An Efficient Real-time System for Active Video Conferencing

Leila Sabeti[1] Jason Z. Zhang[2] Q.M. Jonathan Wu[3]

[1, 3] Department of Electrical and Computer Engineering
University Of Windsor
Windsor, Canada N9B 3P4
sabeti@uwindsor.ca, jwu@uwindsor.ca

Abstract

This paper describes a new approach for implementation of an efficient real-time system proper for using in active video conferencing sessions. A camera tracks presenter's head and its movements and orientations in an unconstrained environment automatically by its pan and tilt actions. Head or face is modeled by an ellipse using least square ellipse fitting algorithm. Linear Kalman filter is employed effectively, in addition to a measurement model based on bimodal color information. Experimental results demonstrate the effectiveness of the proposed system in overcoming challenges that trackers are facing in real world environments.

Keywords: Head tracking, Kalman filtering, video conferencing.

1. Introduction

Video conferencing is increasing attention for its ability in real-time communication over network between two or more people. Many analysts believe that videoconferencing will be one of the fastest-growing segments of the computer industry in near future. To make video conferencing systems, more analogous to physical conferences, using cameras with controllable pan, tilt and zooming abilities are inevitable. These cameras should automatically track the presenter in unconstraint environments. Challenges of this work include object occlusion, changing color with varying illumination, multiple moving people or other moving objects in the background, out of plane rotation and cluttered background.

Automatic human tracking has widely been investigated in robotic vision, active vision, automatic surveillance, facial feature tracking and analysis, 3D head modeling and video coding. Because of its rigid shape and constrained motion, a human head reveals reliable positional information for tracking. Although research related to head tracking using active cameras has not been extensively reported, some published methods and their drawbacks are studied here. Some papers only relied on facial color [4, 5, 6] in their trackers that can fail when the subject turns his face from the camera. There are other papers [8, 4] that didn't address challenges such as face occlusion and multiple moving people in their results. In addition, [9, 5, 6] didn't consider situations when there are other objects with similar skin or hair color in the background. [10] is very efficient in many challenging conditions; however, it can fail when face is mostly occluded. Another efficient work is done by [7]. In this work the combination of the intensity gradients and color histogram are exploited. These modules are orthogonal to each other and when one module fails the other one can come to its aid. However, the highlights in the image background can deduce unreasonably large values of the local gradient that can reduce the performance of the detection and tracking system. Also, searching for the largest sum of the gradient magnitude, increases computational complexity. Kalman filter [2] is a great tool for the stabilization of the tracking systems. It is applied to a few head tracking systems such as [4], but the implementation details are not clarified.

In this paper, a new approach for implementation of a real-time active head tracking system suitable for applications such as video conferencing or distance learning is explained with special attention to applying Kalman filters and direct least square fitting of ellipse. The result is an efficient tracker that overcomes almost all the challenges of an unconstrained environment; which are seen to be addressed only in parts by other similar works in literature.

This paper is structured as follows: our method of face and head detection and modeling is explained in section 2; head tracker implementation scheme and applying Kalman filter into this system is clarified in section 3; experimental result are shown in section 4 and finally the main conclusion and future work appear in section 5.

2. Face and head detection and modeling

In this work, after blob representation, face silhouette is detected and head is modeled by and an ellipse using direct least square fitting algorithm [1].

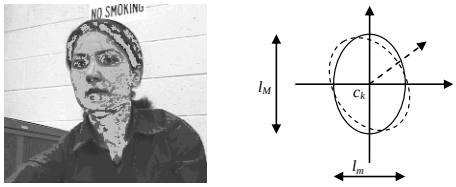

Fig. 1 Blob representation and elliptical model of head

2.1. Blob representation

Blob representation is exploited in this work because of its stability in detecting objects as they change their locations and orientations. For blob representation, color information of the model face and hair in the first frame is modeled by a histogram. The target blob in next frames is then identified by contrasting the histogram of the target blob with that of a template. After representation of the candidate head with pointing to the pixels related to the skin and hair color of the candidate, the contour of the head can be easily obtained by finding the edge points of head pixels. Then an ellipse can be fitted into these edge points using direct least square fitting algorithm to model the candidates head and later Kalman filter is applied to estimate the position of each parameter of this ellipse at each frame when head moves or rotates in the image.

2.2. Ellipse fitting

The literature on ellipse fitting divides into clustering techniques such as Hough-based methods [3] and least squares fitting methods. Least-squares techniques minimize sum of squared algebraic distances from the pixels to the ellipse. It is shown [1] how to fit an ellipse to scattered pixels effectively. This work contrasting previous works that are iterative, is direct and specific to ellipses. This method uniquely yields elliptical solutions that, under the normalization $4ac - b^2 = 1$, minimize the sum of squared algebraic distances from the points to the ellipse. General conic is represented by an implicit second order polynomial:

$$F(\mathbf{a}, \mathbf{x}) = \mathbf{a} \,.\, \mathbf{x} = ax^2 + bxy + cy^2 + dx + ey + f = 0$$

where, $\mathbf{a} = [a\ b\ c\ d\ e\ f]$ and $\mathbf{x} = [x^2\ xy\ y^2\ x\ y\ 1]^{\mathrm{T}}$. $F(\mathbf{a}; \mathbf{x}_i)$ is called the "algebraic distance" of a point (x, y) to the conic $F(\mathbf{a}; \mathbf{x}) = 0$. The fitting of a general conic may be approached by minimizing the sum of squared algebraic distances

$$D_A(a) = \sum_{i=1}^{N} F(X_i)^2$$

To constrain the parameter vector, **a,** so that the conic is forced to be an ellipse the $b^2 - 4ac$ is required to be negative. However, this constrained problem is difficult to solve in general, therefore the equality constraint $4ac - b^2 = 1$ is imposed. Solution results show that there is exactly one elliptical solution to this problem.

A head is modeled as an ellipse E obtained through the elliptical fitting process. Five parameters of the ellipse are adopted in the model: center coordinates $c_k = (x_k, y_k)$, lengths of the minor and the major axes l_m, l_M, and yaw angle a. The ellipse E is represented as $E = \{ x_k, y_k, l_m, l_M, a\}$. Therefore, using Kalman filtering, it is possible to adaptively update the position, shape and orientation of the head being tracked.

3. Applying Kalman filters

Kalman filter [2] is applied here to stabilize the tracking scenario, because color information alone is not robust enough for this purpose. The background, for instance, may contain skin color objects that incorrectly be considered as faces. In these situations, Kalman filter optimally estimates the position and uncertainty of a moving feature point in the next frame that is where to look for the feature, and how large a region should be searched in the next frame, around the predicted position, to be sure to find the feature within a certain confidence. Constant velocity model is considered for the movement of head in image frames and velocity changes are modeled by white Gaussian noise.

3.1. Constant velocity model

Consider a feature point $P_k = [x_k, y_k]^{\mathrm{T}}$, in the frame acquired at instant t_k, is moving with velocity $v_k = [v_{x,k}, v_{y,k}]^{\mathrm{T}}$. If we describe the motion on the image plane with the state vector $x = [x_k, y_k, v_{x,k}, v_{y,k}]^{\mathrm{T}}$, the system model of the Kalman filter will be:

$$p_k = p_{k-1} + v_{k-1} + \xi_{k-1}$$
$$v_k = v_{k-1} + \eta_{k-1}$$

Where ξ_{k-1} and η_{k-1} are zero mean white Gaussian random processes modeling the system noise. In terms of the state vector

$$x_k = \Phi_{k-1} x_{k-1} + w_{k-1}$$

With state transition matrix

$$\Phi_{k-1} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad w_{k-1} = \begin{bmatrix} \xi_{k-1} \\ \\ \eta_{k-1} \end{bmatrix}$$

As to measurements, we assume that a fast feature extractor estimates z_k, the position of the feature point p_k at every frame of a sequence. Therefore, the measurement model of the Kalman filter becomes

$$z_k = H_{k-1}x_{k-1} + \mu_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_k \\ v_k \end{bmatrix} + \mu_k$$

with H_{k-1} measurement matrix and μ_{k-1} a zero mean, white, Gaussian random processes modeling the measurement noise.

3.2. Implementation of Kalman filter

As we mentioned before five parameters are considered for the ellipse *E*. The five components of the state vector are estimated using five independent Kalman filters with each of the filters assigned to one parameter. Using five separate Kalman filters speed up the tracking scenario because, while filters work in parallel, each of them is inverting a smaller matrix according to the following equations:

$$P'_k = \Phi_{k-1} p_{k-1} \Phi^{\mathrm{T}}{}_{k-1} + Q_{k-1}$$
$$K_k = P'_k H^T_{k-1} (H_k P'_k H^T_k + R_k)^{-1}$$
$$x_k = \Phi_{k-1} x_{k-1} + K_k (z_k - H_k \Phi_{k-1} x_{k-1})$$
$$P_k = (I - K_k) P'_k (I - K_k)^{\mathrm{T}} + K_k R_k K_k^{\mathrm{T}}$$

where P'_k and P_k are state covariance matrices, K_k is the gain matrix, Q_k and R_k are the covariance matrices of the noise processes ξ_k and μ_k. The system model of the first Kalman filter for the *x* position of the center of the ellipse is

$$x_k = x_{k-1} + dt.v_{k-1} + \xi_{k-1}$$
$$v_{x,k} = v_{x,k-1} + \eta_{k-1} \qquad (1)$$

Therefore,

$$\begin{bmatrix} x_k \\ v_{x,k} \end{bmatrix} = \begin{bmatrix} 1 & dt \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_{k-1} \\ v_{x,k-1} \end{bmatrix} + \begin{bmatrix} \xi_{k-1} \\ \eta_{k-1} \end{bmatrix}$$

where x_{k-1} is the center of the ellipse, $v_{x,k-1}$ is its velocity and ξ_{k-1} and η_{k-1} are zero mean white Gaussian random processes modeling the system noise. As equation (1) shows, constant velocity model is considered for the feature motion and acceleration is approximated as white Gaussian noise. The measurement model is also

$$z_k = [1 \quad 1] \begin{bmatrix} x_k \\ v_{x,k} \end{bmatrix} + \mu_k$$

where μ_{k-1} is the measurement noise. Other four Kalman filters are implemented in a similar manner using constant velocity model.

Fig.2 Results of tracking and rotation (frames: 1-8-13-19)

4. Experimental results

Figures 2-5 show the results of the simulation of the head tracking system written in C++ and in Microsoft Visual studio environment. We have used EVI-D70 Sony color video camera to grab images. Figure 2 shows the tracking scenario when head moves or rotates. Using bimodal color for blob representation is an improvement over works such as [4, 5, 6] that only relied on facial color because; these trackers can fail when the subject turns his face from the camera. Unlike [8] and [4] that didn't address face occlusion and multiple moving people in their results or [9, 5,6] that didn't consider situations when there are other objects with similar skin or hair color in the background, this work has considered all these situations as figures 2-3 show. One reason is the effective adoption of Kalman filters; otherwise, the ellipse easily switches to either faces, especially when they cross. This is also true for occlusion of face as figure 3 shows. Finally figure 5 shows the camera pan/tilt actions in real-time after person moves in the scene that is also a great extension. Results are successful and stable in all the situations, which show the efficiency of the proposed head tracker. This real-time system is also very fast because of its high-speed measurement system based on skin color detection and employment of five separate Kalman filters. It works as an automatic camera man in video conferencing sessions; which, persistently tracks the presenter while he/she moves or turns his/her face. Also, because of camera's horizontal and vertical movements, this system increases our field of view and explores the remote site.

5. Conclusions

This paper describes a new approach for the implementation of real-time active trackers. In this system a camera detects and tracks face and head of the presenter and follows his/her movements and orientations. In this system,

Fig.3 Occlusion from left and right (frames: 85-89-124-126)

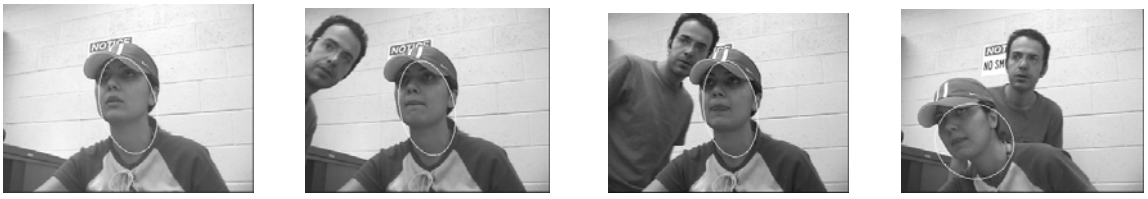
Fig.4 Presence of other people with similar skin color (frames: 13-17-21-35)

Fig.5 Camera movements (frames: 1-20-23-32)

after blob representation, face is modeled by an ellipse and tracked after applying Kalman filters to the tracking system effectively and in addition to our fast measurement model based on skin color information. Video clips[1] taken in various conditions from this real-time system show that our head tracking method is fast, efficient, stable and proper to be used in commercial video conferencing systems. It could also overcome challenges such as object occlusion, multiple moving people or other moving objects in the background, out of plane rotation, cluttered background and existence of other objects with similar skin color information. Future works includes an additional adaptive algorithm for environments with drastic change in illumination condition.

6. references

[1] D. Fitzgibbon, M. Filu, and R. Fisher, "Direct least square fitting of ellipses," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 476-480, 1999.

[2] R. Kalman, "A new approach to linear filtering and prediction problems,"*ASME J. Eng. 82*,1960.

[3] V.F. Leavers, "Shape Detection in Computer Vision Using the Hough Transform," *New York: Springer-Verlag*, 1992.

[4] N. Oliver, A. Pentland, and F. Bernard, "LAFTER: A real-time lips and face tracker with facial expression recognition," *Pattern Recognition*, vol. 33, no. 8, pp. 1369-1382, 2000.

[5] D. Comaniciu and V. Ramesh, "Robust detection and tracking of human faces with an active camera," *Proc. 3rd IEEE Computer Soc. Int'l. Workshop on Visual Surveillance*, pp. 11-18, 2000.

[6] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564-577, 2003.

[7] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 232-237, 1998.

[8] L. Yin and A. Basu, "Integrating active face tracking with model based coding," *Pattern Recognition Letters*, vol. 20, pp. 651-657, 1999.

[9] P. Fieguth and D. Terzopoulos, "Color-Based tracking of heads and other mobile objects at video frame rates," *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 21--27, San Juan, Puerto Rico, 1997.

[10] Y. Wu and T. S. Huang, "A Co-inference approach to robust visual tracking," *ICCV*, Vol. 2, pp. 26-33, 2001.

[1] www.vlsi.uwindsor.ca/~sabeti/headtracker