Fist Tracking Using Bayesian Network

Conference Paper *in* Lecture Notes in Computer Science · October 2005

DOI: 10.1007/11573548_33 · Source: DBLP

4 authors, including:

Peng Lu
Beijing University of Posts and Telecommuni…
14 PUBLICATIONS **55** CITATIONS
SEE PROFILE

Yangsheng Wang
Chinese Academy of Sciences
116 PUBLICATIONS **1,871** CITATIONS
SEE PROFILE

Fist Tracking Using Bayesian Network

Peng Lu, Yufeng Chen, Mandun Zhang, and Yangsheng Wang

Institute of Automation, Chinese Academy of Sciences,
Beijing 100080, China
{plu,wys}@mail.pattek.com.cn

Abstract. This paper presents a Bayesian network based multi-cue fusion method for robust and real-time fist tracking. Firstly, a new strategy, which employs the latest work in face recognition, is used to create accurate color model of the fist automatically. Secondly, color cue and motion cue are used to generate the possible position of the fist. Then, the posterior probability of each possible position is evaluated by Bayesian network, which fuses color cue and appearance cue. Finally, the fist position is approximated by the hypothesis that maximizes a posterior. Experimental results show that our algorithm is real-time and robust.

1 Introduction

With the ever increasing role of computers in society, HCI has become an increasingly important part of our daily lives. One long-term goal in HCI has been to migrate the "natural" means that humans employ to communicate with each other into HCI. In this paper we mainly investigate the fist tracking algorithm, which can be used for HCI.

Tracking objects efficiently and robustly in complex environment is a challenging issue in computer vision. Many researches have been done on this area. Currently, particle filter [1] [2] and mean shift[3] [4]are two successful approaches taken in the pursuit of robust tracking. Particle filters, to apply a recursive Bayesian filter based on propagation of sample set over time, maintain multiple hypotheses at the same time and use a stochastic motion model to predict the position of the object. Maintaining multiple hypotheses allows the tracker to handle clutters in the background, and recover from failure or temporary distraction. However, there are high computational demands in the approach, and this is the bottleneck to apply particle filtering in real time systems. On the other hand, mean shift explores the local energy landscape, using only a single hypothesis. This approach is computational effective, but it is susceptible to converge to local maximum.

In this paper, we propose a novel fist tracking algorithm. First, some hypotheses about the fist's position are generated based on the motion cue and color cue. In this way, the number of the hypotheses is very limited. Then all the hypotheses are evaluated by the Bayesian network, which fuses appearance cue and color cue. Based on Bayesian network the tracking results are more robust.

In the remainder of this paper, section 2 introduces how to generate hypothesis. Section 3 describes the evaluating of the hypothesis by Bayesian network. Some experiments are shown in section 4.

2 Hypothesis Generation

Skin is arguably the most widely used primitive in human image processing research, with applications ranging from face detection and person tracking to pornography filtering. Color is the most obvious feature of the fist. It indicates the possible position of the fist. In order to use color cue of the fist ,a probability distribution image of the desired color must be created firstly. Many algorithms employ a manual process to extract color information in their initialization stage, such as CAMSHIFT. By this way an accurate skin color model can be obtained. But semi-automation is the main shortcoming of these algorithms. For achieving automation, the most popular method is to learn the skin color distribution from a large number of training samples. However, due to the different illumination and different camera lens, this color distribution is not always exact in real condition.

In this paper, we proposed a novel scheme to acquire the color distribution of the fist. Generally speaking, the skin color of hand and face, which belong to the same person, are the same or similar. For face detection, there are many successful algorithms. So we gain the skin distribution from the face of the player instead of the fist. In our scheme, three steps are needed for creating color model. The first stage is face detection. In this stage, our Haar-Sobel-like boosting [5] algorithm is used. Haar and sobel features are used as feature space, and GentleBoost is used to select simple classifiers. Haar features are used to train the first fifteen stages. And then sobel features are used to train the rest fourteen stages. The second stage is face alignment. At this stage, active shape model (ASM) [6] is used. The last stage is creating color model. The hues derived from the pixels of the face region are sampled and binned into an 1D histogram, which is used as the color model of the fist. Through this histogram, the input image from the camera can be convert to a probability image of the fist.

In order to deal with the skin-colored objects in the background, motion cue is used for our algorithm. We differentiate the current frame with the previous frame to generate the difference image using the motion analysis method in [7]. The method is to compute the absolute value of the differences in the neighborhood surrounding each pixel. When the accumulated difference is above a predetermined threshold, the pixel is assigned to the moving region.

Since we are interested in the motion of skin-colored regions, the logical AND operator is applied between the color probability distribution image and the difference image. And as a result the probability distribution image is obtained.

Suppose human hand is represented by a rectangle window, the possible position of the fist is gained as follows:

1, we sample the image from 320x240 to 160x120.

2, a subwindow x, y, w, h, where x and y are the left-top coordinate, w and h are the size of the rectangle, moves on the probability distribution image. And the sum of the pixels in subwindow is calculated. In our experiment, the w is 28 and the h is 35.

3, if the sum is below a certain threshold, it returns to step 2. And if the sum is above the threshold, the CAMSHIFT [8] is applied to getting the local maximum and the local color region size. And the center point of CAMSHIFT region is saved.

4, The pixels in CAMSHIFT region are set to zero. And it goes to step 2.

Based on these saved points, multiple hypotheses of the position and size of the fist are generated. For each saved point (x, y), four rectangle regions are taken out as four hypotheses. These regions have the same center (x, y), but differ in size. So the total number of the hypotheses is $Num \times 4$, where Num is the number of saved point.

3 Inference in the Bayesian Network

For robust tracking, we must eliminate the effect of other skin-color objects, which are also in motion, such as face. So besides skin-color feature more features should be taken into consideration. Fist appearance is the most significative feature and it can be used for differentiating fist from other objects. In our algorithm, two main features, color feature and appearance feature, are employed.

Bayesian network [10] is useful when we are trying to fuse more cue for tracking fist. Thus the posterior probability of fist region given observations of the other variables can be computed as follows:

$$P(X_k | C, A, X_{k-1}, X_{k-2}) \tag{1}$$

where C denotes color cue, A denotes appearance cue, X_{k-1} and X_{k-2} are the previous object state, X_k is the current object state.

The Bayesian network is shown in Fig.1.

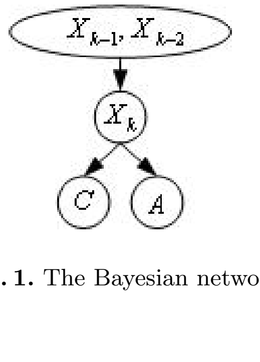

Fig. 1. The Bayesian network

By using conditional independence relationships we can get

$$\begin{aligned}&P(X_k|C, A, X_{k-1}, X_{k-2})\\&\propto P(X_k, C, A, X_{k-1}, X_{k-2})\\&\propto P(C|X_k)P(A|X_k)P(X_k|X_{k-1}, X_{k-2})\end{aligned} \tag{2}$$

3.1 Prior Model

The prior model $P(X_{k-1}, X_{k-2})$ is derived from the dynamics of object motion, which is modelled as a simple second order autoregressive process(ARP).

$$X_k - X_{k-1} = X_{k-1} - X_{k-2} + W_k \tag{3}$$

where W_k is a zero-mean Gaussian stochastic component.

The parameters of ARP model are learned from a set of pre-labeled training sequences.

3.2 Computation of the Color Marginal Likelihood

We define the color likelihood as follows:

$$P(C|X) = \frac{1}{n_c} \sum_{i,j} P_c(i, j) \tag{4}$$

where n_c is the scale of the likelihood, and $P_c(i, j)$ is the pixel in color cue image.

3.3 Computation of the Appearance Marginal Likelihood

Based on assumption of a Gaussian distribution, the probability of input pattern A, which belongs to fist class X can be modelled by a multidimensional Gaussian probability density function:

$$P(A|X) = \frac{exp[-\frac{1}{2}(A - \mu)^T \Sigma^{-1}(A - \mu)]}{(2\pi)^{N/2}|\Sigma|^{1/2}} \tag{5}$$

where μ is the mean vector of class X, Σ is the covariance matrix of class X.

By using PCA to reduce the dimension of X, the $P(A|X)$ is approximately estimated by equation (6), more detail about equation (6) can be found in [9].

$$\hat{P}(A|X) = exp\left[-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i}\right] exp\left[-\frac{\epsilon^2(x)}{2\rho}\right] \tag{6}$$

where $\hat{P}(A|X)$ is the estimation value of $P(A|X)$, $\epsilon^2(x)$ is the residual error, λ_i is eigenvalue of Σ, M is the dimensional of principal subspace, N is the dimension of total subspace.

In our experiment, the appearance parameters μ and Σ are learned from more than 5000 labelled images, which are collected from one hundred persons with three kinds of illumination. Some samples are shown in the Fig.2.

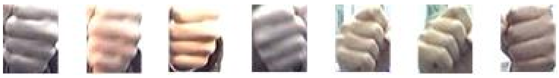

Fig. 2. Some samples used for training

4 Experimental Result

In order to compare with the mean shift algorithm, two experiments are done. In the first experiment, the color-motion based mean shift algorithm is used. When player's fist and face overlap, the tracker loses the fist. Some key frames of this experiment are shown in Fig.3. And it has no chance to recover. This is because the mean shift is a local optimal algorithm. In the second experiment, the proposed algorithm is applied. Based on multi-hypotheses it overcomes the local optimal. And using more cues, our algorithm becomes more robust. Some key frames of this experiment are shown in Fig.4.

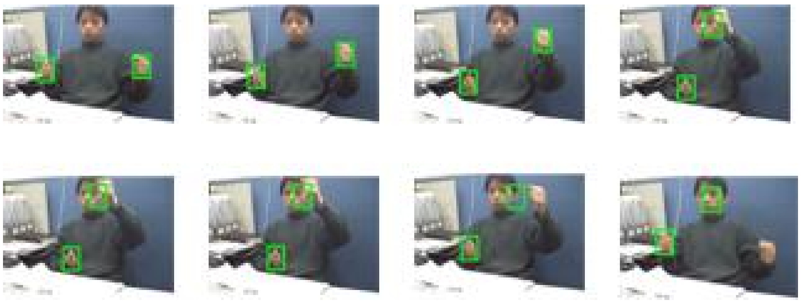

Fig. 3. Tracking failure by color and motion based Mean Shift algorithm

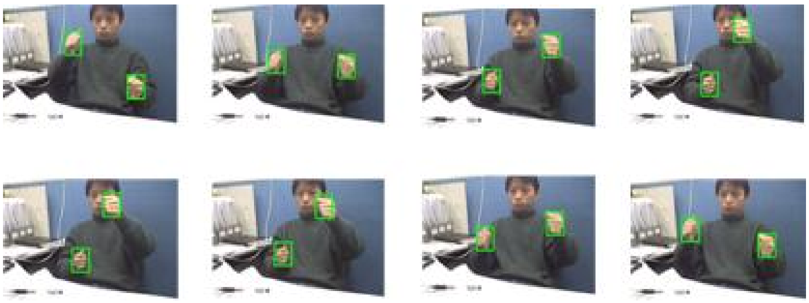

Fig. 4. Tracking by Bayesian network

All the experiments are done on a P4 1.7G machine with $512M$ memory. The normal recognition speed of our algorithm is about 29 fps.

5 Summarize

In this paper,a Bayesian network based fist tracking algorithm is introduced. Comparing with particle filter, the proposed algorithm, which uses more information to generate hypotheses, reduces significantly the number of hypotheses needed for robust tracking. And at the same time it overcomes the shortcoming of local optimal of the mean shift algorithm.

References

1. S. Maskell and N.Gordon, “A Tutorial on Particle Filters for On-line Nonlinear/Non-Gaussian Baysian Tracking”, in Proc. IEE Workshop ”Target Tracking: Algorithms and Applications”, Oct. 2001
2. C.F. Shan, Y.C Wei, T.N. Tan, “Real Time Hand Tracking by Combining Particle Filtering and Mean Shift”, The 6th International Conference on Automatic Face and Gesture Recognition (FG2004).
3. D. Comaniciu and V. Ramesh, “Mean Shift and Optimal Prediction for Efficient Object Tracking”, ICIP, Vancouver, Canada, 2000, pp. 70-73.
4. D. Comaniciu, P. Meer, Mean Shift Analysis and Applications, Proc. Seventh Int’l Conf. Computer Vision, pp. 1197-1203, Sept. 1999.
5. P.Lu, X.S.Huang, Y.S.Wang,“A New Framework for Handfree Navigation in 3D Game,” Proceedings of the International Conference on CGIV04.
6. T.F. Cootes, C.J.Taylor “Statistical Models of Appearance for computer vision,” Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PT, U.K. March 8, 2004.
7. H. P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, “Multi-Modal System for Locating Heads and Faces”, AFG, Killington, Vt, 1996, pp. 88-93.
8. G. R. Bradski, “Computer Vision Face Tracking For Use in a Perceptual User Interface”, Intel Technology Journal Q2, 1998.
9. B. Moghaddam and A. Pentland, “Probabilistic visual learning for object representation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):696-710, July 1997.
10. D. Heckerman, “A Tutorial on Learning With Bayesian Networks”, Microsoft Research Technical Report,MSR-TR-95-06