High Accuracy User's Distance Estimation by Low Cost Cameras

Cheng-Yuan Ko,Chung-Te Li, Chen-Han Chung, and Liang-Gee Chen
Graduate Institute of Electronics Engineering
National Taiwan University
Taipei, Taiwan (R.O.C)

Abstract—In this paper, we propose an algorithm using only two commodity webcams without any calibration to detect the distance between users and the display. According to the experimental results, the error is always less than 2cm and the operation range is up to 406cm.

Our proposed algorithm can provide users' depth information with high accuracy from calibration free stereo capture image pair for the application, such as interactive 3DTV or user-aware auto-stereoscopic display.

Keywords*- calibration free; stereo; distance estimation; low cost; high accuracy;*

I. INTRODUCTION

In recent years, the detection of user's location and motion sensing are very active research areas in Human computer interaction(HCI). The recent introduction of inexpensive depth sensors that work at frame rate offers new opportunities to address this difficult problem. Kinect is a representative product [1]. Many researches use Kinect to get the scene's depth map and extract the user's depth information for other application such as gesture recognition [2]. Although Kinect can provides a convenient, fairly accurate depth measurement method. However, due to the Kinect detects the user real time action with the infrared through a built-in VGA camera sending active laser. It will determine the user position by the laser reflection process within the Kinect scan range, at the same time, all objects are marked "depth field". And the Kinect need a motor on the base to adjust the direction, so the cost is more expensive.

Another method is using stereo camera. Among the previous projects related to people detection and tracking using stereo camera system we find the one by Darrel et al. [3]. Darrel et al. present an interactive display system which can detects and tracks multiple people. Three modules: a skin detector, a face detector and the disparity map provided by a stereo camera to provide the People detection function. First, in the disparity map, their algorithm detects independent objects (blobs), and they are treated as candidate to people. Then, those regions are analyzed by the skin color detection to identify that could be related to skin or not. Finally, for those regions are selected, a face detector is applied. These three parts are merged in order to detect and track multiple people. However, R. Mu˜noz-Salinas et al. [4] consider that a main drawback to their approach can be pointed out as the system relies on a predefined color model to detect skin, degradation on the tracking performance can be expected when the illumination conditions differ significantly from the training ones.

Using Stereo camera captures left view and right view simultaneously and then do the stereo matching process to find out the user's depth is another way [5]. However, for stereo matching requirements, the input of left view and right view should be calibrated.

Therefore, we proposed an algorithm by face and mask based stereo matching fusion that can use uncalibrated capture inputs left view and right view.

The rest of the paper is organized as follows. Section 2 describes the proposed face and mask based stereo matching fusion algorithm. The experimental setup and results are described in Section 3. Finally, we conclude this paper in Section 4.

II. PROPOSED ALGORITHM

In this section, we introduce the proposed face and mask based stereo matching fusion algorithm. We have proposed an algorithm for the user distance estimation by mask based stereo matching (MBSM) [6]. However, for real interactive 3DTV, the MBSM algorithm cannot work very well in the short distance due to the incomplete mask, see Fig.1. Thus, we proposed an algorithm based on data fusion concept to overcome this problem and enhance the accuracy.

The proposed processing can be applied for any system with two commodity cameras. Our proposed algorithm can be composed of four steps: (A) face detection based stereo matching (FDBSM) (B) refined mask based stereo matching (MBSM) (C) multi-cue data fusion for final disparity calculation (D) disparity transfer to user's real distance. The overall system flow is as follow in Fig. 2.

A. Face Detection Based Stereo Matching (FDBSM)

First, we use Haar-like feature classifier [6] to detect the user's face. Refer to [7], the algorithm can be composed of two parts: (1) Face detection without false detection. (2) Calculate disparity of detected face. The main concept of first step is using the geometric relationship between face width and disparity to rule out some impossible pairs which are false detection. So after this step, we can get a pair of robust face detection result. The second step is that to calculate disparity of user's face by using equation as shown as follow:

$$\text{Disparity} = |Lx - Rx| \qquad (1)$$

Where Lx is the center of user's face in left view and Rx is the center of user's face in right view. After getting the disparity, we transform the disparity to the real distance between the user and the camera by geometric relationship. Finally, we can get the user's distance D_F.

B. *Refiened Mask Based Stereo Matching(MBSM)*

We have proposed an algorithm for the user distance estimation by mask based stereo matching (MBSM) in [8]. The MBSM algorithm can be composed of two steps: (1) Do the background subtraction process to get the mask map. (2) Mask based stereo matching to get the user's disparity and use geometric relationship transforming to the user's distance.

In this paper, we refined this algorithm by adding one step t the original first step. Because of in step A have done the face detection, so here we do the connected component with face to refine the masks after background subtraction. Then, we use the noise-free masks to calculate the disparity and get another user's distance D_M. The overall flow chart of proposed refined MBSM is shown in Fig. 3.

C. *multi-cue data fusion for final disparity calculation*

Now, in this step, we have both the user's distance D_F and D_M which are estimated by face detection and MBSM, respectively. The final estimated distance is a fusion of these two values. The following is the expression.

$$D_E = \alpha * D_M + (1-\alpha) * D_F \qquad (1)$$

Note that in equation (1), D_E is the final value of estimated distance. And α is a weighting factor which has Gaussian distribution when D_F between 40cm-130cm. When D_F = 40cm, α = 0, and when D_F = 130cm, α = 1. The expression of α is shown as follow:

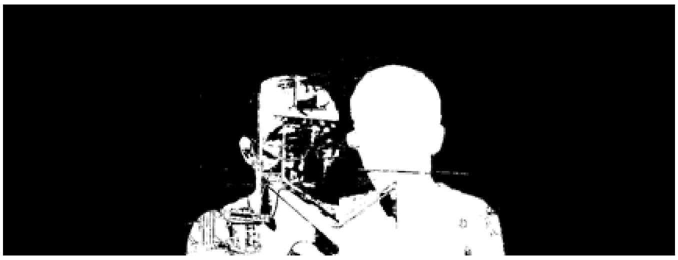
$$\alpha = 2\int_{40}^{D_F} \frac{1}{180} e^{-\frac{\pi(x-130)^2}{180^2}} dx \qquad (2)$$

D. *disparity transfer to user's real distance*

The last step is to transfer disparity to real world distance between cameras and user. The main concept we used in this step is a geometric relationship. In human vision system, the binocular cue is the most important information for human to sense the depth in short distance(<10m). So we use formula (3) to calculate the real world distance:

$$\frac{D_x}{D_{40}} = \frac{x}{40} \qquad (3)$$

Where D_{40} denote that disparity of user in front of camera with 40cm, and x is the distance of user current stand in front of camera. By this equation, the final real world distance can be calculated.



(a) **(b)**

Figure1. Incomplete mask of MBSM in short distance.

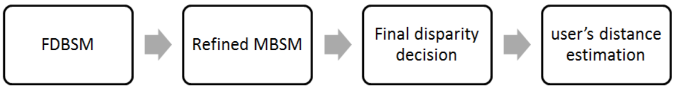


Figure2. The overall system flow for proposed algorithm

III. Experimental Setup and Results

In this section, we present our experimental setup and result. We measured the linearity of actual physical distance versus the estimated distance as the basis to judge the accuracy. In our experiment, we use two Logitech C910 webcams to build a human vision like stereo vision system. In the beginning, the user stands in front of cameras which the distance between user and cameras is 70 cm. And we increase the distance between user and cameras incrementally up to 410 cm every 30 cm. The results are show as Fig. 4 and Table 1.

Fig. 5 showed the observations, the correlation coefficient of physical distance and the estimated distance is up to 0.9999. It is worth mentioning that the advantage of the proposed method is that we can get very good results as long as the difference of horizontal level of two cameras are not too much, because we only use the horizontal disparity to estimate distance.

There is one more important thing is worth to noting. The performance of proposed data fusion algorithm compared with masked based stereo matching (MBSM) we proposed in [8] has a significant improvement. Not only the linearity of actual physical distance versus the estimated distance from 0.9985 to 0.9999, but also to promote a substantial decline in the estimated error in short distance. The results shows that proposed algorithm overcome the incomplete mask problem in MBSM, Fig. 4 shows this result.

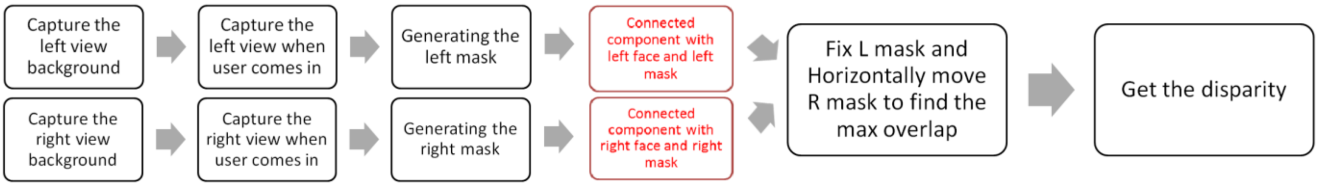
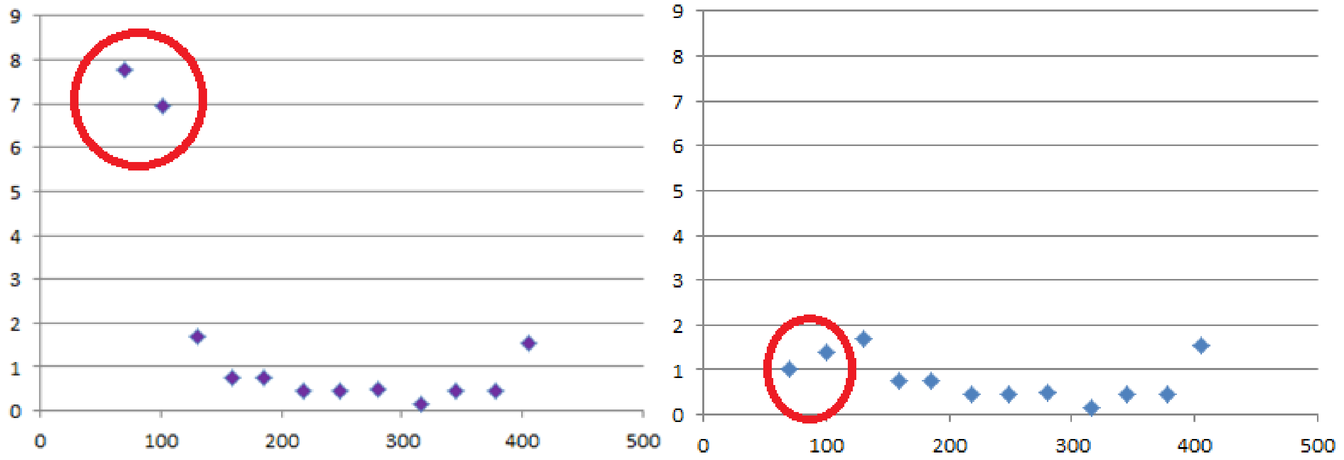

Capture the left view background
Capture the left view when user comes in
Generating the left mask
Connected component with left face and left mask
Capture the right view background
Capture the right view when user comes in
Generating the right mask
Connected component with right face and right mask
Fix L mask and Horizontally move R mask to find the max overlap
Get the disparity

Figure3. The whole algorithm flow of refined mask based stereo matching (MBSM). The red blocks is the difference between refined MBSM and original MBSM proposed in [8].



(a) (b)

Figure4. Comparison of estimated error by two algorithms. X- axis represents the user's distance, Y-axis represents the estimated error. Note that in the short distance, the estimated error is significantly improved in proposed algorithm.(a)estimated error only by MBSM. (b)estimated error in proposed algorithm.

TABLE 1. GROUND TRUTH AND ESTIMATED USER'S DISTANCE

Ground Truth(cm)	Estimated Distance(cm)	Ground Truth(cm)	Estimated Distance(cm)
70	68.99837	249	249.4571
100	98.61867	281	281.4735
131	132.6998	316	315.8353
159	158.2313	345	345.4355
185	184.2547	378	378.4584
218	218.4574	406	407.5446

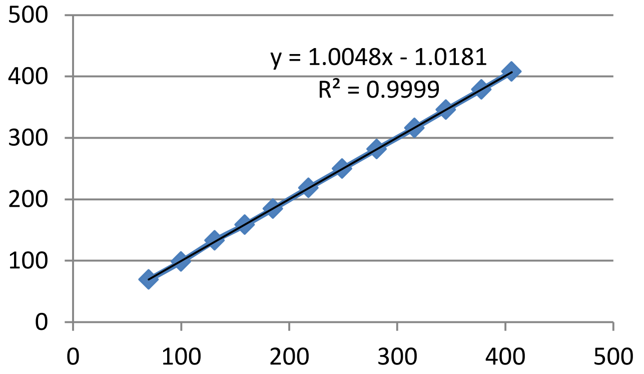


Figure5. Linearity of actual physical distance and the estimated distance by proposed algorithm. X-axis represents distance between user and cameras; Y-axis represents the estimated distance by proposed algorithm.

IV. CONCLUSIONS

In this paper, we proposed a multi cue data fusion method using only two commodity cameras to detect user's location with very low complexity computation and calibration free. According to the experimental result, the correlation coefficient of actual physical distance and the estimated distance is up to 0.9999.

Because of the input of traditional stereo vision system using stereo matching method must use the two calibrated images, the advantage of the proposed method is that we can get very good results as long as the difference of horizontal level of two cameras is not too much.

Proposed algorithm can provide user's depth information with high accuracy from calibration free stereo capture image pair for the application, such as interactive 3DTV.

ACKNOWLEDGMENT

This work is partially supported in part of Himax Technologies, Inc., Taiwan, (R.O.C.).

REFERENCES

[1] MICROSOFT KINECT. http://www.xbox.com/kinect.

[2] Oreifej, Omar, and Zicheng Liu, "HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences," *Computer Vision and Pattern Recognition (CVPR),* June 2013

[3] T. Darrell, G. Gordon, M. Harville, J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection" International Journal of Computer Vision 37 (2000) 175–185.

[4] R. Mu˜noz-Salinas, E. Aguirre, and M. Garc´ıa-Silvente, "People detection and tracking using stereo vision and color," Image and Vision Computing, vol. 25, no. 6, pp. 995–1007, 2007.

[5] Cheng-Yuan Ko, Chung-Te Li, Chen-Han Chung, and Liang-Gee Chen, "3D hand localization by low-cost webcams," *IS&T/SPIE Electronic Imaging (IS&T/SPIE EI),* Jan. 2013

[6] Paul Viola and Michael J. Jones, " Robust real-time object detection," International journal of computer vision, 2004.

[7] Cheng-Yuan Ko, and Liang-Gee Chen, "Acquire User's Distance by Face Detection, in IEEE 17th International Symposium on Consumer Electronics (ISCE), Hsinchu, Taiwan, June 2013.

[8] 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.