Inferring Abnormal Scene Events from Pixel-Wise Change in Dynamic Scence

Ng, J.; Gong, Shaogang

QUEEN MARY
AND WESTFIELD COLLEGE
UNIVERSITY OF LONDON

Department of Computer Science

Research Report No. RR-00-07 ISSN 1470-5559 December 2000

Inferring Abnormal Scene Events from Pixel-Wise Change in Dynamic Scenes

J. Ng & S. Gong

Inferring Abnormal Scene Events from Pixel-Wise Change in Dynamic Scenes

J. NG & S. GONG

Abstract

Monitoring dynamic scenes for unusual events forms an integral part of an automated visual surveillance system. Background modelling have been used in both indoor and outdoor environments to detect and track foreground moving objects. Activities in the scenes are thus only modelled in terms of object positions, velocities and trajectories. We propose a two-stages approach using Gaussian Mixture Models for modelling slow changes in individual pixel intensities and CONDENSATION-based trajectory matching on phase information for capturing the temporal characteristics of changes in pixel history. The dual models allow for the detection of abnormal scene events before any object clustering and modelling are required.

1. Introduction

In visual surveillance tasks, automated systems are confronted with environments under constant change. For such dynamic scenes, visual change is a function of the context (semantics) in the scene which has been observed by the camera and is not necessarily an indication of problematic events occurring in the scene. Rather, a deviation from the established patterns of change in the image might signal an abnormal event under way. For example, constant rapid motion can be observed on a busy road and a sudden absence of motion might reveal an accident, while rapid motion on the sidewalk areas which have previously only been used by slower moving pedestrians is likely to imply abnormal behaviour. Provided that the normal patterns of change can be extracted and learnt, the recognition of "deviant-motion" ought to be used to detect abnormal events.

Previous works addressed the problem of scene-monitoring by explicitly modelling change in terms of the dynamics of moving objects. Object detection and tracking have been performed by numerous methods such as object colour models [8] and background subtraction [10, 5], while the trajectories of moving objects have been modelled using Kalman filters [9] and augmented Hidden Markov densities [4]. Colour models are very useful for distinguishing between different classes of objects based on appearance but it can be difficult to acquire accurate models in unconstrained environments without human intervention. Trajectory-based event recognition relies heavily on the assumption of accurate segmentation. However, it is often the case that the task of object segmentation and object-based event recognition are mutually dependent, which degenerate into a chicken and egg problem. Moreover, explicitly tracking people in busy scenes such as in a shopping mall can be conceptually difficult and computationally intractable.

In fact, object segmentation is not an essential prerequisite for tracking change in the scene. The physical layout and function of most scenes constrain the patterns of change occurring in different areas of the scene. For example, accumulated patterns of change have been found to be highly consistent with spatial location in captured frames [9]. The global dynamics of a scene can therefore be captured by monitoring change independently at the level of individual pixels. To cope with different types of change, we propose a two-stage scheme. Slow change, caused by lighting cycles for example, is modelled with adaptive mixtures of Gaussians which are typically used in background modelling tasks. On the other hand, fast changes require the modelling of temporal structures which are extracted from the rates of change of pixel values. Such temporal structures are learnt and recognised using probabilistic trajectory matching techniques [1]. Deviant-motion is recognised as changes which do not fit pre-learnt patterns.

In Section 2, we make use of Gaussian mixture models in a formal framework for probabilistically detecting the occurrence of deviant-motion in regions of zero to slow change. A novel approach is proposed in Section 3 involving phase information of pixel change and CONDENSATION-based model matching to profile the temporal characteristics of scene events based on pixel-wise information alone. Phase statistics are introduced in Section 4 to determine the pace of change for particular pixels and select appropriate models. Finally, experiments are provided in Section 5 to compare the capabilities of the deviant-motion recognition model to a Gaussian mixture based background modelling system similar to that proposed by Stauffer and Grimson [10].

2. Detecting Change

Dynamic scenes exhibit a wide spectrum of change both in terms of the speed and types of change. Areas of little or no change can simply be modelled according to their appearances or colour distributions. In these cases, static objects found in the background of the scene contribute most to the appearance of specific pixels and any other observed colours are attributed to moving foreground objects. More specifically, given a stream of colour values for a given pixel, $\mathbf{x}_t \in \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t\}$, the variation in the (r, g, b) components of $\mathbf{x}_t$ can be described in terms of Gaussian means μ and covariances Σ. However, illumination specularities or swaying objects such as plants can cause the colour distributions of pixels to become split into multiple modes or clusters [8, 10]. Multiple Gaussian components are therefore necessary to model the more complex distributions. A Gaussian mixture model $p(\mathbf{x}) = \sum_{i=1}^{k} \omega_i \cdot \psi(\mathbf{x}_t, \mu_i, \Sigma_i)$ can be used to describe any irregular distribution, where ω_i represents the mixing parameter and $\psi(.)$ the Gaussian kernel.

In unconstrained environments, the colour distributions of specific pixels rarely remain static. Changes in the lighting conditions or the patterns of sway for multi-modal pixels cause slow shifts in the parameters of the mixture model. First, we make these parameters adaptive in a similar fashion to the online approximation technique described in [10]. We then introduce a novel concept for recognising meaningful change by further analysing explicitly the temporal characteristics of the history for each adaptive pixel colour. More specifically,

1. The means and covariances of Gaussian components are updated according to a pre-determined learning rate α,

$$\mu_t = (1 - \alpha)\mu_{t-1} + \alpha \mathbf{x}_t \quad (1)$$

$$\Sigma_t = (1 - \alpha)\Sigma_{t-1} + \alpha(\mathbf{x}_t \cdot \mathbf{x}_t^{\mathrm{T}}) \quad (2)$$

2. The mixing parameter ω is updated according to whether the Gaussian is responsible for observation $\mathbf{x}_t$ at time t,

$$\omega_{u,t} = (1 - \alpha)\omega_{u,t-1} + \alpha(M_{u,t}) \quad (3)$$

$$M_{u,t} = \begin{cases} 1, & \text{if u is the responsible Gaussian} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

3. The Gaussian components in the mixture are ordered according to the products of (a) their weights, which reflect the amount of time that each have been observed and (b) the inverse of their variances to promote static objects with smaller variances. The first b Gaussians which account for a proportion T of the time are considered as the background components.

$$b = argmin_B \left\{ \sum_{i=1}^{k_{max}} \omega_i > T \right\} \quad (5)$$

4. New observations $\mathbf{x}_t$ which are not represented in the mixture model are assigned to new Gaussians with reasonably large initial covariances. The new component is added to the mixture if the maximum number of components k_{max} has not been reached or otherwise, replaces the weakest component in the mixture. In our case, we use k_{max}=6 so that the mixture model mostly captures the static components responsible for slow change and a few foreground components.

We then use Bayes' rules to formulate the probability of pixel values $\mathbf{x}_t$ belonging to a pre-learnt set of background Gaussian clusters as opposed to the recent foreground components introduced into the mixture:

$$P(\mathbf{x}_t | background) = \frac{p(background|\mathbf{x}_t)P(background)}{p(\mathbf{x}_t)} \quad (6)$$

Fast deviant-motion is therefore detected as pixel values with very low probability $P(\mathbf{x}_t|background)$. The configuration of the background set stores the accumulated history of the frequency of observation of each component in the mixture over a longer time scale. The state of the background set can therefore capture slow changes in the colour distribution of pixels. Depending on the type of change previously observed during the training sequence of the model, the background set can be locked to prevent new Gaussian components from being accepted.

3. Recognising Meaningful Change

Rapidly changing visual phenomena exhibited by the motion of animated objects typically involves both non-rigid deformations [5, 6] and purposeful trajectories [2, 9, 4]. Gaussian mixture models with pre-fixed learning and adaptation rates, as described in Section 2, are not suitable for modelling the wide range of possible observations from both slow and fast changes, therefore are not capable of differentiating normal and abnormal events in dynamic scenes. For event detection, the temporal sequence of change in the appearance of moving objects is more relevant. Furthermore, we suggest that the rate of change, and its derivatives such as phase information, provide a more appropriate measure to extract the structure of generic rapid change from pixel colour values. Phase information Ph can be measured from the response of pairs of quadrature filters [7]. Figure 1 shows the phase information collected from a sample sequence of a person moving from left to right.

Here, we adopt the second derivative in the form of Laplacian of Gaussian $g(y)$ and its Hilbert transform $h(y)$, phase-shifted by 90^o, as a pair of quadrature filters of tem-

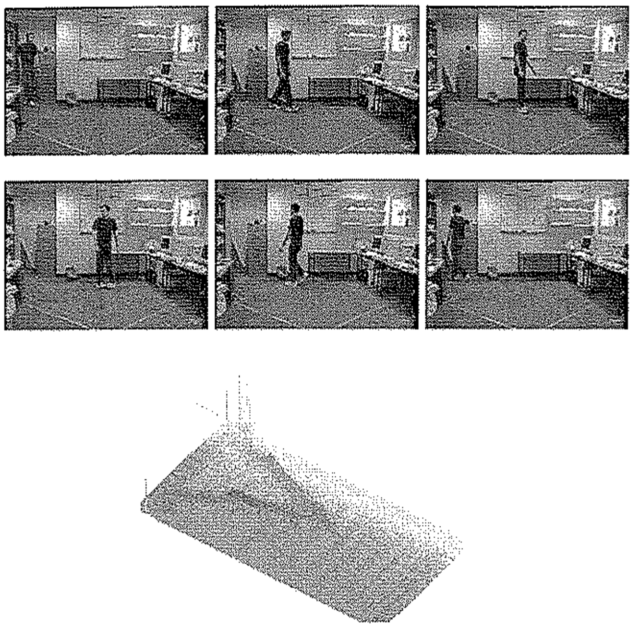

Figure 1: Spatio-temporally correlated phase information from an accumulated sequence of 10 repetitions of a person moving from left to right. The vertical axis represents intensity of phase response, the longer diagonal axis, a concatenated 2D image, and the shorter diagonal axis represents time.

poral size v for extracting phase,

$$Ph(\mathbf{x}_t) = \left[\sum_{i=o}^{v} g(v \times (i-v)/3.5) \cdot \mathbf{x}_{t-i}\right]^2 + \left[\sum_{i=o}^{v} h(v \times (i-v)/3.5) \cdot \mathbf{x}_{t-i}\right]^2 \quad (7)$$

The filter masks $g(y)$ and $h(y)$ are respectively defined as,

$$g(y) = \eta(2y^2 - 1)e^{-y^2} \quad (8)$$

$$h(y) = \kappa y + \lambda y^3 e^{-y^2} \quad (9)$$

where the normalising coefficients are $\eta = 0.9213$, $\kappa = -2.205$ and $\lambda = 0.9780$.

The temporal support of the filters are chosen to be relatively small, about 10 frames for sequences captured at 8 Hz, in order to obtain better responses to the type of fast change being modelled. The continuous filter responses can then be learnt as probabilistic trajectories and used as models for trajectory matching algorithms. Propagating conditional matching densities provides the desired flexibility, both in terms of temporal and amplitude scaling, for modelling detailed variations across different types of change [3, 1, 2]. Furthermore, our implementation of conditional density propagation does not require the segmentation of captured phase information into atomic components.

More precisely, the matching hypotheses or states are defined as $(\mu, \phi, \alpha, \rho)$ where μ is the model being matched, ϕ, the position within the model, α, the amplitude scaling parameter and finally ρ, the temporal scaling parameter. A finite set of k states are then propagated across time according to the observation probability defined in [1],

$$P(\mathbf{y}_t|\mathbf{s}_t) = \exp\left\{-\sum_{j=0}^{w-1} \frac{(\mathbf{y}_{t-j} - \alpha m^{\mu}_{(\phi-\rho j)})^2}{2\sigma_{\mu}(w-1)}\right\} \quad (10)$$

States are randomly chosen from a cumulative probability distribution of the normalised observation probabilities of all the states in the set. Then, states with observation probability higher than a certain threshold of probable match (we use 0.3) are propagated to the next time step according to,

$$\mu_t = \mu_{t-1} \quad (11)$$

$$\phi_t = \phi_{t-1} + \rho_{t-1} + N \quad (12)$$

$$\alpha_t = \alpha_{t-1} + N \quad (13)$$

$$\rho_t = \rho_{t-1} + N \quad (14)$$

where N is additional normal noise for performing local search in the parameter space.

A percentage of the states are reserved for random initialisation. The probability of the change in a given pixel at time t matching the pre-learnt models of normal change is given as the best observation probability over a set of k states,

$$P(\mathbf{y}_t) = argmax_{i=1}^{k}(P(\mathbf{y}_t|\mathbf{s}_{i,t})) \quad (15)$$

Unlike in [11, 2], this probabilistic matching framework does not make use of any prior knowledge as the phase signatures have not yet been discretised into atomic components and the dynamics of the latter are still unknown. However, discretisation and prior knowledge are not important in our case as only a matching probability against all the stored models is required.

4. Selective Event Detection

The appropriate level of modelling for different parts of dynamic scenes is highly dependent upon the nature of change occurring in different regions of the image. Whilst Gaussian mixture models cannot distinguish between different types of change, temporal phase signature is less suited for monitoring slow change. As the magnitude of phase response of quadrature filter pairs can be arbitrary and do not have any real correspondence with qualitative perceptions of zero, slow or fast change, the phase information of particular pixels provide a good cue as to the pace of change observed at the pixel's location. Phase statistics can therefore be accumulated in the form of averages, or alternatively maxima, over the training period of the system as shown in Figure 2.

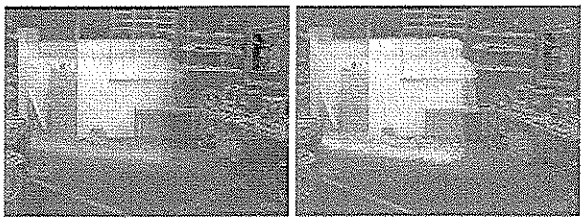

Figure 2: From left to right: (a) The average and (b) the maximum of the filter responses coded as grey-levels from a scene with motion.

The Gaussian mixture-based model is essentially a detector of fast change with poor ability to differentiate between different types of change. However, it can provide an effective mechanism for linking semantic descriptions of change to actual phase information. A phase threshold to determine the occurrence of fast change can be obtained as follows,

$$Ph_{thresh} = \frac{1}{n}\sqrt{\sum_{i=0}^{l} Ph(\mathbf{x}_i)b(\mathbf{x}_i)} \qquad (16)$$

where

$$b(\mathbf{x}_i) = \begin{cases} 1, & \text{if } P_{max}(\mathbf{x}_t) < 0.5 \\ 0, & \text{otherwise} \end{cases} \qquad (17)$$

$$P_{max}(\mathbf{x}t) = argmax_{j=0}^{v}\{P(\mathbf{x}_{t-i}|background)\} \qquad (18)$$

$$n = \sum_{i=0}^{t} b(\mathbf{x}_i) \qquad (19)$$

Pixel values $\mathbf{x}_t$ having a phase energy, $Ph(\mathbf{x}_t)$ and less than threshold Ph_{thresh} are considered to be undergoing slow change and are probabilistically matched with the Gaussian mixture model for that pixel. Phase signature matching is used only for pixel values with phase energies higher than the threshold. The extensive propagation of model matching hypotheses used in phase signature matching can be computationally expensive. The phase threshold is therefore used to extract semantic meaning from phase information and also as a "focus of attention" mechanism to limit abnormal phase signature matching to those pixels in the scene where infrequent changes occur.

5. Experiments

To illustrate the capabilities of our deviant-motion recognition model, here we give some preliminary results from experiments designed to test the model's ability to differentiate between different types of change in the scene and therefore signal abnormal events when they occur. The system is exposed to a training sequence of approximately 1000 frames (from 20 repeated sequences) containing two people who are carrying out their normal routine of entering the office from the door on the right, moving to the left for an inspection of the room and leaving by the same door, as shown in Figure 3. We expect the system to learn the patterns of change in the scene and therefore profile typical behaviour from 20 repetitions by 2 persons, observed during training.

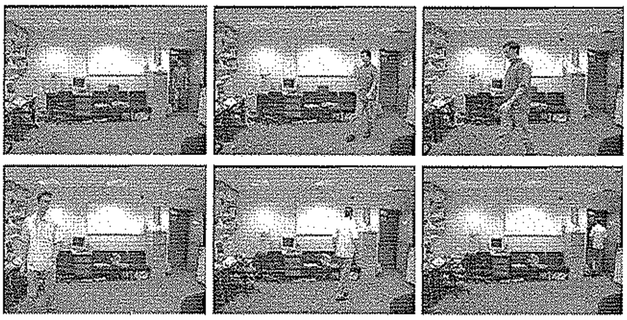

Figure 3: Selected frames from the training sequence.

After training, the system was tested on five different sequences of activities performed by one of the persons from training. The testing sequences contain similar events to the training sequences but with differences in the characteristics of the performed movement so as to render either part or the whole of the activity "abnormal". They are described as:

- Slow movement - The test subject walked at a slower speed while keeping to the same trajectory of motion.
- Fast movement - The test subject walked at a faster speed along the same trajectory.
- Stationary pause - The person walks as usual except for a brief pause in the middle of the right-left movement.
- Jump - A quick jump is introduced in the middle of the right-left movement of the person.
- Falling box - The system was retrained to include a static object (a box) in the lower right corner of the room. The context of the environment allows for movement by the person in the scene. However, the event of the box falling over is not considered as normal.

Table 1 shows the results of deviant-motion recognition over the five test sequences. For the "Slow Movement" sequence, deviant motion has been successfully detected

Table 1: Abnormal event detection results for the test sequences totalling over 900 frames. This is based on the deviant-motion recogniser's (DMR) and Gaussian Mixture Model's (GMM) ability to correctly classify frames containing normal and abnormal motion.

Test Sequence	Abnormal Frames			Normal Frames		
	Total	Recognition Rates(%)		Total	Recognition Rates(%)	
		DMR	GMM		DMR	GMM
Slow Movement	254	64.6	-	11	100.0	100.0
Fast Movement	64	35.9	-	9	55.6	100.0
Stationary pause	74	95.9	-	94	100.0	20.2
Jump	13	84.6	-	98	66.3	9.1
Falling Box	42	100.0	100.0	108	100.0	36.1

in most of the frames containing abnormally slow events. However, the faster abnormal events were not as successfully detected in the "Fast Movement" sequence as the duration of the events were too short compared to the temporal support of the phase filter pair. The temporal support of the filters and the rate of capture of frames should therefore be tuned according to an estimate of the duration of events occurring in the scene. The deviant-motion model perform better in sequences which involve more semantically meaningful deviations from pre-learnt patterns of change, such as "Stationary pause" and "Jump". For the two sequences, most of the frames containing normal movement were recognised as such, while deviant motion were detected in the frames where the extra pause and jump occurred. As the Gaussian mixture models do not possess any knowledge of context, they detect all movement as abnormal events and results are therefore not provided in Table 1. In the "Falling box" sequence, the movement of the person is considered as normal in the particular context of that office environment. . The deviant-motion recognition system correctly matches the per-pixel change occurring in the frames with its pre-learnt models and disregards the movement event as normal. Both the deviant-motion model and the Gaussian mixture model do detect the event when the box falls over.

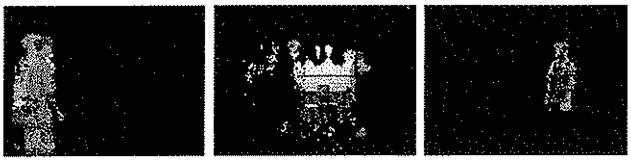

Figure 4: Selected detection results from sequences of (a) slower motion, (b) faster motion and (c) stationary pause from left to right.

Although the detection of deviant-motion is performed at the pixel level, the *simultaneous* triggering of all the pixels involved in the abnormal event, as illustrated in Figure 4, shows that the system can successfully detect unknown deviant-motions both in time and spatial location. The system has therefore been able to capture the structure of normal change from the training sequence. It is then used to distinguish fine-grain differences in the type of change to which it is exposed. Such ability provides additional flexibility over Gaussian Mixture Models for monitoring complex events in dynamic environments. Furthermore, the deviant-motion recognition system selectively detects only abnormal events and can therefore be used as a pre-attentive mechanism for initiating person and object tracking.

6. Conclusion and Future Work

In this paper, a deviant-motion recognition model is presented that does not require object segmentation for differentiating changes caused by different spatio-temporal events and semantics in dynamic scenes. By modelling change at the pixel level alone, we have shown that a system using a combination of mixtures of Gaussians for slow pixel change and phase signature tracking over time for fast changes, is able to infer the structure of change in local areas of images and distinguish abnormal from normal scene and object motion patterns and events, including the introduction of foreign or the removal of existing objects to and from a familiar scene.

Currently, the phase signatures are extracted as continuous streams and used in their entirety as models for trajectory matching. Cross-correlation of the statistics of change across pixels can help to obtain better general parameters for the Gaussian Mixture Models and the phase models.

References

[1] M.J. Black and A.D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and ex-

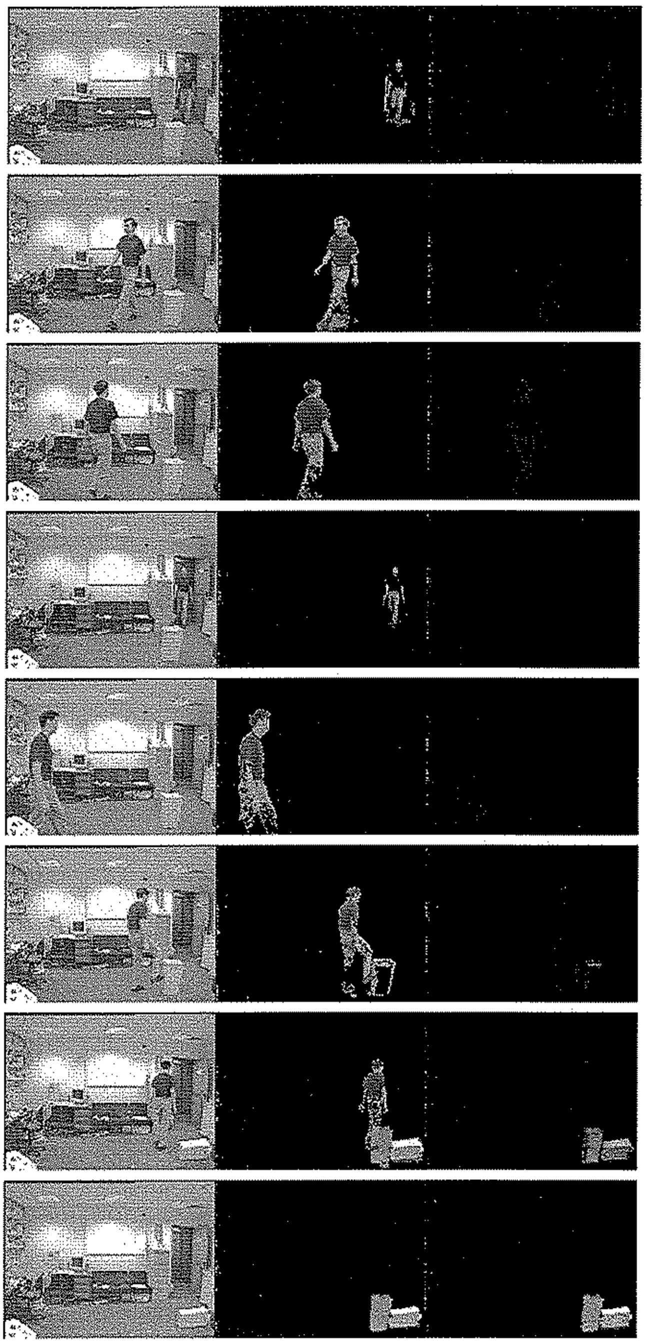

Figure 5: From left to right: Selected frames (a) for comparison of Gaussian Mixture Model change detection (b) and a deviant-motion detector trained to accept the walking person as a normal pattern of change(c).

pressions. In *European Conference on Computer Vision*, pages 909–924, Freiburg, Germany, 1998.

[2] S. Gong, M. Walter, and A. Psarrou. Recognition of temporal structures: Learning prior and propagating observation augmented densities via hidden markov states. In *IEEE International Conference on Computer Vision*, Corfu, Greece, September 1999.

[3] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *European Conference on Computer Vision*, pages 343–356, Cambridge, UK, 1996.

[4] N. Johnson and D.C. Hogg. Learning the distribution of object trajectories for event recognition. *Image and Vision Computing*, 14(8):609–615, 1996.

[5] S.J. McKenna, S. Jabri, Z. Duric, and H. Wechsler. Tracking interacting people. In *IEEE Int. Conf. on Face & Gesture Recognition*, pages 348–353, Grenobles, France, March 2000.

[6] S.J. McKenna, S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, October 2000.

[7] A.V. Oppenheim and R.W. Schafer. *Digital Signal Processing*. Prentice-Hall, 1975.

[8] Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using colour. In *IEEE Int. Conf. on Face and Gesture Recognition*, pages 228–233, Nara, Japan, 1998.

[9] C. Stauffer and W.E.L. Grimson. Using adaptive tracking to classify and monitor activities in a site. In *CVPR*, pages 22–29, Los Alamitos, USA, 1998.

[10] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, pages 246–252, Colorado, USA, June 1999.

[11] M. Walter, A. Psarrou, and S. Gong. An incremental approach towards automatic model acquisition for human gesture recognition. In *IEEE Workshop on Human Motion*, December 2000.