A Real-Time 3D Interactive System with Stereo Camera in the Uncertain Background

Chun-Wei Yu, Che-Wei Chang, and Liang-Gee Chen
DSP/IC Design Lab, National Taiwan University, Taiwan
jamesbond@video.ee.ntu.edu.tw, r02943127@ntu.edu.tw, *lgchen@video.ee.ntu.edu.tw

***Abstract*—This paper proposes a real-time 3D interactive system for human to communicate with the display device by using the stereo camera. The system can automatically detect human actions through online learning process and overcome the uncertain background through proposed algorithms. Additionally, we generate the depth information from stereo camera to improve the accuracy of action detection and also use the tracking methods to achieve real-time application.**

I. Introduction

Based on the growing of the popularity and the stronger function of smart phones, the depth sensor with low power such as the stereo camera is more possible being equipped, for example, the HTC one M8 generation, which uses the stereo camera to generate 3D reconstruction. One of the most famous and robust interactive system between human and TVs is Xbox with Kinect, which is inconvenient for customers to take along because of its huge size. So, it is a trend that anyone can play games with TVs or any other displays by using their body with the portable sensor, stereo camera, in the smart phone. This paper proposes a system, which is aimed at building the bridge between computers and human with stereo camera.

There are three main challenges: one is that the detection of human face and fingertips in an uncertain luminance of the background may cause too many noises, we combine a sequence of filters to de-noise; another is that the real-time processing seems unreachable, the system adopt weighted function as ROI region to reduce the calculation; the last one is that there are countless human gestures, we use the combination of fundamental elements through online learning for users to define their own gesture, this interactive system is the first one to achieve real-time and has high accuracy.

The main advantages of our system are, first, we combine three filters to segment the body shape efficiently. Another one is that we combine hand and face detection with tracking technology to shorten the calculating time in order to achieve real-time performance. We also generate the depth information after detection to raise the accuracy for gesture recognition. Finally we use the concepts of exemplars-SVMs [1] for action definition. The exemplars representation combines the detail information to train and estimate the human action. It can reach high accuracy and train any actions users want.

II. Proposed System Flow

Our interactive system comprises three phases as Fig. 1. First, it begins with de-noise phase, in which it can segment ROI region without noises. Second, the detection and tracking procedures aim at pointing out the face and the fingertips area to do the action recognition. The last one is training and testing phase, in which users may train their own actions followed by the indicator on the screen, then use the classifier built in memory to determine the actions in real time.

A. De-noise Phase

In this section, we may introduce how we do de-noise to overcome the uncertain luminance in the background. There are three steps have been used, the first one is the skin color filtering, also the most important one of all. In the ideal case, the skin color is the only feature that we concern about because the interactive object is human. However, using only skin color filter may be out of our expectation in the reason that there may be other objects' color looks like skin one. Instead of using only skin color filter, we also adopt background subtraction [2] and edge detection algorithm.

The reason of using background subtraction is that the background is always fixed; only the human is moving and should be detected, so we don't need to take care about the unchangeable part. This method can highly rise the de-noise performance.

However, using this strategy may generate some errors because the computer may detect the human as the background based on the similar color, then the target object would be ignored as background. So, we combine the edge detector in the reason that we consider if the background is colorful, it may not be recognized as human, so just use the background subtraction is enough, but, if the color of the background is monotonous as human skin, we use the edge detection to segment human from background. These two methods occur oppositely, so combining them can almost ignore the noise than just using skin color filter.

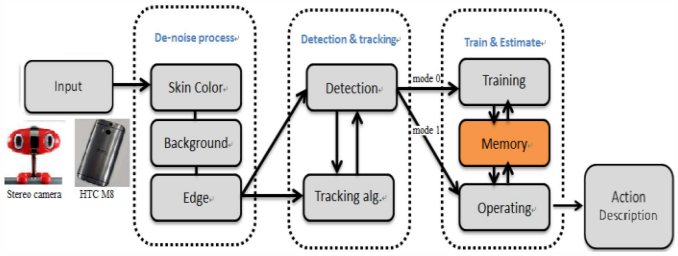


Fig. 1. Block diagram of proposed interactive system

B. Tracking by Detection Phase

In this part, we may describe how we detect the face and catch the fingertips which are the most basic elements for training and operating phases. Before the detection of target elements, we pre-filter the ROI region by using skin color

978-1-4799-7543-3/15/$31.00 ©2015 IEEE

filter, background subtraction [2], and the edge detection mentioned in part A to boost the procedure to realize real-time performance.

The detection of faces has been researched for a long time; we adopt the original method, which uses haar-like features and adaboost learning [3]. The detection of hands is more complicate due to its uncertain gesture and changeable environment. We use the convex hull and defected points, which is implemented in open source of opencv, to represent each finger tips and palm area. The experiment results are showed in Fig. 2.

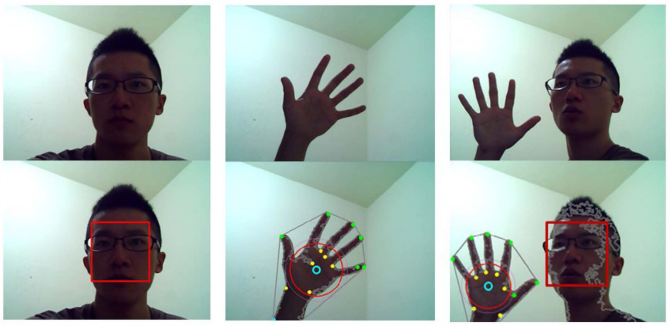
We also apply the tracking algorithm by weighting the whole image using several previous images with the decay in time. The ROI region is proportional to the weighting factors in each position in the each frame. Formula is showing below,

$$\mathrm{ROI}(i, j, t) = \sum_{k=1\sim30} W(i, j, t-k) * 0.5^{k}. \quad (1)$$

After finding the target in the current image, the weight function W is refined as following.

$$W(i, j, t) = \mathrm{ROI}(i, j, t) + \mathrm{Target}(i, j, t). \quad (2)$$

$$\mathrm{Target}(i,j,t) = \begin{cases} \text{true, if I}(i,j) \text{ is target.} \\ \text{false, otherwise.} \end{cases} \quad (3)$$



(a) (b) (c)

Fig. 2. Based on the uncertain luminance background and all of them are accelerated by the tracking algorithm. (a)The haar-like face detector with adaboost learning algorithm [5]. (b)Detection of hand, green and yellow points represent finger tips and convex hulls separately, the blue and red circle shows the center of hand and the palm region respectively. (c)Mixture of the face detection and hand detection algorithm with the tracking algorithm to achieve real-time video processing.

C. *Training and Estimating Phase*

In this section, we show the training and the estimation flow. Instead of directly training the action from features, we use the concept of exemplars-SVMs [3], which means that separating the action into several detail parts, and then the combination of the basic elements can decide the gesture. We track the path and record the depth of fingertips, the center of the hand and the face. We use the relationship of the recorded points as the decision factors.

Both of training and estimating procedures take 2 second video segments as representation of one action. In the training phase, the user may follow the director on the screen to train the actions they want, and in the estimating phase, the system may detect human action in the overlap period of 1 second, the flow is showed as Fig. 3.

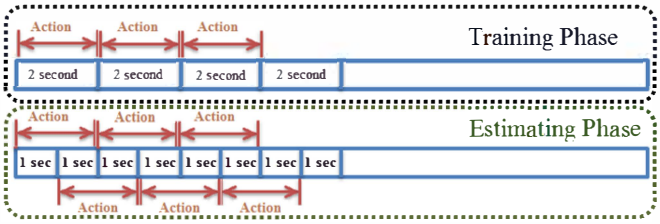


Fig. 3. Actions separation in the training and estimating phase.

III. Experiment Result

In the experiments, we define five actions as waving the hand from left to right with open hand (WLR), waving the hand from right to left with open hand(WRL), pulling with open hand straight forward (PSF), grabbing from open hand to close one straight backward (GBW) and rising the hand (RH).

We train each action 5 times and test each of them for 20 times. We record the video and set the ground truth by ourselves, then compare the ground truth with the estimation result. The result shows as Table. I. The left column means the ground truth and the top row is the detection results. For example, the 23 in WLR to WLR and 4 in WLR to PSF means there should be 27 actions of WLR and the our system correctly classifies 23 to WLR but makes 4 errors to PSF.

IV. Discussion and Conclusion

TABLE I
THE EXPERIMENT RESULTS WITH FIVE ACTIONS

	WLR	WRL	PSF	GBW	RH	Accuracy
WLR	23	0	4	0	0	85.18%
WRL	0	25	3	0	0	89.29%
PSF	3	2	20	0	0	80%
GBW	1	0	0	23	0	95.83%
RH	0	0	0	0	27	100%

The action is more than 20 times because the period of some actions last more than 1 second. The result is as good as we expected, but the WRL and WLR are sometime recognized as PSF or the inverse occurs because we segment each action in the fixed time, showed in Fig. 3. It might cause errors because each action takes different time.

There are two directions should be improved. First, there is a future work that we need to find a way that can separate each action more appropriately in time domain. The second one is that we need to find more comprehensive way to represent an action, make sure each action can be separated well although they looks similar.

V. References

[1] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of Exemplar-SVMs for Object Detection and Beyond," International Conference on Computer Vision, 2011.

[2] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," International Conference Pattern Recognition, UK, August, 2004.

[3] P. Viola, M. J. Jones, "Robust Real-Time Face Detection," International Journal of Computer Vision 57(2), 137-154, 2004.