To appear in: Proc. 1st Int. Conf. on Audio- and Video-based Biometric Person Authentication, Lecture Notes in Computer Science, Springer Verlag, 1997

Non-intrusive Person Authentication for Access Control by Visual Tracking and Face Recognition

Stephen J. McKenna and Shaogang Gong

Machine Vision Laboratory, Department of Computer Science, Queen Mary and Westfield College, Mile End Road, London. E-mail: stephen@dcs.qmw.ac.uk

Abstract. Face recognition systems typically operate robustly only within highly constrained environments. This paper describes work aimed at performing face recognition in more unconstrained environments such as occur in security applications based on closed-circuit television (CCTV). The system described detects and tracks several people as they move through complex scenes. It uses a single, fixed camera and extracts segmented face sequences which can be used to perform face recognition or verification. Example sequences are given to illustrate performance. An application in non-intrusive access control is discussed.

1 Introduction

There has been significant development in recent years in the use of computer vision systems for automatic face recognition. This technology is now beginning to be deployed outside the laboratory in applications such as access control (e.g. [4]). These systems typically assume a single face imaged at high resolution in frontal or near-frontal view. The environments in which they operate are highly constrained compared to the scenarios in which we perform recognition during our daily lives. The work described in this paper aims towards performing face recognition in more realistically unconstrained environments.

We will concern ourselves with how computer vision techniques can be used to implement a person authentication system based upon the kind of camera set-up used in closed-circuit television security systems. For illustration, let us imagine one possible scenario in which a camera is positioned above an entrance through which we wish to restrict access. A vision-based authentication system could allow the person's identity to be verified *transparently* in a 'hands-off' and socially acceptable manner without even requiring the person to stand facing a camera. All that is required of the person is not to be uncooperative by looking away from the camera or by obscuring the face. In the few seconds it takes for a person to approach the entrance the system must:

1. Detect the presence of a moving person-like object and start to track it
2. Locate the person's face and track it
3. Establish the person's identity by performing face recognition

Given a sequence of centred face images normalised in scale, a number of techniques could be applied to obtain real-time recognition. The use of sequences

as opposed to 'snap-shots' aids recognition through the use of temporal information and eases the burden of accuracy placed upon the final face recognition. This paper focuses on the task of obtaining normalised face sequences in order that such a recognition might be performed. Several non-trivial issues must be addressed by a robust solution:

Illumination conditions typically vary to a much greater extent than is normally encountered in laboratory experiments. There is often exterior lighting and a system should operate during the day and the night-time. Problems can be caused by reflected motion in windows, picture frames and other reflective objects. Shadows may be cast in unpredictable ways.

Occlusion can occur when several people are present in the scene. A system needs to handle multiple moving objects and to reason about occlusions.

Image resolution is lower than used by most face recognition systems. The use of image sequences as opposed to isolated snapshots can compensate for this poor spatial resolution.

A single camera: In order to allow easy and economical integration into CCTV security scenarios, a system should operate robustly using only one uncalibrated, fixed camera.

Pose changes: Face appearances from different poses must be associated from a sequence.

The system described here attempts to address these issues and while it does not offer complete solutions to them all, it does achieve accurate face tracking in a wide range of scenarios. The system combines two complementary visual cues: motion and facial appearance. In order to reliably detect significant motion we use an approach based upon the detection of spatio-temporal zero-crossings. A clustering algorithm is used to group the detected motion and moving objects are tracked robustly using Kalman filters. An appearance-based face detector confirms that an object is a person and facilitates face tracking which aids subsequent motion grouping and enables segmented face sequences to be produced. Each resulting face image has a parameter indicating confidence in its segmentation. The face sequences can be used to perform recognition. Such a system allows learning to be performed on-line by building and updating face representations in a manner which is transparent to the user.

Several groups have described methods for tracking people and faces but with objectives other than or in addition to performing face recognition. Darrell et al. describe a face tracking system which is similar in some respects for use in an interactive room [1]. It tracks a single person as he/she walks around a room with well-controlled lighting conditions. A second narrow-angle camera is used to image the face.

The remainder of this paper describes the system in more detail. Some examples are given to illustrate the performance of the tracker.

2 Tracking Faces

This section describes the system for tracking peoples' faces. More detailed descriptions of some aspects can be found elsewhere [5, 6].

2.1 Tracking multiple motions

Visual motion is estimated by convolving the intensity "history" of each pixel $I(x, y, t)$ with the second-order temporal derivative of a Gaussian function $G(t)$ yielding an image of temporal zero-crossings $S(x, y, t)$ [2]:

$$S(x, y, t) = \frac{\partial^2 G(t)}{\partial t^2} * I(x, y, t)$$

This convolution is applied to six consecutive frames in the current implementation. Global illumination changes and changes in the intensity level of static objects are not erroneously detected as motion. Figure 1 shows an image from a sequence of two people walking through our laboratory along with its temporally filtered image (in the middle) and the detected temporal zero-crossings (on the right).

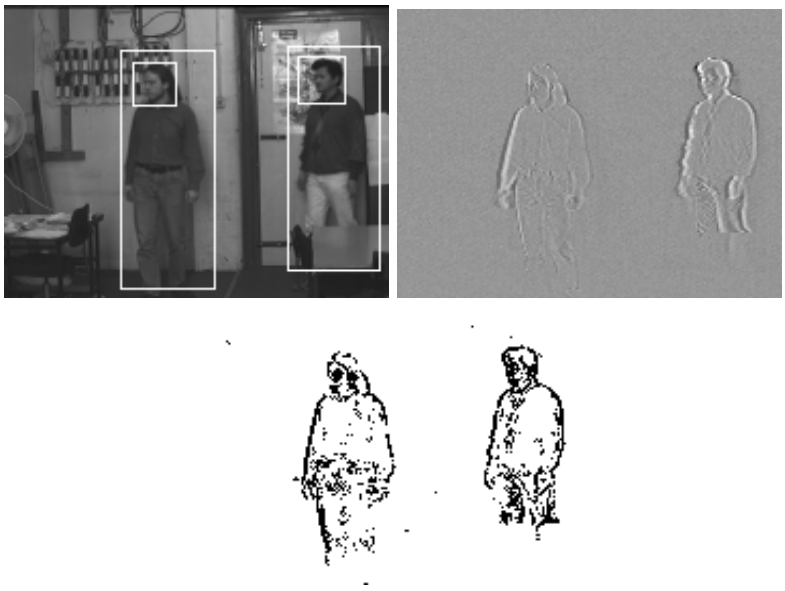

Fig. 1. *The motion-based tracker. Top left: An image from a sequence taken in our lab with bounding boxes for the tracked people and their heads overlaid. Top right: The image after temporal convolution. Bottom: Detected temporal zero-crossings.*

In order to handle multiple moving people in a scene, the detected temporal zero-crossings are clustered into separate areas of motion which ideally correspond to different people. Stationary people, lack of texture, (near-) occlusions and strong shadows can all cause errors in this clustering process. However, most of these errors are corrected through the use of temporal filtering techniques and through integration with appearance-based cues provided by a face model (described in the next section).

The motion pixels are clustered in real-time using a two-level coarse-to-fine method which also calculates a mean position and a diagonal covariance matrix for each cluster. Any clusters which are too small to correspond to people are discarded. The resulting Gaussian clusters are matched from frame-to-frame using time-symmetric matching which yields a temporally consistent list of moving 'objects' [8].

The measurements of the mean and co-variance of each Gaussian cluster are noisy and are treated using Kalman filters. A person's global body motion is modelled as second-order. System noise is modelled for inaccuracy in the assumption of constant acceleration and for 'motion in jerks' [9]. Each cluster's size is modelled as a noisy constant system. Once established, each Kalman cluster is assigned a *persistence* parameter which allows it to track for a short time in the absence of a matching motion cluster. In the absence of a match, the cluster list is updated with the Kalman filters' predicted values for the cluster's mean and co-variance.

2.2 Tracking people and their faces

The above method based solely upon motion cues can already track several moving people within a complex scene. The purpose of the system, however, is to accurately locate and track faces. Simple heuristics based upon knowledge of the human form can be combined with the tracked motion clusters to give a crude localisation of the human head but an accurate and robust face tracking system requires further knowledge which is provided here in the form of a face appearance model implemented using a neural network.

Rowley et al. describe a neural network for face detection in static scenes [7]. (See [5] for a discussion and references on face detection). A similar approach is used here to localise and track faces in dynamic scenes. Localised connectivity (receptive fields) can be used to constrain the network, yielding improved generalisation. Radial basis function networks (RBFN) were also used to perform face detection. These were faster to train at the expense of a slight decrease in accuracy. In the tracking application described in this paper it was useful to trade accuracy for speed. Therefore, a multi-layer perceptron with just 8 hidden units was used to form a compact representation for face detection. It was trained on 9000 face images and a similar number of 'near-face' patterns. These 'near-face' patterns were selected from previously misclassified non-face patterns in an iterative training scheme.

The motion-based tracking is used to 'bootstrap' the face detection neural network using the estimated head region. Once a face is detected, it is tracked

from frame-to-frame using the network. If the face becomes obscured for several consecutive frames resulting in an absence of high confidence face detections, the neural network tracker can lose lock on the face. Recovery is then performed using the head region estimated from the motion-based tracker.

Face tracking is used to provide feedback to the motion clustering process to help deal with occlusions. For example, two people walking arm-in-arm would normally be clustered as a single moving object. However, consistent tracking of their faces constrains the motion grouping process, forcing it to form two Gaussian clusters.

Temporal convolutions are performed on a Datacube MaxVideo250 with the remaining computation on a Themis 10MP host machine. The system is capable of sub-second frame rates.

3 Example sequences

Figure 2 shows an example sequence of a girl being tracked as she approaches the camera. The girl's motion was successfully detected and a Kalman filtered cluster established within the first 15 frames of the sequence. The face was then reliably tracked with the exception of a few frames. At about frame 70 the girl turned her head suddenly to her left and the tracker lost her face. This resulted in inaccurate face boxes with lower confidence values. The tracker had recovered lock on the face by frame 85. A sequence of accurately segmented face images was obtained by simply discarding any images with confidence values below a threshold of 0.99.

Figure 3 shows an example of the system tracking two people as they approach along a corridor. This scene has multiple interior and exterior sources of illumination. There are additional difficulties caused by the reflected motions on the polished floor as well as moving shadows. However, the people are successfully tracked with many frames of the sequences yielding accurate face segmentations. Three frames from the sequence are shown. In the first, the facial resolution is very poor and face tracking is not always accurate. The bounding box for the person on the right-hand-side is artificially elongated due to motion reflected in the floor but this does not prevent correct localisation of his head. In the second frame shown, feedback from the face tracking neural network has been used to prevent the two people being grouped as a single moving object. The confidence in the face detections is low because the faces are far from their frontal views. However, the resulting face images are still well centred as a result of recent high confidence detections. In the third frame shown, one of the people has just walked out of the field of view. The system still attempts to find a matching motion cluster for a few more frames in case the person has ceased to move.

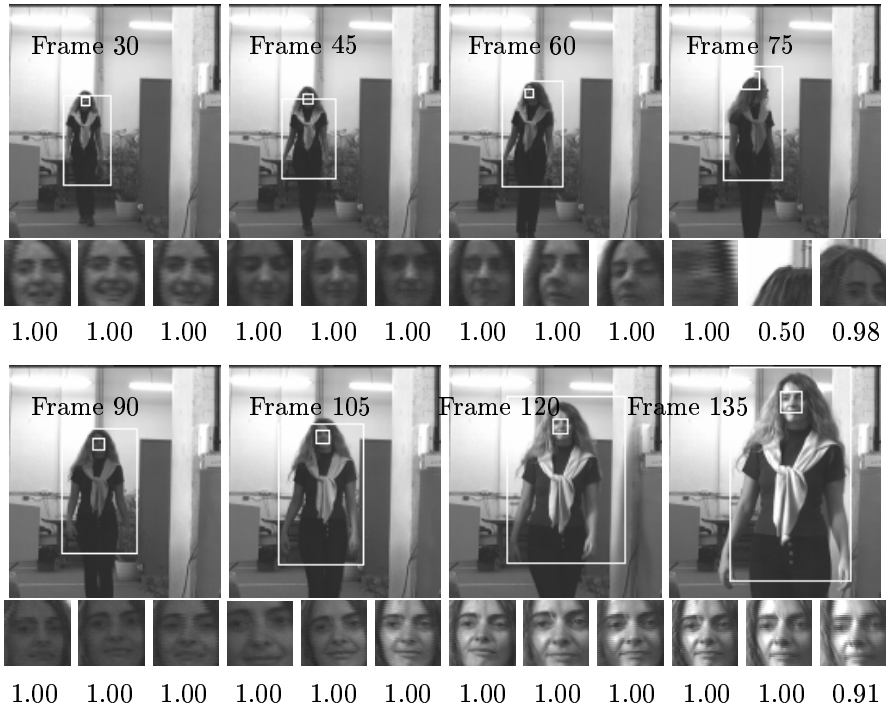

Fig. 2. *A girl is tracked for recognition as she approaches the camera. Bounding boxes for the Kalman filtered motion cluster and the associated tracked face region are shown overlaid on every* 15^{th} *frame. The face is shown centred and normalised in scale every* 5^{th} *frame. The confidence values for each segmented face are also given.*

4 Discussion

The tracker described produces sequences of segmented faces and the confidence measure can be used to obtain accurately segmented sequences. Once normalised for scale these sequences can be used to perform face recognition/verification. A person authentication system could learn and store representations for previously unseen people on-line in a non-intrusive manner. These representations could then be used for recognition in the future. In addition, reliably tracked and identified faces could be used to update representations of known people enabling more robust recognition by adapting the representation as a person's appearance changes over time.

Integration of the face tracking system with face recognition methods is currently underway. Howell and Buxton report some preliminary results using radial basis function networks to perform recognition from sequences produced by the face tracker [3]. We are investigating a range of view-based techniques for identity recognition with the face sequences produced by the tracking system.

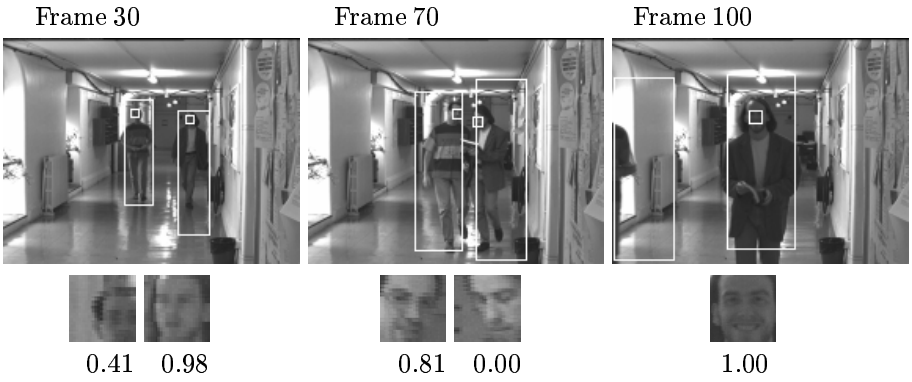

Fig. 3. *Two people are tracked as they walk along a corridor.*

5 Acknowledgements

This research was supported by the EPSRC Integrated Machine Vision grant IMV GR/K44657 "Real-Time Target Identification for Security Applications".

References

1. T. Darrell, B. Moghaddam, and A. P. Pentland. Active face tracking and pose estimation in an interactive room. In *CVPR*, 1996.
2. J. H. Duncan and T.-C. Chou. On the detection of motion and the computation of optical flow. *IEEE PAMI*, 14(3), 1992.
3. A. J. Howell and H. Buxton. Towards unconstrained face recognition from image sequences. In *IEEE Second International Conference on Automatic Face and Gesture Recognition*, 1996.
4. W. Konen and E. Schulze-Kruger. Zn-face: A system for access control using automated face recognition. In *IEEE First International Workshop on Automatic Face and Gesture Recognition*, pages 18–23, 1995.
5. S. McKenna and S. Gong. Tracking faces. In *IEEE Second Internatioanl Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, US, October 1996.
6. S. McKenna, S. Gong, and J. J. Collins. Face tracking and pose representation. In *BMVC*, Edinburgh, Scotland, September 1996.
7. H. A. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. Technical Report CMU-CS-95-158R, CMU, 1995.
8. S. M. Smith and J. M. Brady. A scene segmenter: visual tracking of moving vehicles. *Engineering applications of Artificial Intelligence*, 7(2):191–204, 1994.
9. P. Torr, T. Wong, D. Murray, and A. Zisserman. Cooperating motion processes. In *BMVC*, pages 145–150, Glasgow, 1991.

This article was processed using the LaTeX macro package with LLNCS style