Real-time tracking for an integrated face recognition system [1]

Stephen J. McKenna, Shaogang Gong and Heather Liddell
Dept. of Computer Science
Queen Mary and Westfield College
University of London [2]

Abstract

A real-time tracking system suitable for surveillance applications was implemented on pipeline image processing hardware. Motion estimation was performed using a symmetric spatio-temporal filter in order to determine regions of interest. Normal components of visual motion were obtained at moving 'edge' features and Kalman filtering techniques were used for robust object tracking. The use of such a method within an integrated machine vision system for face recognition was discussed.

Introduction

In order to recognise peoples' faces in realistically unconstrained environments (e.g. such as arise in many security applications), a robust tracking and segmentation is required. This would provide a sequence of roughly segmented face images for recognition or verification purposes. Since people are almost constantly moving, motion estimation provides an effective technique for focusing of attention and discarding cluttered, static backgrounds. Perhaps the simplest approach to detecting areas of motion is to subtract each (possibly smoothed) frame from either a previous frame or from a reference background frame. While such methods have been used to segment regions of interest (e.g. Turk and Pentland (1991), Palavouzis (1994)) they are noisy and do not provide direct estimates of optical flow. More robust motion detection becomes possible by integrating intensity information over longer periods of time.

A large number of approaches for computing dense or sparse optic flow fields have been suggested and a discussion of these is beyond the scope of this paper (see e.g. Barron, Fleet and Beauchemin (1994)). Several tracking systems have used feature based flow methods which required explicit solution of the visual correspondence problem (e.g. Smith (1995)). The method used in the system described here was based on spatio-temporal filtering to detect moving image edge features. While more sophisticated spatio-temporal filtering methods have been suggested for estimation of visual motion (see e.g. Heeger(1988)), the method used was selected for its relative simplicity, ease of implementation and because it provided estimates of visual motion at moving edges where they were likely to be most reliable and relevant.

The following section describes the motion estimation method employed. The use of Kalman filters for object tracking is also outlined. Implementation using Datacube

[1]Supported by EPSRC Integrated Machine Vision project IMV GR/K44657 "Real-Time Target Identification for Security Applications"

[2]Mail address: QMW, Mile End Road, London E1 4NS, U.K. Email: stephen@dcs.qmw.ac.uk

hardware is mentioned and finally the use of such a tracking approach within an integrated machine vision system for face recognition is discussed.

Methods

Buxton and Buxton (1983) used a spatio-temporal Gaussian filter

$$G(x,y,t) = u(\frac{a}{\pi})^{3/2} \exp^{-a(x^2+y^2+u^2t^2)},$$

(where u is a time scaling factor and a gives the width of the filter as $w_m = \frac{2}{\sqrt{a}} = 2\sigma\sqrt{2}$) to define the following symmetric second order temporal edge operator:

$$m(x,y,t) = -(\nabla^2 + \frac{1}{u^2}\frac{\partial^2}{\partial t^2})G(x,y,t) \tag{1}$$

Consecutive image frames $I(x,y,t)$ from an image sequence were convolved with $m(x,y,t)$ yielding spatio-temporally filtered images $S(x,y,t)$. Zero-crossings in such images indicated moving edges (see figure 1). The normal component of optic flow can be estimated at these zero-crossings. Given a 3D convolution of sufficient width, the 3×3×3 neighbourhood $S_{ijk} = S(x+i, y+j, ut+k)$, of a zero-crossing (x,y,t) in S can be approximated using least-squares fitting to a linear polynomial $f_{ijk} = f_0 + f_x i + f_y j + f_z k$, where

$$f_x = \tfrac{1}{18}\textstyle\sum_{ijk} iS_{ijk}, \quad f_y = \tfrac{1}{18}\sum_{ijk} jS_{ijk}, \quad f_z = \tfrac{1}{18}\sum_{ijk} kS_{ijk} \tag{2}$$

The normal component of motion is then estimated as

$$(\dot{x}_\perp,\ \dot{y}_\perp) = \frac{-uf_z}{f_x^2 + f_y^2}(f_x,\ f_y) \tag{3}$$

The locations and normal flow components of spatio-temporal zero-crossings should enable multiple moving objects to be clustered and segmented, at least in cases where occlusions are not too severe and occluding objects have dissimilar visual motion.

Kalman filtering techniques can be used to track objects robustly from measurements of position, motion and shape. In the current implementation only a single moving region of interest is tracked. A dynamic system describes both coordinates of the centre of the object by its position, velocity and acceleration. For the x coordinate (and similarly for y), an update condition is given by (Torr *et al.* 1991):

$$\begin{bmatrix} x \\ \dot{x} \\ \ddot{x} \end{bmatrix}_{k+1} = \begin{bmatrix} 1 & \Delta t & \Delta t^2/2 \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ \dot{x} \\ \ddot{x} \end{bmatrix}_k + \begin{bmatrix} 0 \\ 1/2 \\ 1 \end{bmatrix} \mathbf{v_{acc}} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \mathbf{v_{pos}}$$

where $\mathbf{v_{acc}}$ and $\mathbf{v_{pos}}$ are noise covariances for assumptions about constant acceleration and position estimation.A measurement model is given by $\mathbf{z_{k+1}} = \mathbf{x_{k+1}} + \mathbf{w}$, where $\mathbf{x_{k+1}}$ is the actual state vector at time t_{k+1} and $\mathbf{w}$ is a noise covariance vector for measurement errors. $\mathbf{z_{k+1}}$ is the observation vector at time t_{k+1}. The height and width of the object's bounding box are separately modeled as

$$l_{k+1} = l_k + v_l$$

where v_l is the noise measure for the constant length assumption and the measurement model is $z_{k+1} = l_{k+1} + u$, where u is an error measure in the measurement of length.

The mean of the normal optic flow components given by equation (3) can be used as an estimation for the object's 2-dimensional velocity in the image plane. The temporal difference between velocities of successive frames provides an estimation for acceleration.

Real-time implementation

An initial implementation on a SunSparc 2 workstation ran at approximately 4 seconds per frame due to the high computational demands of performing spatio-temporal convolution (even when decomposed into 1D convolutions). Spatio-temporal convolution was therefore implemented on a Datacube pipeline image processing architecture with a MaxVideo250 board (see figure 2). Motion estimation and updating of the Kalman filters were performed on the host machine (a Themis SPARC 10MP).

Tracking currently runs at 8 frames per second and is not yet fully optimised. The system successfully tracks moving people/heads in indoor lighting conditions against arbitrary, static backgrounds.

Conclusions

A real-time tracking system suitable for surveillance applications was presented. Future work will involve extending the method to handle multiple motions by clustering the flow field in a temporally consistent manner (see e.g. Smith (1995), Shio and Sklansky (1991)) and running Kalman filters to track each object.

While the system as it stands is a fairly generic object tracker, the aim in our application is to track and segment faces (figure 3). Faces could be detected within moving objects by matching against 2D models such as those used for face detection in static scenes by Poggio and Sung (1994) and Rowley, Baluja and Kanade (1994). This matching process would have a greatly reduced search space due to focusing of attention on moving objects. In addition, heuristics could be used to guide the search (e.g. faces appear at the top of the body). The scale for matching is approximately determined by the tracking stage. The 3D pose of the head could also be 'tracked' to determine which 2D model to match. Subsequent recognition would benefit from the fact that scale and pose are already known.

In contrast to the usual approaches to face recognition which use single "snapshot" images, the availability of segmented face sequences allows temporal information to be used in the recognition process. Flow information obtained as part of the tracking process could also be used to guide the motion of feature detectors or fiducial points over time. Tracking of 3D pose could guide transformations of local feature detectors in the spirit of Maurer and von der Malsburg (1995). Alternatively it could guide the selection of appropriate 2D views for matching such as used in the eigenface approach (Pentland, Moghaddam and Starner (1994)). Gong *et al.* (1994) introduced the concept of temporal signatures of face classes and showed how temporal information might be used to constrain the recognition task in a useful way (Psarrou, Gong and Buxton (1995)). The use of temporal information should allow recognition at lower spatial resolutions than would otherwise be possible. This would be of important benefit in many applications.

References

Barron, J. L., Fleet, D. J. and Beauchemin, S. S. (1994) Performance of optical flow techniques, Int. J. of Computer Vision 12.

Buxton, B.F. and Buxton, H. (1983) Monocular depth perception from optic flow by space time signal processing. Proceedings of the Royal Society of London B-218.

Gong, S., Psarrou, A., Katsoulis, I. and Palavouzis, P. (1994) Tracking and Recognition of Face Sequences. In: European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production, Hamburg, Germany, November.

Heeger, D. (1988) Optical flow using spatiotemporal filters, Int. J. of Computer Vision 1, 270-302.

Maurer, T. and von der Malsburg, C. (1995) Single-view based recognition of faces rotated in depth, Int. Workshop on Automatic Face and Gesture Recognition, Zurich, 248-253.

Palavouzis, P. (1994) Head Tracking for Face Recognition. M.Sc. Thesis, Department of Computer Science, QMW, London.

Pentland, A., Moghaddam, B. and Starner, T. (1994) View-based and modular eigenspaces for face recognition, I.E.E.E. Conf. on Computer Vision and Pattern Recognition, Seattle, WA, July.

Poggio T. and Sung K. K. (1994) Example-based learning for view-based human face detection, ARPA94 (II:843-850)

Psarrou, A., Gong, S. and Buxton, H. (1995) Modelling spatio-temporal trajectories and face signatures on partially recurrent neural networks, I.E.E.E. Int. Conf. Neural Networks, Perth, Australia.

Rowley, H. A., Baluja, S. and Kanade T. (1995) Human face detection in visual scenes, Carnegie Mellon Computer Science Technical Report CMU-CS-95-158R.

Shio, A. and Sklansky, J. (1991) Segmentation of people in motion, IEEE Workshop on Visual Motion, Princeton, NJ., 325-332.

Smith, S. M. (1995) ASSET-2: Real-time motion segmentation and shape tracking, I.C.C.V., 237-244.

Torr, P., Wong, T., Murray, D. and Zisserman, A. (1991) Cooperating motion processes. In: B.M.V.C., Glasgow, Scotland.

Turk, M. and Pentland, A. (1991) Eigenfaces for recognition. Journal of Cognitive Neuroscience 3.

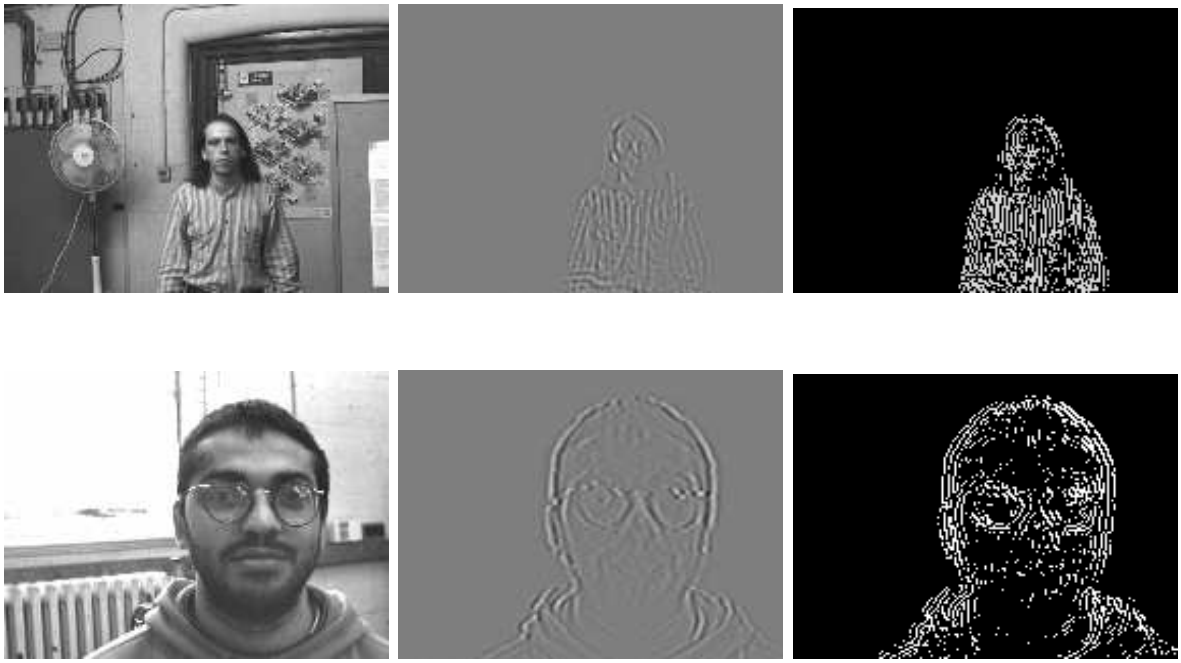

Figure 1: *The images on the left are from sequences taken in our lab. Those in the middle are the results of spatio-temporal convolution. Finally, on the right are the corresponding, thresholded spatio-temporal zero-crossings.*

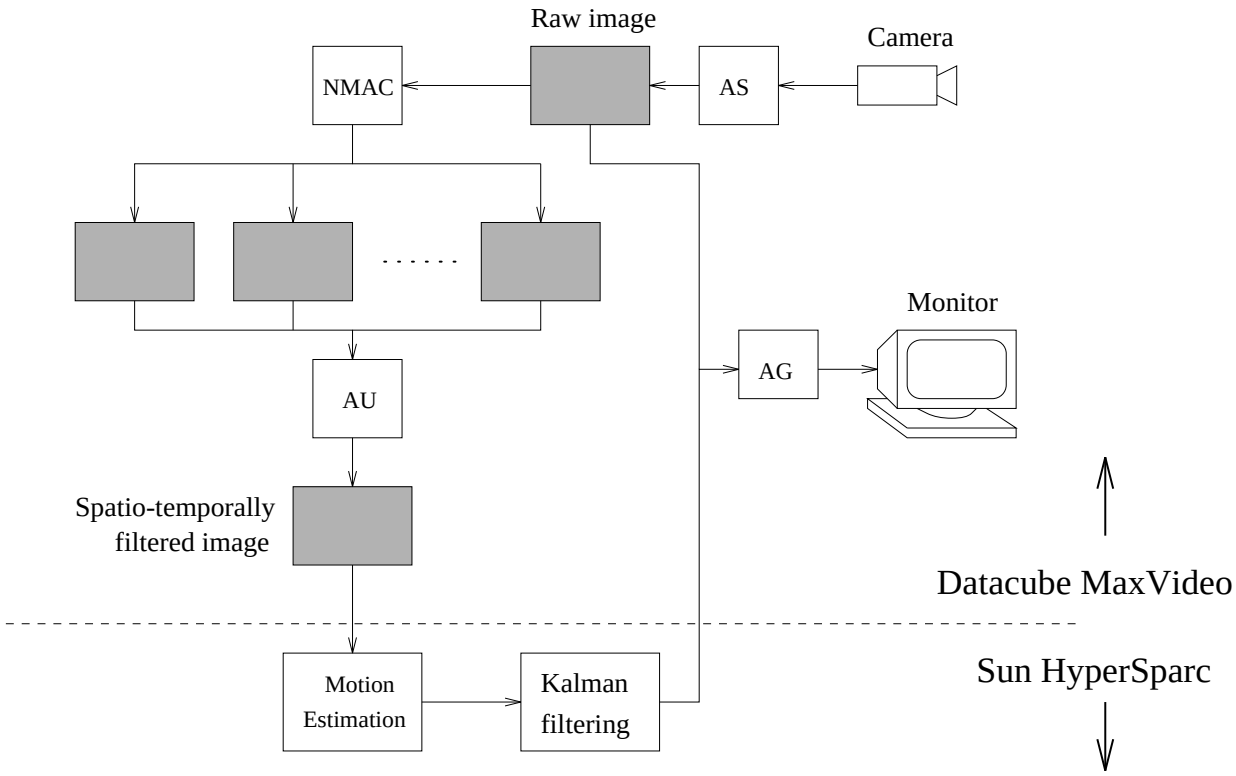

Figure 2: *Implementation on Datacube hardware.*

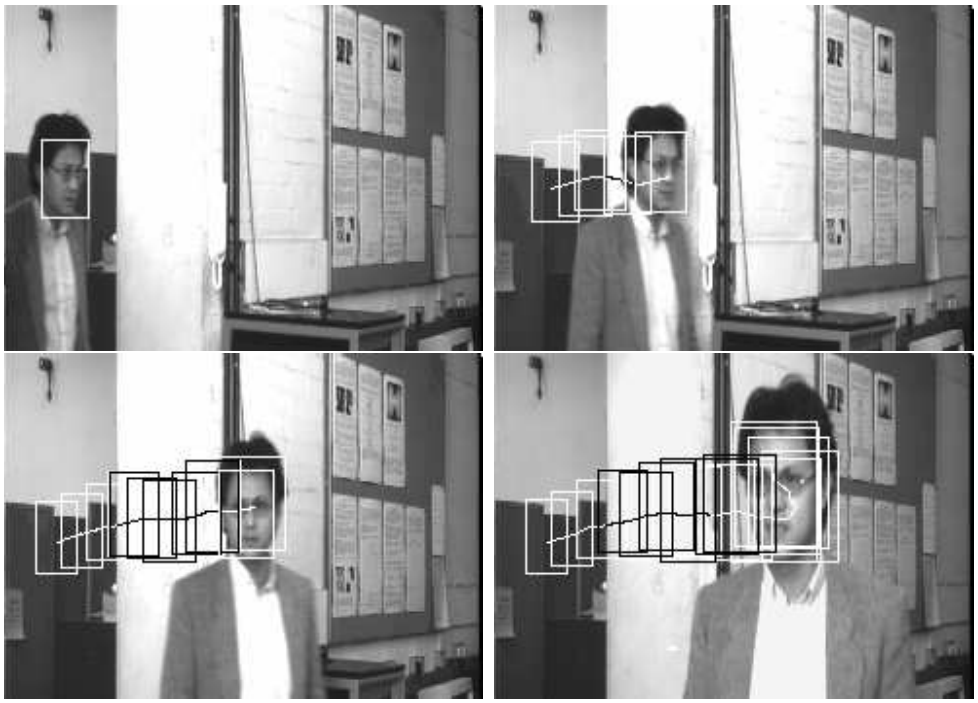

Figure 3: *A segmented face sequence.*