3D HAND LOCALIZATION BY LOW COST WEBCAMS

Cheng-Yuan Ko, Chung-Te Li, Chen-Han Chung, and Liang-Gee Chen
DSP/IC Design Lab, Graduated Institute of Electronics Engineering
National Taiwan University, Taiwan, R.O.C

ABSTRACT

In recent years, depth sensors, such as Kinect provides new opportunities for Human-Computer Interaction (HCI). However, for more universalization of depth sensors in consumer electronics, the cost of the sensors should be considered. In this paper, we proposed an algorithm for 3D hand localization by two commodity low cost webcams. Because of the noise produced by the low cost webcams, the depth quality is not very good. Ac-cording to the poor quality depth map, our algorithm can still to do the 3D hand localization. The proposed algorithm can provides 3D hand localization information for applications, such as interactive 3DTV.

Keywords: 3D hand localization, depth map, low cost, HCI

1. INTRODUCTION

In recent years, motion sensing or gesture recognition is an active research region in Human-Computer Interaction (HCI). There are two important information should be considered in human-device interaction: the distance from user to system and the location of user's hand. Previous papers [1] [2] use calibration-free captures to detect the user's distance from system quickly. However, when we only have the user's distance from system, we can just do some simple interaction with system. Because of hand gesture is one of the most intuitive and nature ways for people to communicate with machines, so system have to get the user's hand 3D localization, and thus the user can do more complex control or interaction with system.

Hand localization is not a new field in Human-Computer Interaction (HCI) or pattern recognition. Ying Wu et al. [3] pro-posed an algorithm based on skin-color detection. They present a tracking system based on the self-organizing color segmentation. A 1-D self-organizing map (SOM) is used to clustering the HSI color space automatically. However, there are still some chal-lenging problems related to hand localization by color based segmentation [4], such as cluttered background with color dis-tracters and changing lighting conditions. To overcome these problems, Enver Sangineto et al. proposed an algorithm based on machine learning. They have shown a hand detection system working in real time and able to localize hands independently of the person's identity and the hand position and which is robust in different lighting condition changes as well as in cluttered back-ground images [5]. Nonetheless, these papers still just can provide 2D hand localization. Thus, for more advanced application, such as interact with virtual object in 3DTV, we have to get the user's 3D hand localization.

2. RELATED WORK

For 3D hand localization, depth sensor should be used to get the depth information of user's hand. Depth sensors are classified with active sensor and stereo camera system. Active sensors such as IR-based camera or Time of Flight (TOF) camera can provide more reliable and good quality depth map. M. Van den Bergh et al. presents a hand gesture interaction system based on a Time of Flight (TOF) camera and a RGB camera. An improved hand detection algorithm is introduced based on depth and adaptive skin color detection [6]. However, the cost of active camera is more expensive to stereo camera system.

In this paper, unlike traditional 2D hand localization which are based on skin-color or feature extraction, or RGB-D cameras use both skin color detection and depth, we only use two low cost commodity webcams to build a stereo system and get the depth map by stereo matching. The rest of this paper is organized as follows. Section 3 describes the proposed algorithm. The experimental setup and results are described in Section 4. Finally, we conclude this paper in Section 5.

3. PROPOSED ALGORITHM

In this section, we introduce the proposed 3D hand localiza-tion algorithm. The proposed processing can be applied for any system with two cameras. Our proposed algorithm can be com-posed of four steps: (1) Calibration and cutting input captures. (2) Stereo matching and face detection. (3) Filter out the background by depth map. (4) Decide the hand region and calculate the real distance from user's hand and webcams. The overall system flow is as follow in Fig.1.

3.1 Calibration and cutting input captures

The objective of stereo camera calibration is to estimate two kinds of parameters of each camera: external parameters and internal parameters. According to these parameters, the 3D posi-tion of a point in a scene will be identified and match in stereo image pair, which can be determined by the method of triangula-tion [7]. In this work, we use OPENCV stereo calibration method to calibrate input captures [8]. We use 30 pairs calibration source image to calibrate these two commodity cameras, and the part of calibration source images are shown as follow in Fig. 2. After calibration process, we cut the leftmost part in left view and the rightmost part in right view to enhance the accuracy of depth which is calculated by stereo matching.

3.2 Stereo matching and face detection

The goal of stereo matching is to determine disparities that are indicating the difference in locating corresponding pixels. We implement the stereo matching algorithm by Yen-Chieh Lai et al. [9]. The stereo matching results after post processing is quite good, as shown in Fig. 3. But in our algorithm, the depth map without post processing is used.

On the other hand, we use Haar-like feature classifier [10] in OPENCV to detect user's face from rectified left view and right view simultaneously. Finally, we get the face location in the depth map.

3.3 Filter out the background by depth map

After above process, we can get the localization of user's face. To avoid the calculation error of depth value on the center pixel of the face, we calculate the average depth value in the center region of detected face, i.e.

$$Estimated\ Depth\ of\ face = \frac{\sum \frac{1}{4}\ number\ of\ pixels\ depth\ value\ in\ detected\ face\ nearest\ to\ the\ center}{\frac{1}{4}\ \text{number of pixels in detected face}} \quad (1)$$

And here, we require the user to cooperate in one aspect which is the user's hand is in front of user's face (it is a reasona-ble requirement in HCI). According to this, we can remove all the background behind user. In theory, we can segment the hand region easily right now, but because of noisy depth map, we can only get the result as shown in Fig. 4.

3.4 Decide the hand region and calculate the real distance from user's hand and webcams

In this step, we decide the hand region by finding the area which is max and the pixel value in each pixel are almost the same. Then, we considered it as the hand region and rule out other regions. Now, we can calculate the depth information of user's hand easily, and from the regression of face disparity and real distance [2], we can estimate the distance between user's hand and webcams to achieve 3D hand localization.

4. EXPERIMENTAL SETUP AND RESULTS

In this section, we present our results and describe the ex-perimental setup in detail. We use two commodity webcams to build a stereo vision system. The resolution of input images for left view and right view are both VGA (640x480). After the calibration and cutting process, the resolution of left view and right view are both dropped to 528x464. The operational range in this experiment is about 30 cm to 150 cm. We use the regression line which is made by face detection result [2], i.e.

$$y = 0.0037x + 0.0001 \quad (2)$$

Here, "y" represents the reciprocal of disparity of user's face and "x" represents (the real distance between user and webcams)/30 as shown in Fig. 5. According to the depth value of user's hand in depth map, we can reverse the disparity to the actual distance easily.

The experimental results are as follow in Fig. 6, Fig. 7 and Fig. 8. We choose three representative hand shapes to test our 3D hand localization system, they are paper, scissors and rock. We can find out that for each case, we can detect the hand localization by poor quality depth map very accurately.

5. CONCLUSIONS

In this paper, we proposed an algorithm for 3D hand localization by two commodity webcams. The proposed algorithm is suit for any low cost cameras to do the 3D hand localization. According to the experiment results, it is robust to any hand shape.

Our proposed algorithm can provides 3D hand localization information for applications, such as interactive 3DTV.

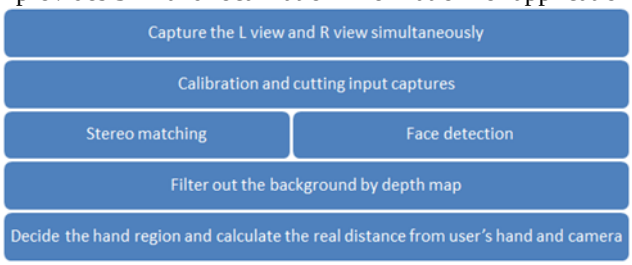
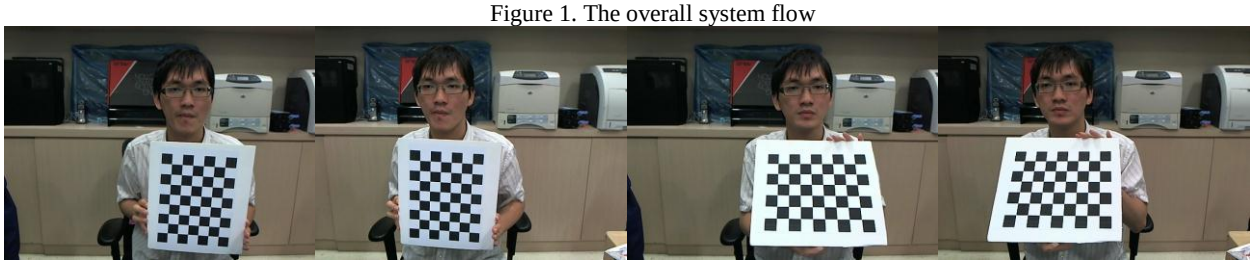


Figure 1. The overall system flow



(a) (b) (c) (d)

Figure 2. Part of calibration source images. (a)(b) and (c)(d) are two calibration source pairs respectively. (a)(c) are in left view and (b)(d) are in the right view.

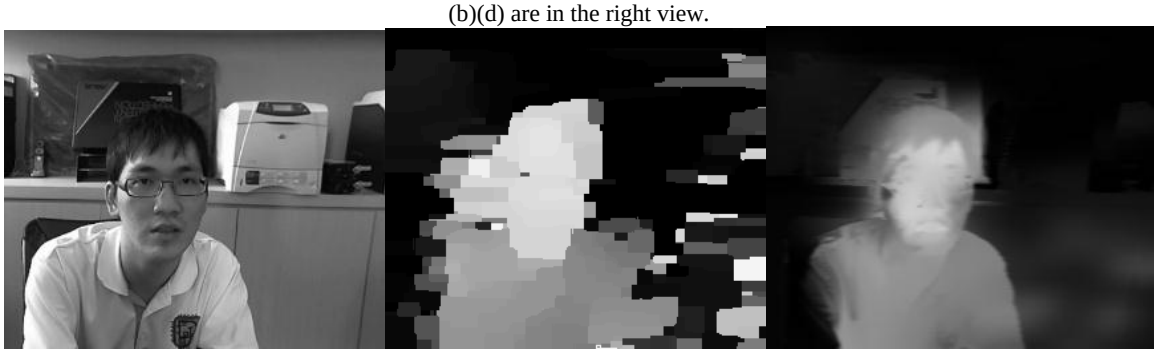


Figure 3. Post processing for depth map. (a) input capture (b) poor quality depth map without post processing (c) depth map with post processing

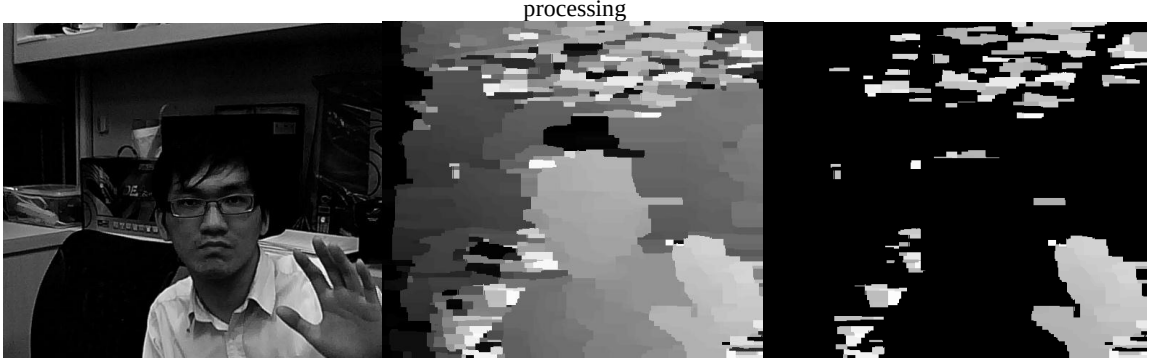


Figure 4. Filter out the background only by depth of user's face. (a) input capture (b) poor quality depth map (c) filter out the background only by depth of user's face

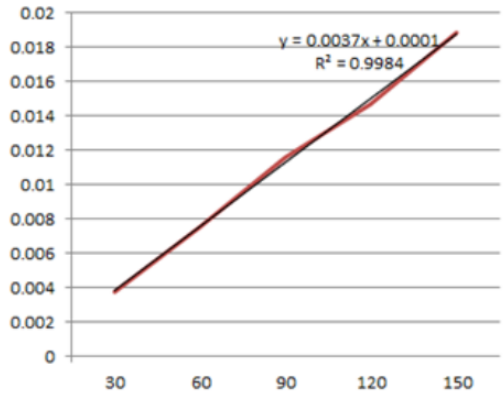
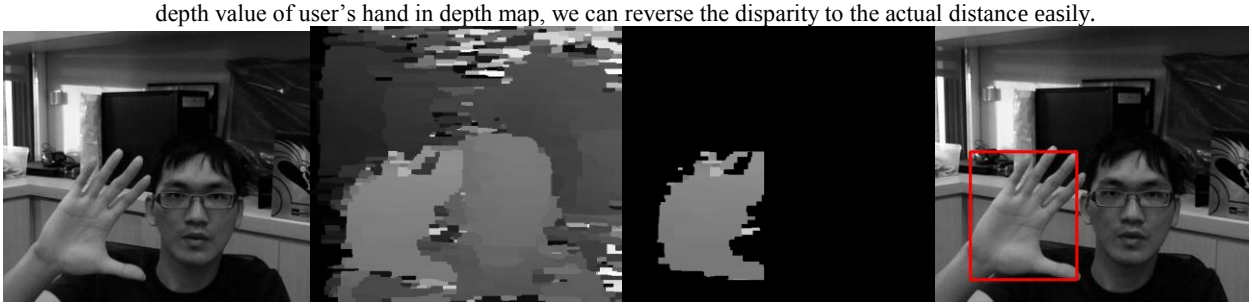
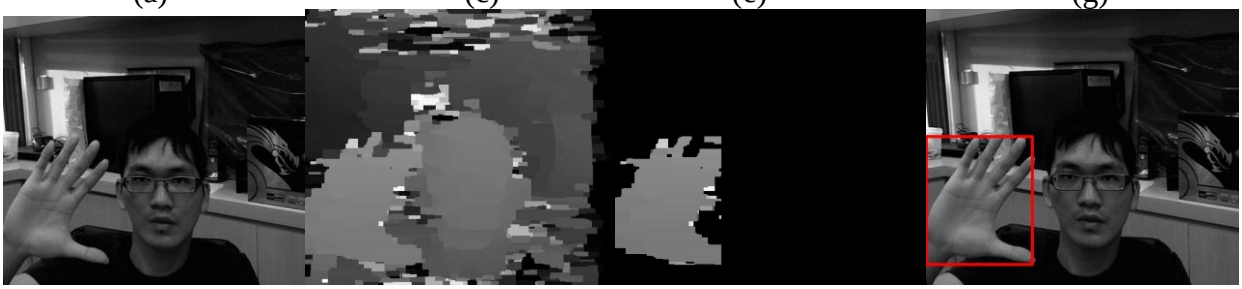


Figure 5. Linearity of actual physical distance and the reciprocal of disparity by face detection. Y-axis represents the reciprocal of disparity, X-axis represents distance between user and cameras. We use these data to find the regression line so that according to the depth value of user's hand in depth map, we can reverse the disparity to the actual distance easily.

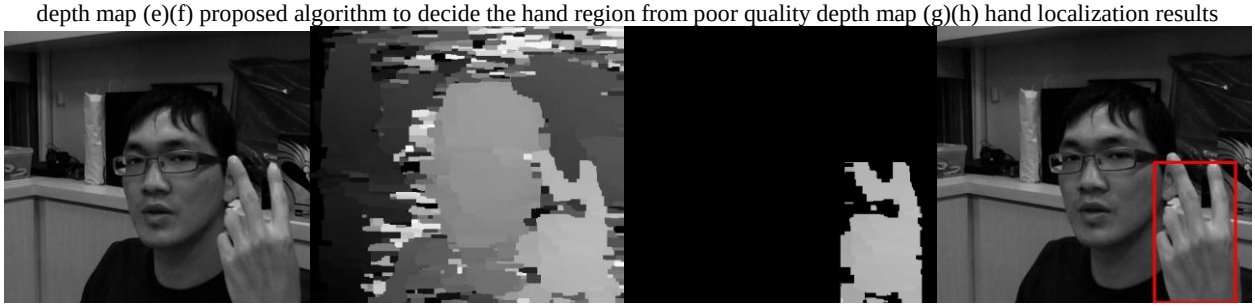


(a) (c) (e) (g)

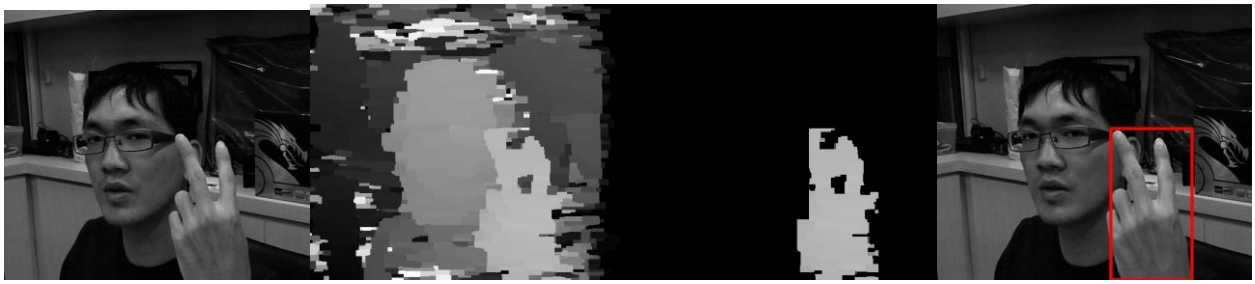


(b) (d) (f) (h)

Figure 6. Paper shape for testing 3D hand localization system. (a)(b) input capture for left view and right view (c)(d) poor quality depth map (e)(f) proposed algorithm to decide the hand region from poor quality depth map (g)(h) hand localization results

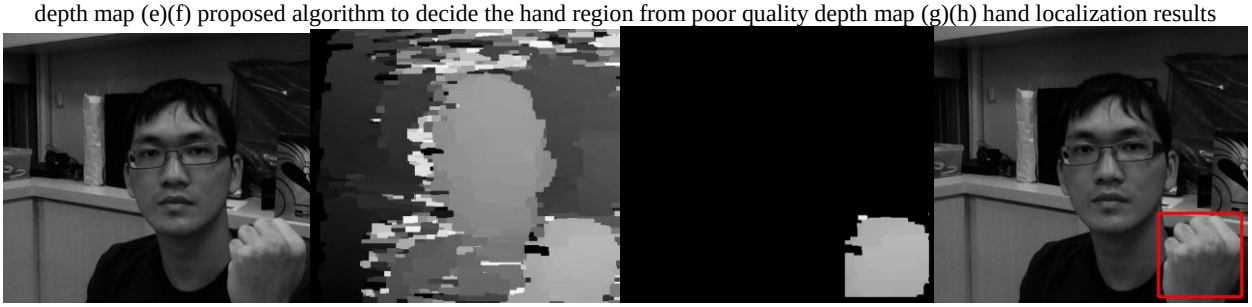


(a) (c) (e) (g)

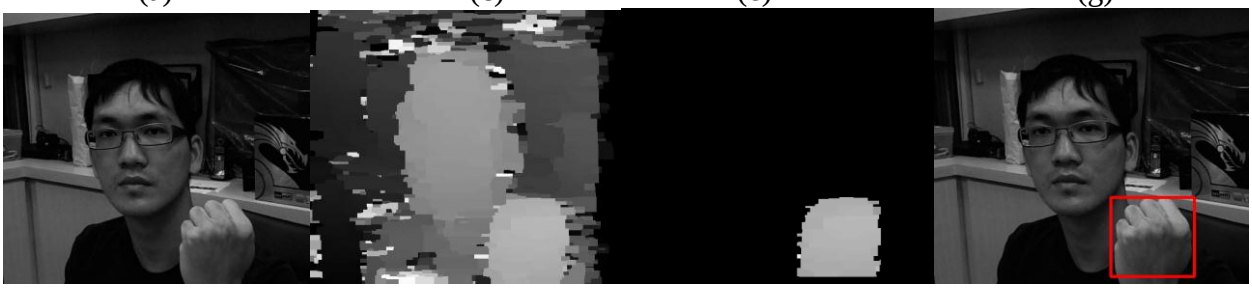


(b) (d) (f) (h)

Figure 7. Scissors shape for testing 3D hand localization system. (a)(b) input capture for left view and right view (c)(d) poor quality depth map (e)(f) proposed algorithm to decide the hand region from poor quality depth map (g)(h) hand localization results



(a) (c) (e) (g)



(b) (d) (f) (h)

Figure 8. Rock shape for testing 3D hand localization system. (a)(b) input capture for left view and right view (c)(d) poor quality depth map (e)(f) proposed algorithm to decide the hand region from poor quality depth map (g)(h) hand localization results

6. REFERENCES

[1] Cheng-Yuan Ko, Chung-Te Li, Chien Wu, and Liang-Gee Chen, "An Efficient Method for Extracting the Depth Data from the User," in *International Conference on 3D systems and Applications (3DSA)*, Hsinchu, Taiwan, June 2012.

[2] Cheng-Yuan Ko, Chung-Te Li, and Liang-Gee Chen, "Acquire User's Distance by Face Detection," *submitted.*

[3] YingWu, Thomas S. Huang "Robust Real-time Human Hand Localization by Self-Organizing Color Segmentation," *IEEE Int'l Workshop on Recognition, Analysis and Tracking of Face and Gestures in Real-Time Systems*, Corfu, Greece, 1999.

[4] Rick Kjeldsen, John Kender, "Finding Skin in Color Images," in *Proc. the Second International Conference on Automatic Face and Gesture Recognition*, pp.312-317, 1996.

[5] Enver Sanginetoa, Marco Cupellib, "Real-time viewpoint-invariant hand localization with cluttered backgrounds," *Journal of Image and Vision Computing*, Volume 30 Issue 1, Pages 26-37, January, 2012

[6] M. Van den Bergh and L. Van Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," In *Proc. of the IEEE Workshop on Applications of Computer Vision (WACV 2011)*, 2011

[7] J. Weng, P. Cohen, and M. Herniou, "Camera Calibration with Distortion Models and Accuracy Evaluation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 10, pp. 965-980, Oct. 1992

[8] Guo Yan Xu, Li Peng Chen, and Feng Gao, "Study on binocular stereo camera calibration method," in *International Conference on Image Analysis and Signal Processing (IASP)*, 2011

[9] Yen-Chieh Lai, Chao-Chung Cheng, Chia-Kai Liang, and Liang-Gee Chen, “Efficient message reduction algorithm for stereo matching using belief propagation,” in *International Conference on Image Processing (ICIP)*, 2010.
[10] Paul Viola and Michael J. Jones, “Robust real-time object detection,” on *International journal of computer vision*, 2004.