Precise Searching and Segment of the Hand and Fingers on Gesture Recognition

Zhou Hang, Ruan Qiuqi

Institute of Information Science, Beijing Jiaotong University, Beijing, P.R.China

Abstract

In this paper, hand gesture recognition and tracking can be separated into two stages: the location of the hand and the tracking of their fingers. In the first stage background is checked and the hand area is found. In addition, the center of palm can be extracted as a preparatory step for tracking. Then a blob-ridge model is used to represent the hand that consists of five base states for feature extraction. A state-space model estimate method attempts to control the influence of uncertain environmental conditions. We proposed a new histogram matching algorithm and a searching location processing. Our system's initial algorithm gets most valuable hand feature, so lots of training is reduced compared with neural methods and it has satisfied the real-time requirement. The experimental results demonstrate the efficiency of the system.

Keywords: Gesture, histogram, window, location

1. Introduction

Much of the previous work in tracking and recognition has concentrated on rigid models of non-natural objects with corners and edges. Visual recognition systems have often exploited these features and the assumption of rigidity.

R. Kjeldsen and J. Kender [1] presented a real-time gesture system, which is used in place of the mouse to move or resize windows. In that system, the hand is segmented from the background using skin color then the hand's pose is classified by using a neural net. J. Triesch and C. Ven Der Malsburg [2] developed the combination of motion and color, so that stereo cues were used to track and locate the human hand. The hand posture recognition was based on elastic graph matching which can work in the presence of complex backgrounds in real time. But it is prone to noise and sensitive to the change of the illumination because its skin color detection was based on a defined prototypical skin color point in the HS plane. Davis and Shah [3] used markers for tracking fingertips and used the fingertip trajectories for recognizing seven gestures. In our paper, combined with the color segment, we proposed a hand searching method and finger prediction to solve the disturbance of illumination and lots of frames.

2. Hand Region Segment

Hand is a highly deformable articulated organ with many degrees of freedom. It can be used in expressing information for various purposes through different postures and motions. We represent the hand by a hierarchy of stable features at different scales, which captures the shape combined with hand skin color cues.

2.1 Color Space

We use the HSV color space [4] instead of the RGB color space. At each frame, we locate palm and finger regions by using the range of the skin color and the templates of shape. Using a background subtraction applied to color representation, hand regions can be extracted in *HSV* color space.

$$x = S*(V/100)*\cos(H)$$
$$y = S*(V/100)*\sin(H)$$
$$z = w*V$$
$$0 \le w \le 1$$
$$0 \le H \le 2\pi, 0 \le s \le 100, 0 \le v \le 100$$

where w is a normalization weight for V's change.

2.2 Initiation Processing

At first, the frame performs an initial segmentation using background appearance models [5]. We define the background as the hand (including fingers) and the background as anything else in the scene. Pixelwise background model is initialized with pixel statistics from each image in a sequence of an empty scene. The candidate images, compared with the model and pixel that differ significantly from the corresponding model pixel, have their color information added to a foreground histogram-based model. The map is

combined by using Bayes Rule to give a posterior probability map of foreground regions from which connected components are extracted.

The center of a user's hand is given as the point whose distance to the closest region boundary is the maximum. We calculate it with implement ting by a morphological erosion operation of an extracted hand region. First, we obtain a rough shape of the user's palm by cutting out the hand region at the estimated wrist (Fig. 1). Then apply a morphological erosion

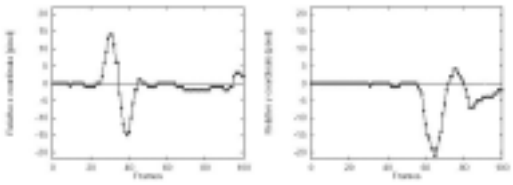

Figure1: The false rate of locating center of the hand

operator to the obtained shape until the region becomes smaller than a predetermined threshold value. Finally, the center of the hand region is given as the resulting region's center of mass.

We formulate the hand segmentation problem as follows: The hand is known to be in an image. The hand color is unknown in advance (different environments may result in different hand colors), but is assumed to be largely consistent within the image. In addition, we are concerned with initial hand segmentation, not subsequent hand tracking. Therefore we limit ourselves to a single image. Under these conditions, we want to segment the hand from the background, i.e. for each pixel in the image; we want to classify it as either a hand pixel or a background pixel (Fig.2).

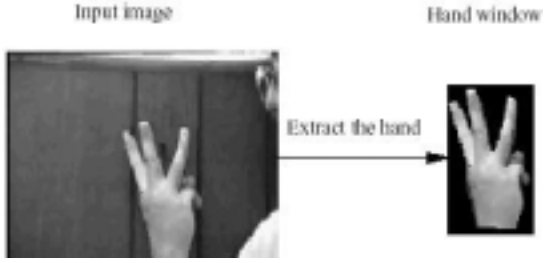

Figure 2: Hand area segment

2.3 Blob-Ridge Hand Model

Considering the relative orientation, position and scale, we adopted a straightforward view-based model [6] called blob-ridge model (Fig.3). The following features are extracted as: 1) finger tips as even finer scale blobs; 2) the fingers as ridges at finer scales; 3) the palm as a coarse scale blob. They are showed in figure1.

A parameter vector $V = (x, y, s, a, l)$ is defined to model translations, rotations and scaling transformations of the hand. (x, y) is the global position; s is the size; a is the orientation of the hand; discrete state $l = 1,2...5$. In the form of a probabilistic priority, we defined skin color

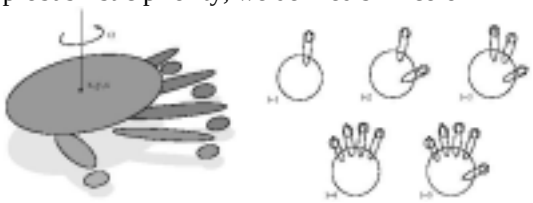

Figure 3: Five states of blob-ridge model

information as follows: a) In about 25 frames, hands were segmented from background and histograms over the chromatic information (u, v) were accumulated for skin regions R_s and background R_b ; b) histograms were summed up and normalized to unit mass; c) the skin probability of any pixel (u, v) was estimated:

$$w_{skin}(u,v) = \frac{\max(0, aR_s(u,v) - R_b(u,v))}{\sum_{u,v} \max(0, aR_s(u,v) - R_b(u,v))},$$

where $a = 0,1$ is a constant determining factor of the discrimination between skin and background color.

2.4 Histogram Matching

The histogram is based on hue and saturation only. Each axis of the histogram is discredited into a number of bins yielding an $N \times M$ array. The histogram for a given patch is populated by adding 1 to the bin that represents each pixel's hue and saturation values. The algorithm moves through the image and presents the control histogram (C) and the current histogram from the image (I) for the comparison. Histogram intersection is used to compare them. The match number of instances between C and I is defined as [4]:

$$M_{C,I} = \frac{\sum_{i,j} \min(H^C(i,j), H^I(i,j)}{\sum_{i,j} H^C(i,j)}$$

In experiment, an initial control seed must be designated by the user. The user is the one whose gesture to be recognized. The initial step is necessary for a rapid location. Although we can omit the step instead of auto location, but the latter's prove a lower efficiency for the next recognition cycles. After the first iteration, the algorithm must take such smart choice that the control seed may be used in the next iteration. A threshold hit occurs when the result of comparing a histogram patch with the whole control histogram is equal to the preset threshold value. The threshold can be defined by training in advance according to reference for a matching decision. Sometimes, isolated patches and other slice from the complex background may confuse the result. These false positives can be well reduced if single patches

are removed before picking a new control seed. Since the hand area is a rigid connected object, those patch without any eight neighbor connected should be eliminated (Fig.4).

3. Fingertip Location and Searching

Figure 4: Final segment from complex background

After deciding the hand regions approximately in an input image, we search for fingertips within those regions. This search process is more computationally expensive than arm extraction, so we should define the windows of searching for the fingertips. A search window based on orientation is estimated as the extracted wrist region's principal axis from the image moments up to the second order. We then search for fingertip within the new window. A cylinder with a hemispherical cap approximates a finger shape, and the projected finger shape in an input image appears to be a rectangle with a semicircle at its tip, so we can search for a fingertip based on geometry.

Based on a robust state-space estimation algorithm [7], we proposed a searching algorithm to reduce the search area to a smaller search window centered on predictions of the fingertip position. Based on their locations in the previous frame, one frame's fingertip location is predicted. The first frame is resumed in the center of the searching window. First we measure each fingertip's location and velocity and defined the state vector $\mathbf{x}_t$: $\mathbf{x}_t = (x, y, v_x, v_y)^T$, where x , y denote the location of fingertip; v_x, v_y denote the velocity in t th frame. Then we define observation vector $\mathbf{y}_t$ to represent the location of a fingertip detected in the t th frame. Two vector's relation is showed in the following equations:

$$\mathbf{x_{t+1}} = \mathbf{F}\mathbf{x_t} + \mathbf{G}\mathbf{w_t}, \qquad (1)$$

$$\mathbf{y_t} = \mathbf{H}x_t + v_t \qquad (2)$$

where $\mathbf{F}$ is the state transition matrix, $\mathbf{G}$ is the driving matrix, $\mathbf{H}$ is the observation matrix, w_t is the system noise for $\mathbf{x_t}$ in velocity direction and v_t is the observation noise.

Because the time is very short, the fingertip motion in the successive image frames can be assumed to be straight approximately. Then we get these vector matrixes:

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & \Delta T & 0 \\ 0 & 1 & 0 & \Delta T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{0} & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}^T, \ \mathbf{H} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & 0 \end{bmatrix}$$

The filtering formulation assumes the model parameters $\{F, G, H, R, Q\}$ to be accurate [8]. When this assumption is violated, the filter would be lower performance and then one of the model parameter is motivated to consider robust variants, which attempt to limit the effect of model uncertainties on the overall filter performance. We treat given parameters {*F*, *G*} as nominal values and assume that the actual values lie within a certain set around them. Replacing formulation (1) with the state-space model:

$$\mathbf{x_{t+1}} = (\mathbf{F} + \delta F_i)\mathbf{x_t} + (\mathbf{G} + \delta \mathbf{G_i})\mathbf{w}_t$$

where the situation variables {F, G} are modeled as:

$$[\delta F_i \quad \delta G_i] = M\Delta_i[E_f \quad E_g] \qquad (3)$$

for matrices $\{M, E_f, E_g\}$ and for an arbitrary contraction Δ_i , $\| \Delta_i \| \leq 1$. In our experiment, the value is 0.68. When our model changes dramatically in a particular time, we can allow the quantities $\{M, E_f, E_g\}$ to vary with time. The model (3) allows the designer to restrict the sources of uncertainties to a certain range space, and to assign different levels of distortion. The uncertainties can be due to the changes in lighting conditions, the background and object moving independently from each other, or to the user's pointing finger abruptly changing directions at variable speeds and accelerations.

4. Experiment Result

For this experiment, at first, we perform an initial segmentation using background appearance models. With the corresponding model, the likelihood that each pixel belongs to foreground or background is calculated. Then we find the center of the hand so help the next step to segment the hand and its fingers.

After converting the RGB color space, we detect the region of skin with histogram matching. The initial patch of skin called the control seed is used to initiate the iterative method. Blob-Ridge hand model is well done to elaborate the hand for tracking. In tracking hand fingers, the confidence in the tracking is used to specify an angle α (between 15^o and 180^o), which

constraints the directions of the search window placement. For a period of test, we decided to assign the following value to $\{M, E_f, G_f\}$ with the effect of the movement and illumination. Searching windows

$$M = [1 \quad 0.5 \quad 0.5 \quad 0.25 \quad 0.25 \quad 0.125]^T$$

$$E_g = [0 \quad 0 \quad 0.316 \quad 0.316 \quad 10 \quad 10]$$

$$E_f = [0 \quad 0 \quad 2 \quad 2 \quad 4 \quad 4]$$

Figure 5: Key Frame of Tracking Fingers

Algorithm	False Neg	False Pos	Error
Bayes(Actual)	7.5%	4.0%	11.6%
Bayes (Upper ound)	3.1%	3.6%	6.7%
Simple Method (Upper Bound)	6.4%	11.9%	18.3%
Our Algorithm	4.5%	3.4%	5.8%

Table 1: Algorithm Comparison

are placed within $\pm\alpha^{\circ}$ of the directions of the movement direction at a randomly generated angle. Figure 5 is the tracking of hand fingers. The fingertip and its most directions are located accurately. Table 1 is the comparison among this system and other Neural-based algorithm.

5. Conclusions

The work described offers a robust procedure for segmenting the hand in HSV color space. For locate the hand finger for translation, the state-space estimation algorithm creates a search window to reduce the processing area in the real-time system. The application of Blob-Ridge Model achieves perfect effect for segment. The final task of the translate work of gesture is to locate the precise position. With these coordinates and the angle, the information of gesture is recognized. Instead of the Alberto's histogram method [9] that is suit for plain moving recognition, we used the histogram intersection for a given patch. The procedure does not need a complex training step and as initial control is quite sufficient to identify a large skin area. The skin color initial processing in our experiment is not an important role and sensitive to illumination variety. In future, we should seek a more suitable motion model for different luminance and background.

6. Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 60441002).

References

[1] R. Kjeldsen and J. Kender, "Finding skin in color images," *Proceedings of International Conference on Automatic Face and Gesture Recognition*, (Killington, Vt.), pp.312–317,1996.

[2] Jochen Triesch, Christoph von der Malsburg, "Classification of hand postures against complex backgrounds using elastic graph matching", *Image and Vision Computing* 20, pp 937–943, 2002.

[3] J. Davis and M. Shah, "Visual gesture recognition," *IEE Proc. - Vision, Image, Signal Processing*, April, vol. 141, pp.101–105, 1994.

[4] J. D. Foley, A. Van Dam, S. K. Feiner, J. F. Hughes. "Computer Graphics: Principles and Practice," *2nd ed. Addison-Wesley*, Mass, pp. 590, 1993.

[5] G. McAllister, S.J.McKenna, and I.W.Reketts. "Towards a non-contact driver-vehicle interface," *IEEE Conference on Intelligent Transportation Systems*, Dearborn, Michigan, October, 2000.

[6] L. Bretzner and T. Lindeberg, "Qualitative multi-scale feature hierarchies for object tracking," *Journal of Visual Communication and Image Representation*, pp.115–129, Nov. 2000.

[7] A.H.Sayed, "A framework for state-space estimation with uncertain models," *IEEE Transactions on Automatic Control*, Vol. 46, no. 7, pp.998-1013, July 2001.

[8] R.K.Mehra, "On the identification of variances and adaptive Kalman filtering, "*IEEE Transaction on Automatic Control*, AC-15, pp.175-183, 1970.

[9] Tomita, A., Jr.; Ishii, R., "Hand shape extraction from a sequence of digitized gray-scale images Industrial Electronics, *Control and Instrumentation," IECON '94., 20th International Conference* , Vol. 3 , 5-9 pp.1925 – 1930, Sept. 1994.