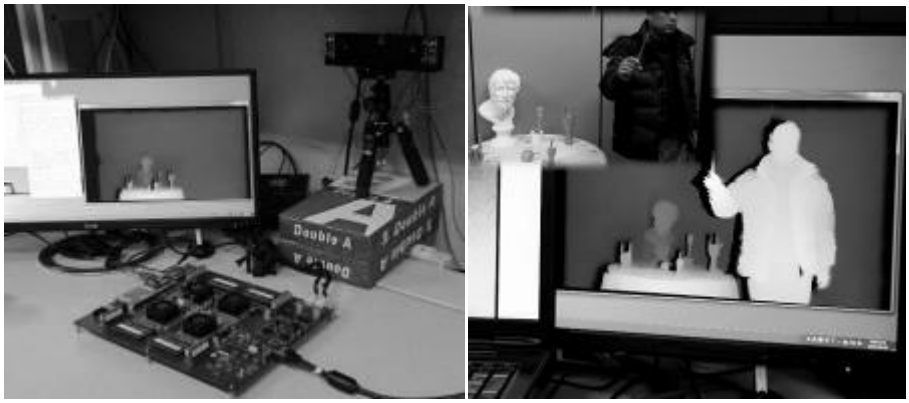
The developed depth map generation algorithm has 19 parameters, and parameter sensitivity was analyzed. Using the Middlebury stereo vision database [12], these parameters were optimized as shown in figure 11.



< Figure 11 Depth Map Generation Parameter Optimization >

4.2 Hardware Implementation

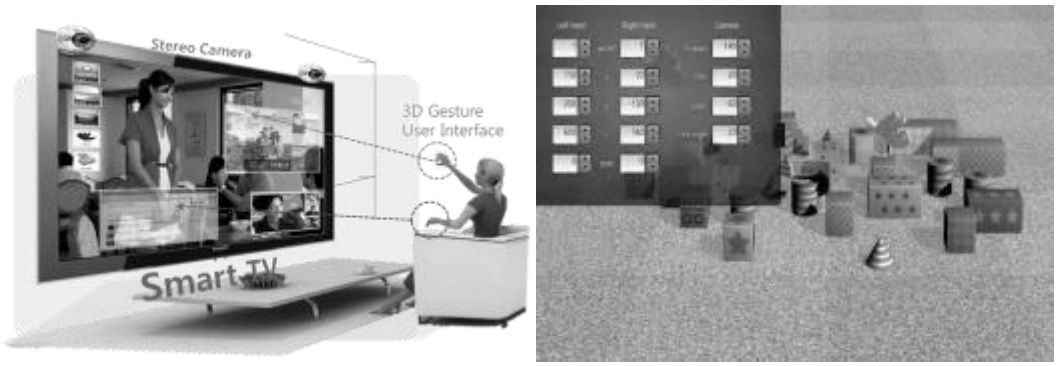
The depth map generation algorithm was implemented on FPGA [11] as shown in figure 12. This hardware can generate 1280x720, 30Hz depth map.



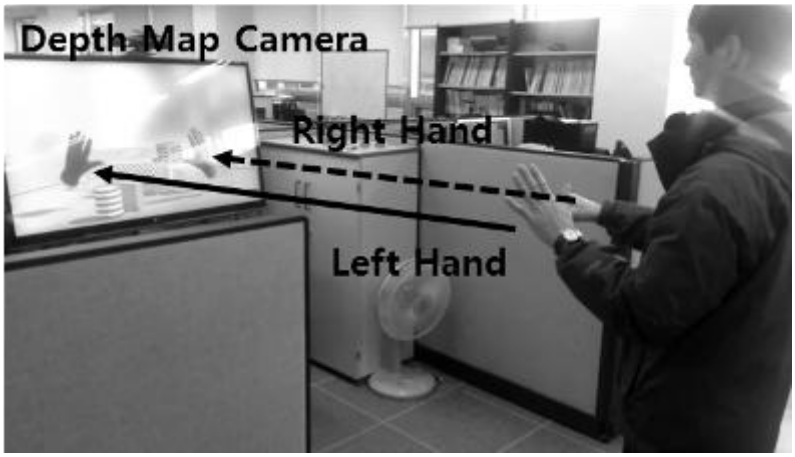
< Figure 12 Hardware Implementation and Depth Map Result >

5. 3D VIRTUAL OBJECT MANIPULATION APPLICATION

For a user interface of smart TV we developed an application contents as shown in figure 13. Using the left and right hand 3d position, grabbing and releasing gesture, user can manipulate virtual blocks with each hand, change view point, zoom in and zoom out by both hands multi-grabbing and releasing. As this contents developed for 3d display, user can see 3d picture if wearing 3d glasses.



< Figure 13 Smart TV User Interface and Developed Application >



< Figure 14 3D Virtual Block Manipulation >

6. CONCLUSION

In this paper a high quality depth map generation, gesture recognition and a 3d user interface application for smart TV were introduced. For a stable depth map generation, we developed an infrared projection module with designed pattern DOE, a stereo image matching algorithm and image processing module. The pattern projection module consisted of DOE which is designed thin etching glass and 830 nm wavelength infrared laser diode. By capturing pattern projected stereo image pair, a stereo matching algorithm was developed which contains wavelet transform, census transform, raw cost calculation, cost aggregation and left-right consistency check. This algorithm was optimized and realized into FPGA hardware module. This module can generate 1280x720 30Hz depth map. Using the stereo image and generated depth map, user detection, upper body posture and hand pose gesture recognition algorithm were developed. A 3d user interface was designed with grabbing and releasing gesture recognition. A virtual block 3d manipulation application was shown with these results. Using these results this interface will be applied to virtual surgery.

7. ACKNOWLEDGEMENT

This work was supported by ETRI R&D Program [15ZC1400, The Development of a Realistic Surgery Rehearsal System based on Patient Specific Surgical Planning] funded by the Government of Korea

8. REFERENCES

[1] http://en.wikipedia.org/wiki/Kinect
[2] https://www.leapmotion.com
[3] http://www.mesa-imaging.ch/home/
[4] Jaboyedoff, Michel Oppikofer, Thierry Abellá, et al., Use of LIDAR in landslide investigations: a review, *Natural hazards*, 61 (1), 5-28, 2012
[5] Kazmi, W., Foix, S., Alenya, G., Andersen, H.J., Indoor and outdoor depth imaging of leaves with time-of-flight and stereo vision sensors: Analysis and comparison, *Journal of photogrammetry and remote sensing*, 88, 128-146, 2014
[6] Jiang, Jun Cheng, Jun Zhao, Haifeng, Stereo Matching Based on Random Speckle Projection for Dynamic 3D Sensing, *Machine Learning and Applications,11th International Conference*, 1, 191-196, 2012
[7] Pages, J., Salvi, J., Collewet, C., Forest, J., Optimised De Bruijn patterns for one-shot shape acquisition, *Image and vision computing* 23(8), 707-720, 2005
[8] Jeong Jae-Chan, Shin Hochul, Chang Jiho, et al, High-Quality Stereo Depth Map Generation Using Infrared Pattern Projection, *ETRI journal* 35(6), 1011 - 1020 , 2013
[9] Choi, J., Yoo, S.J., Baik, S.W., Shin, H.C., Han, D., Rotation Invariant Multiple Face-detection Architecture for Smart TV, *IERI Procedia*, 6, 33 - 38 , 2014
[10] Young-Keun Kim, Ho Chul Shin, Jae Il Cho, Real-Time Human Body Posture Estimation Using a Stereo Vision Embedded System, *Computational Advances in Multi-Sensor Adaptive Processing*, Dec., 145 - 148, 2007
[11] Seung-min Choi, Jiho Chang, Dae Hwan Hwang, A FPGA based Real-time Post-Processing Architecture for Active Stereo Vision, *ISCE*, P2-46609, 2014
[12] http://vision.middlebury.edu/stereo