Real-time Deformable Hand Tracking System Based on Monocular Camera

Chen-Lan Yen, Wen-Hung Ting and Chia-Chang Li
Industrial Technology Research Institute of Taiwan
{chester, wenhungting, lijc}@itri.org.tw Tel: +886-6-6939000

***Abstract*— In this paper, a real-time markless hand tracking system is proposed based on a monocular camera. The proposed method captures the deformable hand posture and filters background noise by analyzing the motion and skin features obtained from video sequences. In the system, users are allowed to control the mouse cursor in the application by moving their hands without retrieving hand gestures from a training database. Experiments show that the proposed system could robustly extract the hand trajectories with low computational complexity to achieve real-time interaction applications.**

I. Introduction

Hand gesture is one of the common modalities for human communication in our daily lives. Hand movements and gesticulations appear to represent most aspects of the expression, such as visual, temporo-spatial and emotion etc. [1]. Recently in the HCI area, efforts in novel research of natural, convenient and efficient user interfaces [2] which capture hand gestures and poses are gradually taking place.

Due to the variant and complex appearance of a moving hand, it is difficult to track the hand posture especially in a random selected environment. Electronic data gloves is one of the common solution for capturing hand's movement and expression, but at the same time, it typically constrained users with much uncomfortable restriction. On the other hand, non-contact sensors such as cameras offer users more freedom of expressing themselves without physical restriction; however, more conditions of environment have to be set properly for a real-world application comparing to the glove-based method.

Generally, camera-based hand-tracking research could be categorized into two groups. Model-based methods use fixed 3D hand models to deal with the changing appearance of continuous hand motion [3] and track hand gestures with numerous features in high dimension, therefore its expensive computation always leads to a lower speed and efficiency. Alternatively, the second method is appearance-based tacking. Some researches of this track show good results of real-time tracking with presumed constant hand postures [4] and some researches apply fast and efficient algorithms such as KLT optical flow, SIFT [5] et cetera. Nevertheless, these methods usually lose their tracking target when the featured contour suddenly flip or change its shape abruptly. Moreover, to select enough quantities of features from a moving hand in color or monochrome images is difficult. Camshift [6] algorithm is able to tacking the skin feature of dynamic hand motion. However, when the background objects have the same appearance of hands, the algorithm is likely to fail because of the ambiguous color features. Moreover, most of existing methods depend on manual marking initial feature points or area and are not able to resume automatically once the tracking fails or otherwise it relies on a large database with a huge amount of trained data.

In this paper, we developed a real-time deformable hand tracking algorithm MSED (Motion-Skin Extended Determination) based on monocular view. The visual system is created under the view of bionics on the features from skin and motion contours extracted from foreground and background objects. The approach has been successfully implemented on a front-view human-computer interactive system which efficiently tracks user's hand trajectories for smoothly simulating desired mouse cursor movement in a common OS. Users are able to interact with the system with their hands naturally and directly without an extra device or previous training data.

The paper is organized as follows: The next chapter introduces our approach by two steps: 1. how to segment the hand region; 2. how to determine the target object by MSED algorithm. The third chapter is our experimental results from several videos including a list comparing our method to other existing algorithms. The last chapter gives the conclusion and briefly introduces our future work.

II. Materials and Methods

The proposed system could be divided into two modules: Skin-Motion Acquisition and Hand Tracking. The block diagram of procedure is shown in Fig. 1.

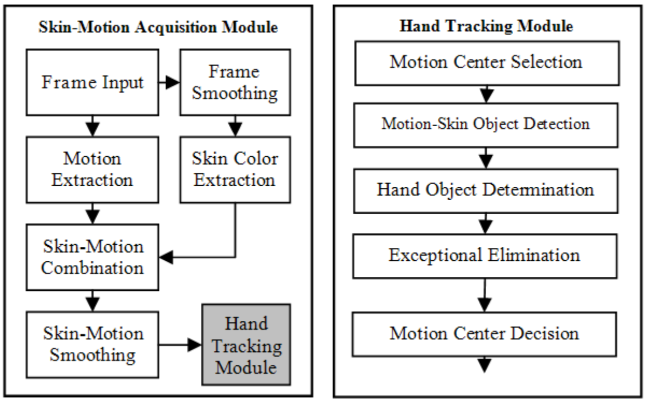


Fig. 1 The block diagram of hand tracking

09-0106520655©2009 APSIPA. All rights reserved.

A. Skin and Motion Objects Acquisition

Initially, the original color image is converted to gray image F , and the motion feature is detected by subtracting last two consecutive images in time sequence. The motion image M , as shown in Fig. 2b, can be obtained as (1):

$$M(k) = T(|F(k) - F(k-1)|) \quad (1)$$

Where k denotes the frame number, $|\cdot|$ indicates the absolute value, and $T(\cdot)$ is the threshold function. In this system the threshold value we use is 6.

In order to unite the discrete skin regions to an identified unit, the raw color image is smoothed before applying skin detection. A mask S is defined and is centered on (x, y). The average of pixel value is computed in the mask as (2).

$$\bar{e}(x, y) = \frac{1}{p} \sum_{(x,y)\in S_{xy}} e(x, y) \quad (2)$$

Where p is the pixel number in the mask S , and S_{xy} denotes the set of pixels in the mask. The smoothed image I can be obtained by using (2) to smooth every pixel in original image.

$$I = \{c \mid c \in \bar{e}(i, j)\},\ 0 < i < w, 0 < j < h \quad (3)$$

Where w, h is the size of the raw image. Then the image V , shown in Fig. 2(c), is derived of skin objects from image I according to the characteristic of skin color in YCbCr color space[7]. Fig. 2(d) is the skin-motion image N from intersecting motion image M and the skin objects image V ,

$$N = M \cap V \quad (4)$$

After all, N is smoothed again by (3) to filter the discrete region and is reset to the final binary result as shown in Fig. 2(f).

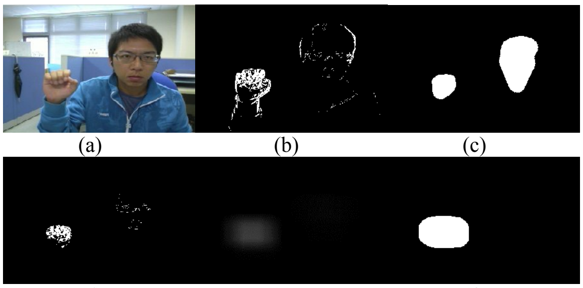


Fig. 2 (a) original image, (b) motion image, (c) skin image, (d) skin-motion image, (e) smoothing of (d), (f) final result

B. Hand Tracking

To determine the exact location of the target hand using previous result is the subject in this section. The center of hand has now been located without any comparison of hand features or templates, therefore, the significant points of hand, for examples, forefinger, thumb, and etc., can not be recognized accurately by previous methods. However, the result is sufficient qualified for our goal: to locate target hand posture.

To track hand using the combination of only skin and motion objects may encounter several difficulties. First of all, not all the motion of skin contours are from the same target object. Secondly, skin objects may superimpose each other in many kind of cases, and so do motion objects. To overcome these difficulties, the following chapter is divided into four stages: 1. Motion Center Selection, 2. Motion-Skin Object Detection, 3. Hand Object Determination, and 4. Exceptional Elimination.

B-1. Motion Center Selection

A skin-motion object rarely exists uniquely in a real-work application. To determine the motion center of current moving hand, minimum distance between possible target objects in current frame and the previous target center in the last frame is calculated by the equation:

$$\vec{M}' = \arg\min_{\vec{P}_m^i}\left(d_i = \left\|\vec{P}_m^i - \vec{M}\right\|_2 \mid i \in C\right)$$
$$\left\|\vec{P}_m^i - \vec{M}\right\|_2 = \sqrt{\left(p_x^i - m_x\right)^2 + \left(p_y^i - m_y\right)^2} \quad (5)$$

Where C denotes the numbers of motion objects and $\vec{P}_m^i$ indicates the mass center of the i^{th} motion object in current frame. d_i denotes the 2-Norm between the previous target motion center $\vec{M}$ and the i^{th} candidate of the current motion center. After computing (5), the motion center $\vec{M}'$ in the current frame is determined.

B-2. Motion-Skin Object Detection

A hand object's existence or superimposition is poorly detected by only the numbers of skin objects in a frame. For instance, when the hand moves near frame boundaries or the hand superimposes a face contour, or the variation of illumination between frames, the hand detection by object numbers could easily fail. From that fact, the status of the hand motion must be considered as a key factor of the detection.

Initially, the hand status is assumed to be in motion. The moving skin object in current frame is extracted by the motion center $\vec{M}'$ as shown in (6) and then determine the current moving skin object $\vec{T}_m'$:

$$\vec{T}_m' = \arg\min_{\vec{P}_s^i}\left(d_i = \left\|\vec{P}_s^i - \vec{M}'\right\|_2 \mid i \in K\right) \quad (6)$$

Where , K denotes the numbers of skin objects and $\vec{P}_s^i$ indicates the mass center of the i^{th} skin object in current frame. This decision process is named MSD, Motion-Skin Determination, for contracted form.

B-3. Hand Object Determination

The skin object $\vec{T}_m'$ obtained from the motion center $\vec{M}'$ doesn't guarantee to successfully track the hand with less momentum. It is observed that the disappeared mass center of the hand may be find again via the past time sequence. The previous mass center of hand was defined as (7)

$$\vec{T}_s' = \arg\min_{\vec{P}_s^i}\left(d_i = \left\|\vec{P}_s^i - \vec{T}_s\right\|_2 \mid i \in K\right) \quad (7)$$

Where, $\vec{T}_s$ is the previous identified mass center of the hand, $\vec{T}_s'$ is the current mass center of the skin object closest to the motion center.

The identified mass center of hand $\vec{T}_s'$ tends to stop; on the contrary, the current moving skin object $\vec{T}_m'$ tends to keep moving. The position of the target hand can be determined by considering the discrepancy of $\vec{T}_s'$ and $\vec{T}_m'$ in 3 different cases:

(a). The hand is moving without overlapping other static skin objects. At the time, $\vec{T}_m'$ presents the same position as $\vec{T}_s'$ and the target hand object is then identified.

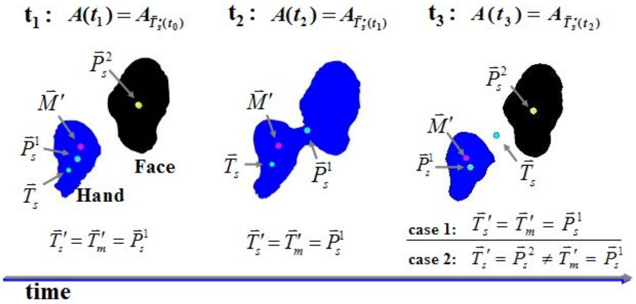
(b). The non-motion target hand is individual to other moving skin objects. In this case, $\vec{M}'$ is replaced by $\vec{T}_s'$ while $\vec{T}_m' \neq \vec{T}_s'$. Hence the current target stays in the static hand objects without considering other motion objects.

(c). The hand and the other skin objects are moving simultaneously but are not overlapping to each other. This is an extension of case (a) and (b).

After the position is determined, $\vec{T}_s$ is updated to $\vec{T}_s'$. In those cases that a frame contains only non-overlapping contours, the hand object can be tracked successfully with MSD method; however, in other cases, the MSD method might fail. In practice, superimposing one skin contour into another usually causes misidentification among common hand tracking algorithms. Therefore, discovering the common parameters are necessary for successfully tracking the target hand object to distinguish different motion types from complex environments. As a result, an expansive determination condition is acceded to MSD and the details are described in the following section.

B-4. Exceptional Elimination

Fig.3 is a time sequence which shows the possible results by applying only the MSD algorithm when a hand contour intersects a non-motion face object. $A(t)$ denotes the area of the target skin object in time t and the rest parameters shown in Fig 3 are defined in the following section. The presented results are divided into 3 phases t1, t2 and t3 for the exposition of the exceptional elimination method:



t_1: $A(t_1) = A_{\vec{T}_s'(t_0)}$
t_2: $A(t_2) = A_{\vec{T}_s'(t_1)}$
t_3: $A(t_3) = A_{\vec{T}_s'(t_2)}$
$\vec{P}_s^2$
$\vec{M}'$
$\vec{P}_s^1$
$\vec{T}_s$
Face
Hand
$\vec{T}_s' = \vec{T}_m' = \vec{P}_s^1$
$\vec{M}'$
$\vec{T}_s$
$\vec{P}_s^1$
$\vec{T}_s' = \vec{T}_m' = \vec{P}_s^1$
$\vec{P}_s^2$
$\vec{M}'$
$\vec{P}_s^1$
$\vec{T}_s$
case 1: $\vec{T}_s' = \vec{T}_m' = \vec{P}_s^1$
case 2: $\vec{T}_s' = \vec{P}_s^2 \neq \vec{T}_m' = \vec{P}_s^1$
time

Fig. 3. A time sequence of two intersecting objects: the hand and the face contours. Two objects are moving closer at t1; uniting at t2; and separating at t3.

(a). Phase (t_1) - Approach

$\vec{P}_s^1$ and $\vec{P}_s^2$ are the mass centers of the hand and the face contours in the time sequence in Fig. 3. $\vec{P}_s^1$ is moving toward $\vec{P}_s^2$ at t1. After goes through the processes from section B-1 to B-3, the target hand object is determined by

$\vec{T}_s'(t_1) = \vec{T}_m'(t_1) = \vec{P}_s^1(t_1)$.

(b). Phase (t_2) - Union

In this state, $\vec{T}_s'(t_2)$ and $\vec{T}_m'(t_2)$ refer to the union object $\vec{P}_s^1(t_2)$ which is consisted with phase(t_1) since there exists only one object at t2.

(c). Phase (t_3) – Separation

Due to the variation of camera frame rate and speed of the target object, applying MSD method produces two possible results at t_3 according to the different location of $\vec{T}_s$. In the first case, after two contours separating from each other, the previous skin center $\vec{T}_s$ locates closer to the hand contour center $\vec{P}_s^1$. At the time, the motion skin center $\vec{T}_s' = \vec{T}_m'$. Consequently, the hand object is successfully identified as the target object. However, when $\vec{T}_s$ is closer to $\vec{P}_s^2$, $\vec{T}_s'$ and $\vec{T}_m'$ will locates in the head and hand contour individually. Thus $\vec{T}_s' \neq \vec{T}_m'$ and the target object is mismatched to the head object. This unexpected result happens especially when the hand moves relatively faster to the camera frame rate and the face contour is larger then the hand.

As the result, equation (8) is formed for guarantying that the object is tracked consistently.

$$f = \frac{\phi}{\sqrt{A(t)}} = \frac{\left\|\vec{T}_s'(t) - \vec{T}_m'(t)\right\|_2}{\sqrt{A_{\vec{T}_s'(t-1)}}} \tag{8}$$

Where A(t) is the area of the previous located target contour. The value f depends on the distance between the current motion target and the skin target contours by considering their contour size. This consideration is generated because $\vec{T}_s$ is tend to be located closer to the bigger contour at the time two objects are severing. When the value f is relatively larger then a threshold, the next target is determined by considering only $\vec{T}_m'$. This method is named MSED, Motion-Skin Extended Determination.

III. Results and Discussions

The environment of experiments is a platform that consists of Intel Core Duo CPU 3.0GHz PC, 2G RAM, and a webcam with frame size of 320x240 and frame rate of 30 fps. We use two videos to verify our proposed system, and the ground truth is established based on human evaluation. Video1 consists of 177 frames to test the hand moving across face; Video2 consists of 211 frames to test the objects moving in the background.

Fig. 6 shows the experimental results of the pure motion center (Motion), the motion center of MSD (MSD-M), the skin center of MSD (MSD-S), MSED and ground truth for video 1. X axis indicates time and Y indicates the distance between each center and the top left corner of the image. Using the pure motion center causes the position varying violently due to the hand and the face both move. The trajectory of MSD-S is more stable but it effects discontinuous position when the hand is moving across the face (49th and 60th frame). The trajectories of MSD-M and MSED are both visually similar to the ground truth. Table 1 and Table 2 show the correlations and RMSE (Root Mean Square Error) between each trajectory and the ground truth. MSD-S has a higher correlation but a higher RMSE since discontinuous position. Both of MSD-M and MSED have higher correlations and lower RMSE.

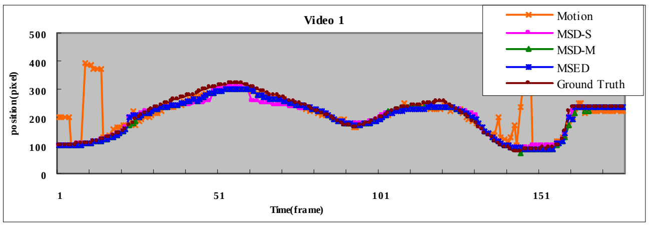


Fig. 6 The results of each steps in our method for video 1.

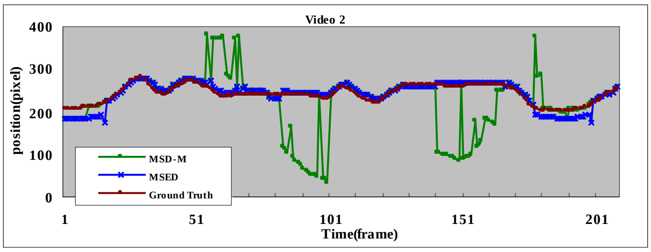


Fig. 7 The results of each steps in our method for video 2.

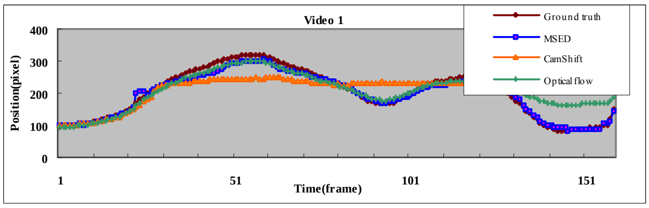


Fig. 8 The results of MSED, camshift and optical flow

The experimental results of MSD-M, MSED and ground truth for video 2 are shown as Fig. 7. Comparing with MSED, it is observed that the trajectories of MSD-M are tending to lose its association with the original target when the objects in the background move. In the meantime, MSED remained stable and has highly related to the ground truth.

Fig. 8 compares the results among MSED and other two tracking methods, camshift and optical flow. Applying camshift can neither track successfully when the hand overlaps the face (33rd frame), nor fix the error when the hand and the face separate. Optical flow shows the result by tracking ten feature points. It achieves the desired result at first, but the feature points are lost when the hand is covered.

MSD-M and camshift can be used for tracking only one skin region. Optical flow overcomes the overlapping of skin regions, but the feature points will be lost when the target is covered. MSD-S seems to achieve satisfied result but it causes discontinuous points when the hand is moving across the face. The experimental results showed MSED is suitable most moving mode and environment. The computing time of MSED is about 16ms. The CPU and RAM usage is less than 10% and 30MB respectively.

Table I The correlations between each trajectory and ground truth

Corr	Motion	MSD-S	MSD-M	MSED
Video 1	0.54085	0.98398	0.99252	0.99255
Video 2	-	-	0.11294	0.93730

Table II The RMSE between each trajectory and ground truth

RMSE (pixel)	Motion	MSD-S	MSD-M	MSED
Video 1	65.9005	16.8239	11.6719	11.6049
Video 2	-	-	74.6390	22.8404

Table III Compare between our method and other tracking algorithm

	CamShift	KLT	MSD-S	MSD-M	MSED
One skin region	v	v	v	v	v
Complex Background		v	v		v
Target deforming	v		v	v	v
Stable and continuous					v

IV. Conclusions and Future work

In this paper, we proposed a markless real-time hand tracking method based on detecting motion and skin features from monocular camera and is capable to work on an low-cost hardware platform without interference from deforming hand shapes or moving background objects. In the accomplished front-view system, users can interact with the computer by directly controlling mouse cursor via their desired hand trajectories. Future work will be analyzing the hand gestures in the extracted region and computing 3D information from a stereo view for a better human-computer interaction scenario.

References

[1] Axel G.E. Mulder (1996). Handgestures for HCI. Technical Report, NSERC Hand Centered Studies of Human Movement project. Burnaby, BC, Canada: Simon Fraser University.

[2] R. J. K. Jacob, “User interfaces,” in Encyclopedia of Computer Science, 4th ed, A. Ralston, E. D. Reilly, and D. Hemmendinger, Eds: Grove Dictionaries Inc., 2000.

[3] Matthieu Bray, Esther Koller-Meier and Luc Van Gool. Smart particle filtering for 3D hand tracking. Proc. Of 6th IEEE Conference on Automatic Face and Gesture Recogntion. 2004. pp. 675~680.

[4] Qing Chen , N.D. Georganas, E.M Petriu, “Real-time Vision based Hand Gesture Recognition Using Haar-like features”, IEEE Transactions on Instrumentation and Measurement -2007

[5] David G. Lowe, "Object recognition from local scale-invariant features," International Conference on Computer Vision, Corfu, Greece (September 1999), pp. 1150-1157.

[6] Mathias K · · olsch andMatthew Turk. Fast 2d hand tracking with flocks of features and multi-cue integration. In IEEE Workshop on Real-Time Vision for Human-Computer Interaction, Washington, DC, 2004.

[7] C. Garcia, and G.Tziritas. “ Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis.” in IEEE Transactions on Multimedia Vol. 1 , No. 3 , pp. 264-277, 1999.