An Efficient Method for Extracting the Depth Data from the User

Cheng-Yuan Ko, Chung-Te Li, Chien Wu, and Liang-Gee Chen*

DSP/IC LAB, Graduate Institute of Electronics Engineering, National Taiwan University, Taiwan,
{kevin, ztdjbdy, mpemial, lgchen}@video.ee.ntu.edu.tw

Abstract

In this paper, we propose a background subtraction based 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 distance and the reciprocal of disparity is up to 0.9985.

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.

Keywords --- background subtraction, calibration free, people detection, depth

1. 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][3]. 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. [4]. 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. [5] 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. However, for stereo matching requirements, the input of left view and right view should be calibrated.

Therefore, we proposed a background subtraction based method with mask based stereo matching 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 background subtraction based method with mask based stereo matching algorithm. The experimental setup and results are described in Section 3. Finally, we conclude this paper in Section 4.

2. Proposed algorithm

In this section, we introduce the proposed background subtraction based method with mask based stereo matching algorithm. The proposed processing can be applied for any system with two cameras. Our proposed algorithm can be composed of two steps: (1)Do the background subtraction process to get the mask map (2) Mask based stereo matching. The overall system flow is as follow in **Fig.1**.

2.1. Background subtraction

Background subtraction is a widely used approach in video surveillance system for detecting moving objects from static cameras. The basic principle of the approach is that to detect the moving objects from the difference between the reference frame and a current frame. We often call that the reference frame "background image", or "background model" [6]. In our proposed algorithm, the first step is when the system is start-up, the scene of the environment will be shot individually by the left and right camera, and the cameras continue shooting. At this point the user enters the environment, two cameras each captures the environment contains the user's information. Now, we have two captures for left view and right view, we call the captured images when system was start-up as "background image", and call the captured images after user entered as "current image". Here, we do the background subtraction process, if the RGB pixel value

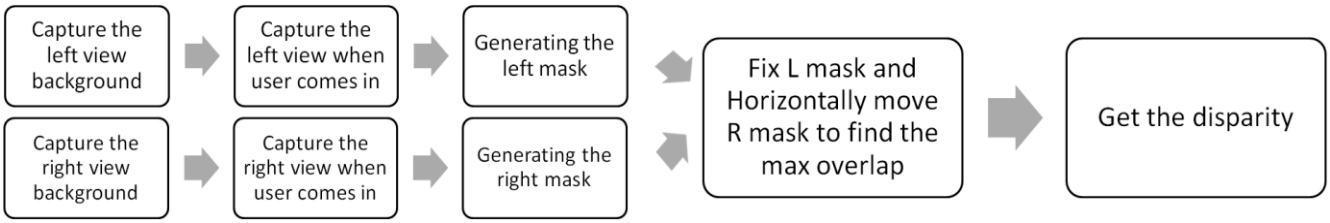

Fig.1 The overall system flow chart for proposed algorithm.

difference in "background image" and "current image" is more than a certain threshold value we set, it is considered as the part of the user's body. Then, it produced the two individual masks as **Fig. 2**.

2.2. Mask based stereo matching

After background subtraction process, we get two results of right mask and left mask which are shown as Fig. 2 (c),(f). Now we want to calculate the disparity from left view and right view to estimate the user's distance from camera. A simple method to calculate the disparity from these two masks is that for left mask and right mask, we find out the centroid for each mask, and represents $L_c(L_x,L_y)$ for left mask centroid and $R_c(R_x,R_y)$ for right mask centroid respectively. Then the x-axis difference of right mask centroid and left mask centroid is disparity, i.e.

$$\text{Disparity} = |L_x - R_x| \quad \textbf{(1)}$$

However, the accuracy of this method is not good enough because of L_x and R_x cannot be calculated preciously without any process for these two masks. So we proposed mask based stereo matching method to get the disparity preciously without calculate the centroid. First, we fix left mask and move the right mask horizontally to find the largest overlap region of two masks. Then, we can calculate the right mask moving for the number of pixel to get the maximum overlap region is the disparity.

3. Experimental results

In this section, we present our experimental result and measure the linearity of actual physical distance versus the reciprocal of disparity as the basis to judge the accuracy. In our experiment, we use two Logitech C910 webcams to build a stereo vision system. In the beginning, the user stands in front of cameras which the distance between user and cameras is 80 cm. And we increase the distance between user and cameras incrementally up to 410 cm every 30 cm. The results are show as **Fig. 3** and **Table. 1**.

Fig. 3 showed the observations, the correlation coefficient of distance and the reciprocal of disparity is up to 0.9985. 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, as shown in Fig 2. (a) and (d).

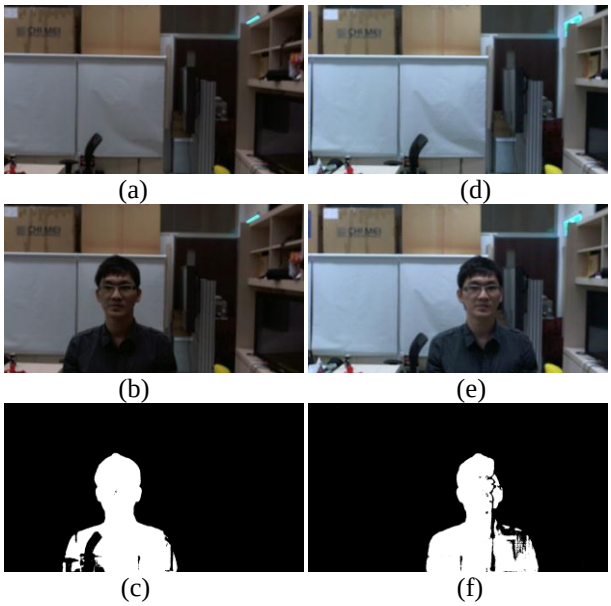

Fig.2 Background subtraction flow and results. (a) the "background image" of right view (b) the "current image" of right view (c) the right view result mask after background subtraction (d) the "background image" of left view (e) the "current image" of left view (f) the left view result mask after background subtraction

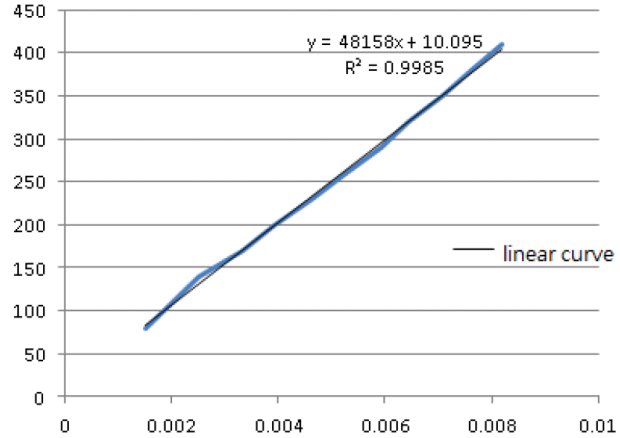

Fig.3 Linearity of actual physical distance and the reciprocal of disparity. X-axis represents the reciprocal of disparity, Y-axis represents distance between user and cameras.

We can find out that the chair in left view is not as high as it in right view.

3. Conclusions

We proposed a background subtraction based method using only two commodity cameras to detect

Distance(cm)	Disparity(pixel)
80	665
110	565
140	398
170	301
200	254
230	216
260	190
290	169
320	155
350	141
380	131
410	122

Table.1 measurement results of actual physical distance(cm) vs. disparity(pixel).

user's location with very low complexity computation and calibration free. According to the experimental result, the correlation coefficient of distance and the reciprocal of disparity is up to 0.9985.

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] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. "Real-time human pose recognition in parts from single depth images." IEEE CVPR, 2011

[3] Amit Bleiweiss* Dagan Eshar, Gershom Kutliroff, Alon Lerner, Yinon Oshrat, Yaron Yanai, "Enhanced Interactive Gaming by Blending Full-Body Tracking and Gesture Animation" 2010 *ACMSIGGRAPH ASIA Sketches*

[4] 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.

[5] 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.

[6] M. Piccardi, "Background subtraction techniques: a review," in Proc. IEEE Int. Conf. Systems, Man, Cybernetics, 2004, pp. 3099–3104.