9th IEEE/ACIS International Conference on Computer and Information Science

Movement Tracking in Real-time Hand Gesture Recognition

Hong-Min Zhu and Chi-Man Pun

Department of Computer and Information Science
University of Macau, Macau SAR, China
{ ma86560, cmpun}@umac.mo

***Abstract*—To translate the gesture performed by the user in a video sequence into meaningful symbols/commands, feature extraction is the first and most crucial step in such systems which measures the detected hand positions and its movement track. We propose an efficient approach based on inter-frame difference (IDF) to handle the hand movement tracking, which is shown to be more robust in the accuracy aspect compared to skin-color based approaches. Computational efficiency is another attractive property that our approach greatly improves the processing frame rate to fulfil the demand of a real-time hand gesture recognition system.**

Keywords- hand gesture, hand detection, movement tracking, inter-frame difference

I. INTRODUCTION

As an attractive alternative of traditional human-computer interaction interface (HCI) such as keyboards, mice, wands and joysticks, the use of human movements, involving arms, face, and/or body, and especially hand gestures, has retrieved more and more attention in recent years according to their naturalness and convenience in practice. This serves as a motivating force for research in modeling, analyzing and recognition of hand gestures.

Recognizing gestures is a complex task which involves many aspects such as motion modeling, motion analysis, pattern recognition and machine learning [1, 2]. Meaningful hand gestures can be classified as static hand postures and temporal gestures (motion patterns). Hand postures appear statically as a combination of different finger states, orientations, and angles of finger joint [3]. Hand modeling is the main and also the most difficult task in recognizing postures, as fingers are usually articulated, self-occlusion makes the detection and tracking local hand configuration challenging. Temporal gestures, on the other hand, represent the meaning of gesture by hand movement rather than the details of finger appearance. In computer vision (CV) based solutions, hand gestures are captured by web cameras which used as input. Due to the resolution of the videos, only a general sense of the figure state can be detected, we focus on the gesture represented by hand movement in this paper. Procedures in a general framework of gesture recognition system are show in Fig.1.

While leave the learning of gesture models and the recognition of extracted hand movement track as our future work, this paper gives a solution to the first stage of the system: hand detection and movement tracking. The recognition rate of the system will highly depend on the accuracy of this feature extraction stage, and the processing frame rate of the system is also dominated by the efficiency of hand detection.

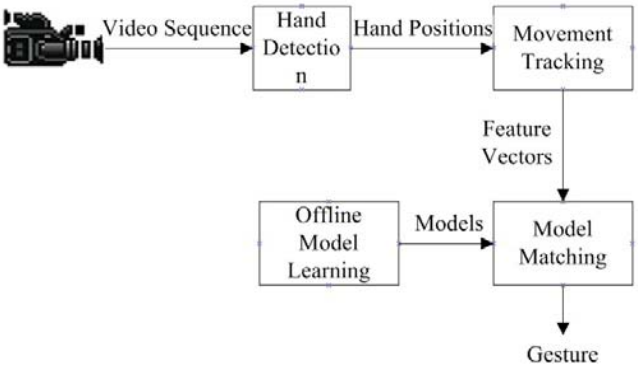

Figure 1 framework of hand gesture recognition system

The rest of the paper is organized as follows. In Section II, some of related works are presented, as well as existing problems and challenges. In Section III, the system context and video data set are defined. The details of our proposed hand detection and movement tracking algorithms are given in section IV. Experimental results are presented in Section V. And we finally conclude the paper Section VI.

II. RELATED WORKS

Temporal hand gestures are performed by hand movement which is described by a serial of hand position sequence. While the detection of gesturing hand in a given video sequence is the first step of movement tracking, there are some challenges with respect to the environments or users, and constrains/assumptions may necessarily need to be defined in specific applications:

1. The user is usually required to gesture in front of the camera in a reasonable distance.
2. The lighting condition should be uniform in the background.

978-0-7695-4147-1/10 $26.00 © 2010 IEEE
DOI 10.1109/ICIS.2010.13

3. The background should not be too complex, and other movable distracters may not be allowed.
4. The gesture should be performed in a reasonable speed compared to the frame rate of camera.

There are two main approaches for detecting the hand in a video frame. The first one is skin color based hand detection (SCHD). The boundaries of skin clusters can be trained from a set of sample frames/images to form the skin color model. J. Alon, V. Athitsos [4] use a generic skin color histogram to compute the skin likelihood image. The statistical skin color model [5] is learned by manually label pixels in the training set as skin/non-skin pixels, this is a trivial work and the model may not be robust on a different test data set.

J. Kovac and P. Peer [6] build a skin classifier by explicitly define the boundaries of skin cluster in *RGB* color space as a set of rules:

P(R, G, B) is classified as a skin pixel if:
$R > 95$ and $G > 40$ and $B > 20$ and
$max\{R, G, B\}-min\{R, G, B\} > 15$ and
$|R-G| > 15$ and $R > G$ and $R > B$

As will be demonstrated in section V, there are some drawbacks of SCHD solutions:

- It's computationally expensive as it's a pixel-wise classification approach.
- Skin-liked objects may also be classified as part of hand.
- Lighting reflection may fail the skin detection. A few works considered this problem such as [7].
- The range of skin color varies among different human species.

The second approach is background subtraction based hand detection (BSHD). When the camera is fixed and a video frame contains only pure background is available, foreground objects in other frames can be extracted by subtracting them with the background frame [8, 9]. BSHD overcomes SCHD both on performance and efficiency aspects since it takes correlation between frames into consideration, objects are detected by the measurement of inter-frame difference.

III. System Domain

Our proposed movement tracking approach is motivated by J. Alon, V. Athitsos's work [4], they developed a robust framework for spatiotemporal gesture recognition which works well when the gestures are performed in front of moving, cluttered backgrounds. Three main components are involved: a spatiotemporal matching algorithm that can accommodate multiple candidate hand detections in every frame, a classifier-based pruning framework that enables accurate and early rejection of poor matches to gesture models, and a sub-gesture reasoning algorithm that learns which gesture models can falsely match parts of other longer gestures. The system is tested in two settings: a gesture recognition setting, where users perform gestures corresponding to the 10 digits, and an ASL sign retrieval setting.

Our solution focus on the hand detection and movement tracking problem, and the easy test data set of 30 video sequences from their approach will be used to test the reliability of our solution. Users wear shot sleeves and each of 10 users gesturing 3 video sequences. In each sequence, the user signed once each of the 10 Palm's Graffiti digits (Fig. 2), the start location of each digit is indicated by the white point and each one is signed by one stroke (see "4" and "5").

For each sequence, there is a corresponding ground truth data file which indicates the start and end frame of each gestured digit. Currently in movement tracking stage, we assume the gesture segmentation is known and make use of this ground truth data to clip a video segment from the original sequence, only one gestured digit will be presented in the segment. This segment serves as the input of our hand detection and movement tracking procedure.

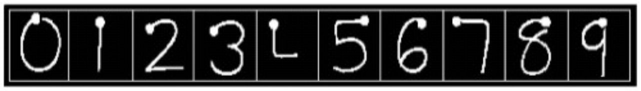

Figure 2. Palm's Graffiti Digits.

IV. Proposed Solution

Our proposed solution for gesture tracking involves two steps as show in Fig. 3, where hand detection and movement tracking are separated by the point line.

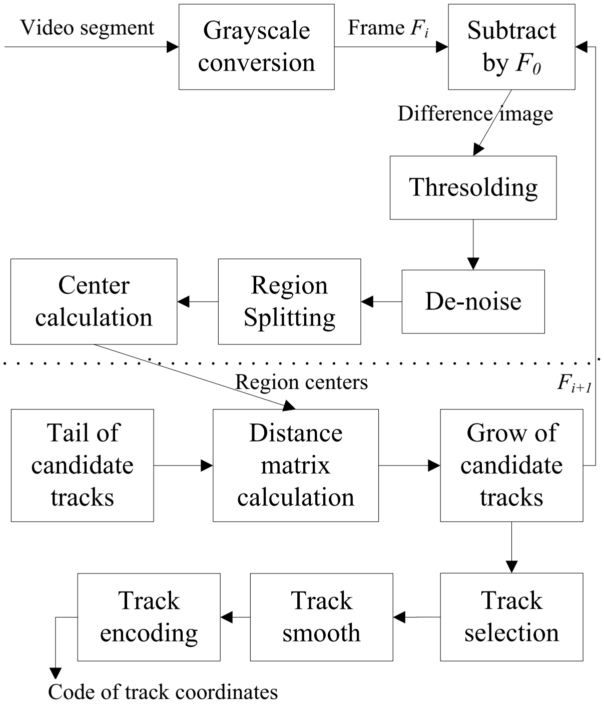

Figure 3. Workflow of gesture tracking

A. *Inter-frame difference based hand detection (IDFHD)*

Our hand detection algorithm is motivated from BSHD according to its robustness and reliability. Rather than a pure background frame is available in BSHD, the user present in all frames of a video segment. Take the first frame from segment as a "background" and subtract it from all subsequent frames, only moveable objects will be detected instead of all foregrounds. Since the gesture is expressed by hand movement track, the gesturing hand is the most active object to be detected. The algorithm of our proposed IFDHD is summarized as follows:

Algorithm I – hand detection

```
Input: frames $F_{i=1...N}$ of a video segment
Step 1: convert frame $F_1$ into grayscale.
For each frame $F_i$ ($i>1$), do
Step 2: convert frame $F_i$ into grayscale.
        Intensity difference image $D_0 = | F_i - F_1|$.
        Binary image $I = (D_0 > T_0)$.
Step 3: do image opening operation on $I$ followed by image closing.
Step 4: large regions are splitted into regions with max size of boundary box as 60x80.
Step 5: for each region, calculate the center coordinate.
Output: center coordinates of each region in each frame.
```

The IDF is calculated in step 2 by subtraction between two frames, we take the fact that the gesturing hand is the most active object in the scene compared to the body and face. In case of body/face moves shortly, most pixels in those areas don't have a significant change since the skin/cloth has a uniform color distribution. After thresholding the difference image by a constant T_0, the resulted binary image will at most reserves the boundary of the body. On the other hand, the whole hand can be detected once it moves to a new position. In step 3, image opening removes the isolated foreground pixels, while image closing connect nearby skin regions which are separated by isolated points/lines to form larger regions. As the detected arm and hand may be in one connected region, step 4 split such large regions into smaller ones so that the calculation of region center coordinates in step 5 will be more precise, which will be used to form the track of a gesture. The output of hand detection procedure maintains multiple center coordinates in each frame.

B. *Movement tracking*

After we get the output features from hand detection procedure as a serial of centers coordinates of detected regions in each frame, movement tracking algorithm follows to construct multiple gesture tracks as candidates and finally select one that best describes the hand movement. While multiple hand regions can be detected in each frame, we consider two regions in successive frames with shortest path of their center points belong to the same gesture candidate.

There are self-occlusions of 2D projected hand during gesturing, and the number of detected hand regions in each frame may be different. This problem challenges movement tracking that: 1) the gesture being tracked may be cut because of hand region lose. 2) Invisible region belongs to the tracked gesture may appear again after a while. 3) Regions in the tracked gesture may be not extracted from the gestured hand.

According to these uncertainties, we keep track of multiple gesture candidates, and tracks of gestures are updated based on the distance between region centers and the tail of each tracks, instead of regions detected in previous frame.

We denote *C (I, M)* as *M* detected regions centers in frame *I*, and *G (N)* as *N* gesture candidates being tracked. The movement tracking algorithm is given as:

Algorithm II – movement tracking

```
Input: region centers detected in each frame.
Step 1: initialize the start of $N=M$ gestures $G (N)$: $G (n) =C (I=1, m)$.
Step 2: For each frame $I>1$, do
    Step 2.1: construct $L (N)$ which are tail locations of each gesture $G (n)$.
    Calculate matrix $D(N, M)$: distances between $C(I, M)$ and $L(N)$.
    Step 2.2: repeatedly select $D(n, m)=min(D(N, M))$.
    If $D(n, m)< T_0$: append $C(I, m)$ to $G (n)$ and delete $D(n, m$.
    Else: initialize $G (N+1)$ with start location of $C(I, m)$.
Step 3: select $G (N_0)$ that has the maximal standard deviation.
Step 4: smooth movement track $G (N_0)$ and interpolate it to $N_g$ coordinates.
Output: encoding of movement track.
```

As described in step 2, each detected region center in current frame will either belongs to a gesture being tracked or starts as a new gesture candidate depend on the distance measurement. The problem that a continuous gesture cut by occluded region is handled in this algorithm, as long as the current region doesn't appear far away, the gesture can also be reconnected. When all track candidates are formed, one possible solution to select the best track is based on the path lengths of candidates since movement performed by hand has the longest path. However, this idea is not robust when an object keep moving in a small area that the candidate can also has a long path. In our solution we use instead the standard deviation of the center coordinates along the gesture path, which is shown to be more reliable.

To learn the gesture models from a set of movement tracks, statistical tools such as hidden markov model [10] require the training data to be of uniform feature size. We apply B-form spline approximation to smooth the movement track and interpolate it to a certain number of points in step 4. The output of movement tracking algorithm is the encoding of motion vector, we classify the movement into eight orientations.

V. Experimental results

To demonstrate the reliability and efficiency of our proposed solution in this section, we firstly compare the accuracy of SCHD and IDFHD, and follow the result of movement tracking based on these two hand detection approaches respectively.

A. *Experiment on hand detection*

Fig.4 shows the result of SCHD in a video frame. We can see from Fig.4(b) that many objects in the background are also classified as skins, such as the door and the pillar behind the person. Another problem is skin pixels may be rendered nearly white under the lighting reflection effect as shown in Fig.5, fingers can't be detected by skin classification rules.

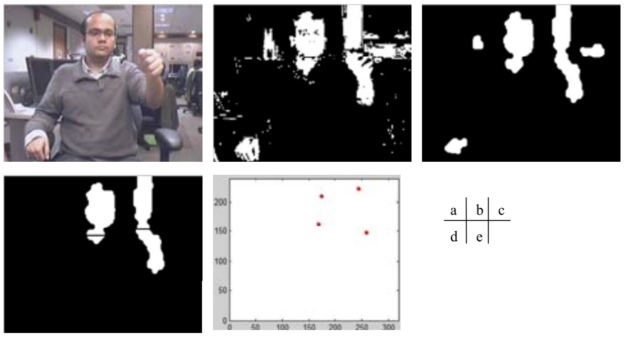

Figure 4. Result of SCHD: a) original frame. b) Skin pixel classification. c) De-noise and region connection. d) Region splitting. e) Centers of each region

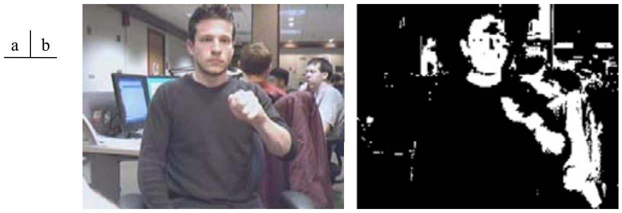

Figure 5. Effect of lighting condition: a) original frame. b) Skin pixel classification

SCHD works on each frame independently, which didn't take the correlation among frames into consideration. Video streams are constructed by continuous frames from which the hand detection should benefit to greatly improve the efficiency.

The result of IDFHD is given in Fig.6. We consider the first frame (Fig.6(a)) from the video segment as the "background" frame, and the 11th frame is subtracted by the "background" frame to get the pixel intensity difference image (Fig.6(c)). After thresholding (Fig.6(d)), the binary image only reserves the boundary of the body. After de-noise and region connection process, only two regions are kept in Fig.6(e).

Compared to SCHD, those static skin-colored objects no long be detected in this approach, and the lighting reflection effect can not lead the hand detection to be failure any more.

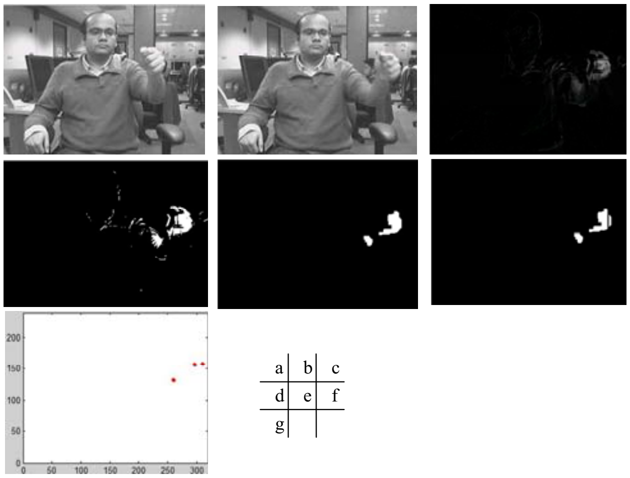

Figure 6. Result of IFDHD: a) 1st frame. b) 11th frame. c) subtraction of (a) and (b). d)threshold of (c). e) de-noise. f) region splitting. g) centers of regions.

B. *Experiment on movement tracking*

Movement tracking based on SCHD and our proposed IDFHD are tested on some selected digits which are performed by different users, the resulting tracks of hand movement are shown in Fig.7 and Fig.8 respectively. Detected cumulative region centers are given in the last frame of each video segment (see first row in Fig.7 and Fig.8).

We can see from Fig.7 that the first three digits (2, 5 and 8) can be tracked correctly, while the fourth digit, which is also a digit “2” but is gestured by a different user, can not be detected successfully. This is caused by the lighting reflection in SCHD as we have discussed. While tested on the same video segments, our IDFHD based movement tracking improves the performance as shown in Fig.8. Although the first three digits tracks are less precise than Fig.7, they are still recognizable after the track smoothing. And the last digit can be well tracked.

Another attractive property of IDFHD based movement tracking is efficiency, which is the main concern of real-time hand gesture recognition system. Table 1 summarizes the time cost of two corresponding experiments. We can see that SCHD based movement tracking has an average processing speed of 2.6fps, while our proposed solution greatly improves the efficiency by around 60fps, which is more than 20 times faster.

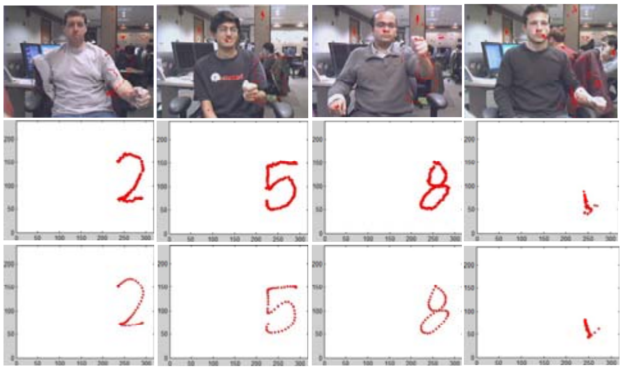

Figure 7. SCHD based movement tracking. First row: last frame of video segment. Second row: detected digit track. Third row: smoothed track.

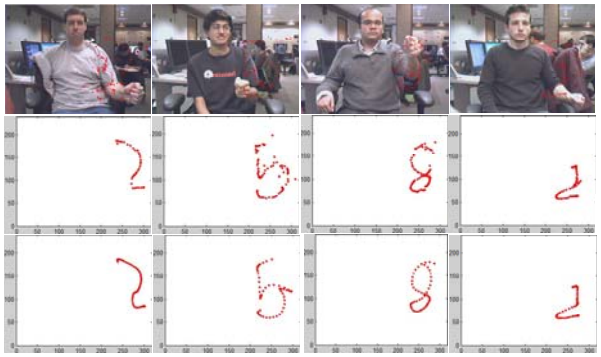

Figure 8. IFDHD based movement tracking. First row: last frame of video segment. Second row: detected digit track. Third row: smoothed track.

Gestured digit		2	5	8	2
	frames	71	81	94	68
SCHD	time	26.43s	29.21s	37.07s	24.92s
	fps	2.69	2.77	2.54	2.73
IDFHD	time	1.28s	1.35s	1.52s	1.08s
	fps	55.47	60.0	61.84	62.96

Table 1. Efficiency of SCHD and IDFHD

VI. Conclusion

We proposed an inter-frame difference based hand movement tracking solution in this paper, which serves as feature extraction stage in hand gesture recognition system. Experiments demonstrate that the performance of our solution overcomes the skin color based movement tracking, specifically on efficiency aspect.

In our future work, we are going to learn the gesture models for each of ten digits, and design the model matching procedure to recognize the extracted movement tracks as meaningful gestures.

REFERENCES

[1] A. Erol, G. Bebis, M. Nicolescu *et al.*, "Vision-based hand pose estimation: A review," *Comput. Vis. Image Underst.,* vol. 108, no. 1-2, pp. 52-73, 2007.

[2] S. Mitra, and T. Acharya, "Gesture Recognition: A Survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* vol. 37, no. 3, pp. 311-324, 2007.

[3] Y. Wu, and T. S. Huang, "Hand modeling, analysis and recognition," ***IEEE SIGNAL PROCESSING MAGAZINE,*** vol. 18, no. 3, pp. 51-60, 2001.

[4] J. Alon, V. Athitsos, and Q. Yuan, "A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,* vol. 31, no. 9, pp. 1685-1699, 2007.

[5] M. Jones, and J. Rehg, "Statistical Color Models with Application to Skin Detection," *Int'l J. Computer Vision,* vol. 46, no. 1, pp. 81-96, Jan. 2002.

[6] J. Kovac, P. Peer, and F. Solina, "Human Skin Colour Clustering for Face Detection," *The IEEE Region 8 Computer as a tool EUROCON 2003,* vol. 2, pp. 144-148, Sept. 2003.

[7] M. Soriano, B. Martinkauppi, and S. Huovinen, "Skin detection in video under changing illumination conditions," *Pattern Recognition, 2000. Proceedings. 15th International Conference on,* vol. 1, pp. 839-842, Sept. 2000.

[8] S. Jabri, Z. Duric, H. Wechsler *et al.*, "Detection and Location of People in Video Images Using Adaptive Fusion of Color and Edge Information," *15th International Conference on Pattern Recognition (ICPR'00)* vol. 4, 2000.

[9] T. Horprasert, D. Harwood, and L. S. Davis, "A Robust Background Subtraction and Shadow Detection," *Proc. Asian Conf. on Comp. Vision*, January 2000.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*: John Wiley & Sons, 2001.