Virtual Mouse----Inputting Device by Hand Gesture Tracking and Recognition[+]

Changbo HU, Lichen LIANG, Songde MA, Hanqing LU

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, P.O. Box 2728, Beijing 100080, China
cbhu,lcliang,masd,hqlu@nlpr.ia.ac.cn

Abstract. In this paper, we develop a system to track and recognize hand motion in nearly real time. An important application of this system is to simulate mouse as a visual inputting device. Tracking approach is based on Condensation algorithm, and active shape model. Our contribution is combining multi-modal templates to increase the tracking performance. Weighting value is given to the sampling ratio of Condensation by applying the prior property of the templates. The recognition approach is based on HMM. Experiments show our system is very promising to work as an auxiliary inputting device.

1 Introduction

Hand gesture based human-computer interfaces have been proposed in many virtual reality systems and in computer aided design tools. In these systems, however, the user must wear special physical devices such as gloves and magnetic device. Comparatively, vision based system is a naturer way. It is very attractive to utilize hand gesture as a kind of "mouse" using only visual information. But it is in fact an inherently difficult task although it is very easy for human being. One obvious difficulty is that hand is complex and highly flexible structure. Tracking and

[+] This work is funded by research grants from the NSFC （No.69805005） and the 973 research project（G1998030500）

recognizing hand motion is the basic techniques needed for this task. Several attempts to recognize hand gesture can be referred in [1,2,3,4].

Generally, gesture researches divide the recognition process into two stages. First, some low-dimensional feature vector is extracted from an image sequence. Most classical method is to segment the object out from image. And the information is obtained to describe the object. Second, recognition is preformed directly or indirectly on these observation data. However it is highly desirable to develop systems where recognition feeds back into the motion feature extraction because the motion style is tightly related to the activity. A great potential advantage of the multi-model approach is that recognition and feature extraction are preformed jointly, and so the form of the expected gesture can be used to guide feature search, potentially making system more efficient and robust.

In this paper we apply active shape model in Condensation framework to deal with hand shape tracking and apply multi-modal templates to use prior knowledge to increase system performance. This extended Condensation algorithm can achieve higher accuracy with lower size of sample-set. Standard Condensation algorithm is used to produces an approximation of an entire probability distribution of likely object position and pose, represented as a weighted sample-set. The weighted mean of this sample-set is used as a starting approximation of ASM. The output ASM is a refined estimation of the object position and pose, which is expected to have a high accuracy. This refined estimation is added to the sample-set with a relatively high weight. The advantages of this method are that: since the Condensation is just to provide a coarse estimation, the size of samples-set can be reduced, so increasing the computational speed. ASM produces a refined estimation, and adding this result to the sample set will improve the sampling of next frame. HMM net is used to train and recognize the hand motion continuously. Although performance of HMM largely depend on training, HMM can be performed in a real time. In general, the system does work well, in near real-time (10 frames/sec on PII 400without special hardware), and it can copes with cluttered backgrounds.

2 Related Works

Active Shape Models (ASM) proposed by Cootes[5] is a successful method to track deformable objects. It can get a high accuracy and can cope with clutter. But its

tracking performance greatly depends on a good starting approximation, so the object movement must be not too large, that limits its application. Random sampling filters [6,7] were introduced to address the need to represent multiple hypotheses when tracking. The Condensation algorithm [7] based on factored sampling has been applied to the problem of visual tracking in clutter. It has the striking property: despite its use of random sampling which is often thought to be computationally inefficient, the Condensation algorithm runs in near real time. This is because tracking over time maintains relatively tight distributions for shape at successive time-steps, and particularly so given the availability of accurate, learned models of shape and motion. The Condensation algorithm has a natural mechanism to trade off speed and robustness. Increasing the sample set size N can lower the tracking speed, but obtain a higher accuracy.

3 Our Approaches

3.1 Multi hand templates and PCA representation

Assuming one hand model is described by a vector x_e, a training set of these vectors is assembled for a particular model class, in our case the hand in its various different poses. The training set is aligned (using translation, rotation and scaling) and the mean shape calculated by finding the average vector. To represent the deviation within the shape of the training set, Principle Component Analysis (PCA) is preformed on the deviation of the example vectors from the mean. In order to do this the covariance matrix S of the deviation is calculated:

$$S = \frac{1}{E}\sum_{e=1}^{E}(x_e - \bar{x})(x_e - \bar{x})^T \qquad (1)$$

The t unit eigenvectors of S corresponding to the t largest eigenvalues supply the variation modes; t will generally be much smaller than N, thus giving a very compact model. It is this dimensional reduction in the model that will enable simple gesture recognition. A deformed shape x is generated by adding weighted combinations of v_j to the mean shape:

$$x = \bar{x} + \sum_{j=1}^{t} b_j v_j \qquad (2)$$

where b_j is the weighting for the j^{th} variation vector.

In our case, we set $t=5$. The hand model uses a mean shape and the first five modes of variation (the five eigenvectors that correspond to the largest five eigenvalues). The model hand shape is described solely in terms of these vectors and the mean shape. The model can be projected onto an image by specifying its location, in terms of scale s, rotation θ, x-translation t_x, and y-translation t_y, and its pose in terms of the variation vector weights $b_j \; i = 1,..5$.

3.2 Active Shape Model for hand contour representation

Active Shape Models (ASM) [5] were originally designed as a method for exactly locating a feature within a still image, given a good initial guess. A contour, which is roughly the shape of the feature to be located, is placed on the image, close to the feature. The contour is attracted to nearby edge in the image and can be made to move towards these edges, deforming (within constraints) to exactly fit the feature. The process is iterative, with the contour moving in very small steps.

The ASM uses a Point Distribution Model (PDM) to describe the shape of the deformable contour. The model is described by a vector $x_e = (x_1, y_1, ...x_N, y_N)$, representing a set of points specifying the outline of an object.

Given that the hand model's current projection onto the image is specified by s, θ, t_x, t_y and $b_i \; i = 1,..5$, we use the ASM algorithm [5] to approach a new position. The algorithm is used iteratively in order to converge on a stable solution.

3.3 Condensation Tracker

A full description and derivation of the Condensation algorithm is given in [7]. We describe here our improvement of the Condensation algorithm.

Given time-step t, Condensation algorithm is activated which produces an approximation of an entire probability distribution of likely object position and pose, represented as a weighted sample-set. The weighted mean of this sample-set

is used as a starting approximation of ASM. The output of ASM is an estimation of the object position and pose. This estimation is added to the sample-set.

The advantages of this method are that: since the Condensation is just to provide a coarse estimation, the number of samples can be reduced, so increasing the tracking speed. ASM provides a fine estimation, and this result can guide the Condensation algorithm in next time-step. Following gives a synopsis of the algorithm.

From the "old" sample set $\{s_{t-1}^{(n)}, \pi_{t-1}^{(n)}, n = 1,...N\}$ at timstep t-1, construct a "new" sample set $\{s_t^{(n)}, \pi_t^{(n)}, n = 1,...N\}$ for time t.

Construct the n^{th} of N-1 new samples as follows:

1. Select a sample $s_t'^{(n)} = (x_t'^{(n)}, i)$ as follows:
   (a) Generate a random number *j* with probability proportional to $\pi_{t-1}^{(j)}$. This is done efficiently by binary subdivision using cumulative probabilities.
   (b) Set $s_t'^{(n)} = s_{t-1}$
2. Predict by sampling from $p(X_t \mid X_{t-1} = s_t'^{(n)})$ to choose each $s_t^{(n)}$. In our case, the new sample value may be generated as

   $s_t^{(n)} = s_t'^{(n)} + Bw_t^{(n)}$ where $w_t^{(n)}$ is a vector of standard normal random variates, and BB^T is the process noise covariance.
3. Measure and weight the new position in terms of the measured features Z_t: $\pi_t^{(n)} = p(Z_t \mid X_t = s_t^{(n)})$
   then normalize so that $\sum_n \pi_t^{(n)} = 1$
4. Once the N-1 samples have been constructed: estimate the mean of the sample-set as

   $$\mathrm{E}(X_t) = \sum_{n=1}^{N-1} \pi_t^{(n)} s_t^{(n)}$$

5. Activate ASM algorithm with initialization being $E(X_t)$, and get a refined estimation $\hat{X}_t$
6. Add $\hat{X}_t$ with its weight $\pi_t^{(N)}$ to the sample set at time t.

This algorithm is also used iteratively until convergence or reaching time limit.

3.3 Apply template knowledge to weight sampling

In order to increase the quality and speed of the tracker, a prior weight related to template knowledge is multiplied to sampling weight at different position. Assume the knowledge of the template is vector x. and x satisfy the Gaussian distribution

$$p(x)=\frac{1}{2\pi}\|\Sigma\|^{-1/2}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \qquad (3)$$

where $\mu=\frac{1}{n}\sum_{k=1}^{n}x_k$, $\Sigma=\frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)(x_k-\mu)^T$ is the mean and variance learned beforehand from examples.

When the $X_t=s_t^{(n)}$ is sampled, a number of points $a_i, i=1..N$ are randomly selected inside the shape, for simplification we take the prior weight of this sampling as the mid-value of $p(a_i), i=1..N$.

Using skin color and texture of the template in this way can improve the tracking performance dramatically.

3.4 Mouse Action Recognition using HMM

Three actions are currently defined to the virtual mouse: left button down, right button down. and mouse moving. According our hand tracker, observation vector $(t_x, t_y, \theta, s, b_1, b_2, b_3, b_4, b_5)$ is concatenated to a time series. In HMM we use only vector $(b_1, b_2, b_3, b_4, b_5)$ to train and recognize mouse action. (t_x, t_y) is used to compute the position of the mouse. We use three-state, first order discrete HMM to perform this task. The topological net is as figure 1.

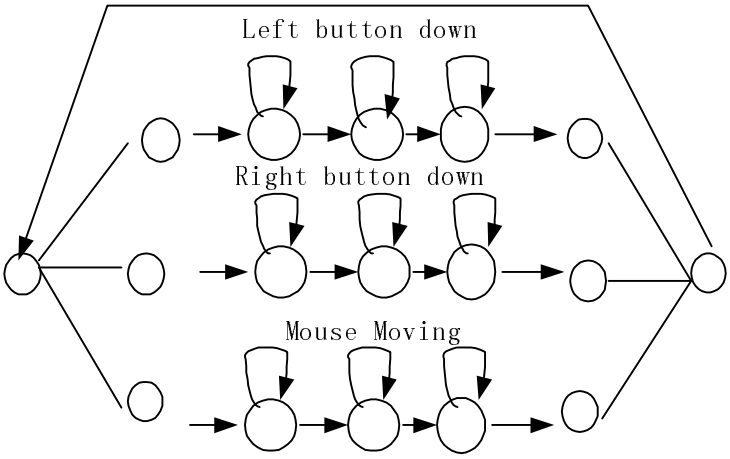

Fig. 1. . HMM network for continuous mouse action recognition

4 Experiments

The system run on a PII450 PC machine, about 10 frames can be tracked using our tracker and after about 2 frame delay the mouse's action is determined. The function of system is like this: the application shows a window, displaying the video images being seen by a camera. When a hand moving on the desk in the view of the camera, the hand was located and tracked until it moved out of view. Another window shows an mouse icon moving according the position; the color and pattern of the icon are changing with the action of the mouse.

The hand model includes 51 points. When motion is detected, the initial sampling is concentrated around the motion region. The size of sample-set for initial localization is 1500. Once the hand is located, the size of sample-set is reduced to 200. Three of the hand models are shown in Figure 2.

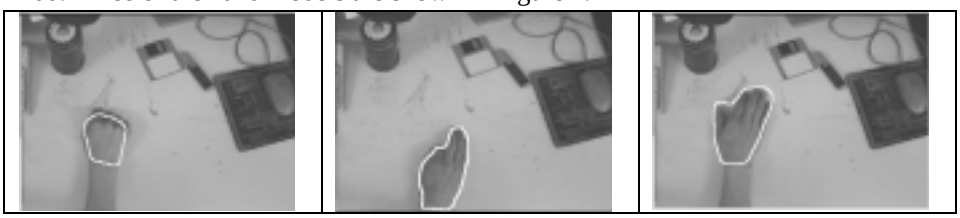

Fig. 2. Three ASM hand models

5.Conclusion

A system for hand tracking and gesture recognition has been constructed. A new method which extends the Condensation algorithm by introducing Active Shape Models and multi-modal templates, is used to fulfill this task. The system works in near real time, and virtual mouse position and action are interestingly controlled by hand.

Reference

1. Rehg, J. and Kanade, T. Visual Tracking of high dof articulated structures: An application to human hand tracking. In Eklundh, *Proc. 3rd ECCV*, 35-46, Springer-Verlag, 1994.
2. Freeman, W.T. and Roth, M. Orientation histograms for hand gesture recognition. In Bichsel, *Intl. Workshop on automatic face and gesture recognition*, Zurich. 1996.
3. Cohen, C. J., Conway, L., and Koditschek, D. .Dynamical system representation, generation, and recognition of basic oscillatory motion gestures. In *Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition*, 151-156, 1996.
4. Yaccob, Y. and Black, M. Parameterized modeling and recognition of activities. In *Proc. 6th Int. Conf. on Computer Vision*, 6, 120-127, 1998.
5. Cootes, T., Hill, A., Taylor, C. and Haslam, J. The use of active shape model for locating structures in medical images. *J. Image and Vision Computing*, 12,6, 355-366, 1994.
6. M. Isard and A. Blake. Visual tracking by stochastic propagation of conditional density. In *Proc. 4th ECCV*, 343-356, Cambridge, England, Apr. 1996.
7. G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. of Computational and Graphical Statistics*, 5(1):1-25, 1996.
8. U.Grenander, Y. Chow, and D.M. Keenan. *HANDS. A Pattern Theoretical Study of Biological Shapes.* Springer-Verlag. New York, 1991.
9. B.D. Ripley. *Stochastic simulation.* New York: Wiley, 1987.