Learning Pixel-Wise Signal Energy for Understanding Semantics

Jeffrey Ng and Shaogang Gong
Department of Computer Science
Queen Mary, University of London
London E1 4NS, UK
{jeffng,sgg}@dcs.qmw.ac.uk

Abstract

Visual interpretation of events requires both an appropriate representation of change occurring in the scene and the application of semantics for differentiating between different types of change. Conventional approaches for tracking objects and modelling object dynamics make use of either temporal region-correlation or pre-learnt shape or appearance models. We propose a new pixel-level approach for learning the temporal characteristics of change at individual pixels. Gaussian Mixture Models are used to model slow long-term changes in pixel distributions while pixel energy histories are used to extract fast-change signatures from short-term events and modelled by CONDENSATION matching.

1 Introduction

In visual surveillance, automated systems are confronted with environments under constant change. For such dynamic scenes, visual change is not necessarily an indication of the occurrence of problematic events but rather, is a function of the context (semantics) in the scene. Deviations from 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.

Previous works addressed the problem of scene-interpretation by explicitly modelling change in terms of the dynamics of moving objects. Object detection and tracking have been performed by numerous methods such as colour object models [11] and background subtraction [12, 9], while the trajectories of moving objects have been modelled using Kalman filters [12] and augmented Hidden Markov densities [8]. However, appearance models are difficult to obtain in unconstrained environments such as shopping malls. And the application of region-growing techniques on collections of pixels obtained from background subtraction artificially induces spatial correlations which complicates the disambiguation process for object or group based trajectory event recognition in free-flowing group-based behaviours.

It is argued here that temporal information contained in the colour signal of individual pixels constitutes a more attractive alternative for understanding events

than spatial connectivity or proximity. Pixel signal energy, computed from the local colour history of the pixel, provides a condensed temporal measure of change. Although the latter is related to computing visual motion such as optical flow from motion-energy filters [6], we are not interested in establishing correspondence in local pixel neighbourhoods. Rather, we are only interested in extracting reliable temporal change at individual pixels. We then consider that meaningful events rather than simply motion in the image sequence should be modelled through understanding the energy history of each individual pixel. This is to some extent reminiscent of the notion of 'topic spotting' in speech recognition, i.e. extracting meaning without explicit modelling the details. Furthermore, pixel energy constitutes a good measure for exploiting synchrony in pixel-events, which have been extensively researched in the psycho-physical literature[13]. Synchronous recognition of pixel-events addresses the limitation of the short-sighted view of single pixels in the scene and provides a more flexible framework for understanding global events as opposed to related spatio-temporal motion-energy measures [1].

To cope with different types of pixel-change, we propose a two-stage scheme. In Section 2, we make use of adaptive Gaussian mixtures (GMM) for modelling long-term colour distributions of pixels, especially slow change caused by lighting cycles. While GMMs provide the platform for long-duration scene analysis, they have also been probabilistically formulated for detecting faster short-term change to perform more computation intensive synchronous energy-history recognition. A novel approach is proposed in Section 3 involving energy histories of pixel change for CONDENSATION-based recognition. Finally, experimental results are provided in Section 4 to investigate the relationship between low-level energy information and high-level semantics for understanding scene events. The technique is also compared to traditional GMM background modelling.

2 Detecting Change

Dynamic scenes exhibit a wide spectrum of change both in terms of the speed and nature of the change occurring in individual pixels. While fast short-term change is mostly characterised by its temporal profile, long-term change is slower in nature and affects pixel's colour distribution. Whether these components are generated by scintillating static objects in the background or cyclically moving objects, they can be modelled by Gaussian Mixture Models. More specifically, given a stream of colour values for a given pixel, $\mathbf{x}_t \in \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_l\}$, the variation in the (r, g, b) components of $\mathbf{x}_t$ can be described in terms of Gaussian means μ and covariances Σ. Illumination specularities or swaying objects such as plants induce multiple modes into the colour distributions of pixels [11, 12], which require multiple Gaussian components. A Gaussian mixture model $p(\mathbf{x}) = \sum_{i=1}^{k} \omega_i \cdot \psi(\mathbf{x}, \mu_i, \Sigma_i)$ can be used, 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 of objects 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 [12]. New visual evidence is approximated with uniform Gaussian

clusters of pre-set variance according to the amount of noise present in the particular capturing setup. Offline methods such as k-means clustering or EM are not fast enough for computing thousands of mixtures for separate pixels. The clusters are then adapted to the particular distribution of the pixels.

For a new pixel $\mathbf{x}_t$, the closest Gaussian with Mahalanobis distance smaller than 2 s.d.($\sim$ 98% confidence) is selected as responsible. A learning rate α is used to constrain the pace of change of the means and covariance of the Gaussian as well as the mixture parameter ω of all the Gaussians to promote long-term changes of the distribution over short-term variation,

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

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

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

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

A confidence factor T is used to identify pre-dominant components in the distribution and is expressed in terms of the overall ratio (0-1) of predominant clusters. The Gaussian components in the mixture are ordered according to the product of (a) their weights, which reflect observation frequency 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 observations are considered as predominant.

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

New clusters are generated for observations $\mathbf{x}_t$ which do not fit current clusters. Once a limit k_{max} is exceeded, the weakest, less important, cluster is replaced for computational reasons. In our case, we use k_{max}=6 so that the mixture model mostly captures static components responsible for slow change.

We then use Bayes' rules, instead of simple Mahalanobis distance as in [12], to formulate the probability of pixel values $\mathbf{x}_t$ belonging to a pre-learnt set of long-term Gaussian clusters as opposed to recent foreground components. The prior is readily available in the form of the mixture parameters.

$$P(\text{long-term}|\mathbf{x}_t) = \frac{p(\mathbf{x}_t|\text{long-term})P(\text{long-term})}{p(\mathbf{x}_t)} \quad (6)$$

where

$$p(\mathbf{x}_t|\text{long-term}) = \sum_{i=1}^{b} \frac{1}{2\pi^{3/2}|\Sigma_t^i|^{3/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_t - \mu_t^i)^{\mathrm{T}}(\Sigma_t^i)^{-1}(\mathbf{x}_t - \mu_t^i)\right) \quad (7)$$

The configuration of the predominant set stores the accumulated history of the observation frequency of each component in the mixture over a long time scale. The state of the set can therefore capture slow changes in the colour distribution of pixels. Depending on the surveillance task, the predominant set can be locked so that new clusters are reported as abnormal, e.g. the introduction of a parcel in a busy scene. In the short term, the long-term models can detect non-fitting fast change which are subsequently modelled with energy histories in the next section.

3 Recognising Meaningful Change

Rapidly changing visual phenomena exhibited by the motion of animated objects typically involve both non-rigid deformations [9] and purposeful trajectories [5, 12, 8]. Illumination specularities further complicate the task of understanding scenes from purely visual data. Without higher-level knowledge provided in the form of pre-learnt object and trajectory models, it is very difficult to interpret frame-wise data. Indeed, semantics used for understanding scenes and classifying events operate on object-level information which is not readily available in low-level pixel data. However, the temporal sequence of change in pixel data can provide a better cue as to the type of event occurring at the pixel's location. Pixel energy extracts the signature of change occurring at any time instant. Furthermore, we propose that histories, or temporal sequences, of pixel energy provide a generic means of extracting signatures from short-term visual change. Figure 1 shows pixel energy collected from a sample sequence of a person moving from left to right and back.

Figure 1: Rows from top to bottom: (a) Selected frames from a left-right-left walking sequence. (b) Pixel-energy data is encoded in grey-level using a log-scale to show small scale structures. (Black indicates high response). Reflective edges can be seen with small responses.

Pixel energy Pe can be measured from the response of pairs of quadrature filters of temporal width v[10] and filter cut-off $ct = 3.5$,

$$Pe(\mathbf{x}_t) = \left[\sum_{i=o}^{2v} g\left(\frac{ct \times (i-v)}{v}\right)\mathbf{x}_{t-i}\right]^2 + \left[\sum_{i=o}^{2v} h\left(\frac{ct \times (i-v)}{v}\right)\mathbf{x}_{t-i}\right]^2 \tag{8}$$

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

$$g(y) = \eta(2y^2 - 1)e^{-y^2} \tag{9}$$

$$h(y) = \kappa y + \lambda y^3 e^{-y^2} \tag{10}$$

where the normalising coefficients are $\eta = 0.9213$, $\kappa = -2.205$ and $\lambda = 0.9780$ [4].

Most energy-based approaches suffer from scale problems for tuning the temporal width of the filters. Spatio-temporal filters used for computing optical flow require multiple banks at different scales [3]. However, we have found that using a temporal width of 10 frames for our sequences captured at 8Hz, provide acceptable, although not optimal, energy histories for a variety of events. Figure 2 shows typical energy histories extracted from different pixels in another right-left-right

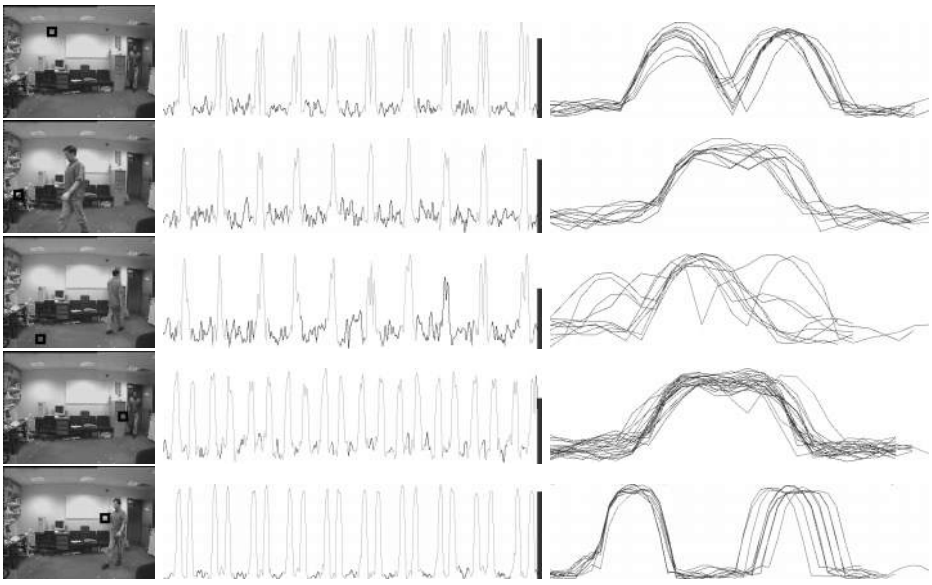

Figure 2: From left to right: (a)Selected frames of the sequence are shown with a black square for selected pixels; (b)Distinctive and repeatable structures from energy histories of the entire sequence for the pixel. Gray parts indicate fast change while black parts indicate slow change, obtained from GMM (akin to background subtraction); (c)Normalised and super-imposed energy-histories of fast-change.

sequence. Pixel-energy information can be seen to possess distinctive repeatable signatures caused by different patterns of change from the view-point of the pixels.

The Gaussian Mixture Models of long-term colour distributions of pixels can be used to detect sections of slow change in the energy signal and therefore allow for segmentation into discrete pixel-events of fast-change. The third column of Figure 2 shows energy histories segmented from the continuous energy signals. The discrete pixel-event energy signatures obtained for fixed pixels from a training sequence of "normal" activities are used as models for classifying new activities as normal(known) or abnormal(unknown). Essentially, semantics are being tied to specific energy histories through supervised learning. Probabilistic trajectory matching provides the mechanism for matching new observations to pre-learnt models [7, 2, 5]. Multiple hypotheses are generated to match a backward window on the signal against template windows in the models. The propagation of random samples allows for concurrent hypotheses to be maintained while providing temporal and amplitude scaling for signal-matching cross-correlation flexibility.

More precisely, the matching hypotheses or states are defined as $(\mu, \phi, \alpha, \rho)$ where μ is the model (pixel-event energy history) being matched, ϕ, the position of the correlation window in the model, α and ρ are the amplitude and temporal scaling parameters respectively. A finite set of k states are then propagated across

time according to a cross-correlation observation probability as defined in [2],

$$P(\mathbf{y}_t|\mathbf{s}_t) = \exp\left\{ -\sum_{j=0}^{w-1} \frac{(\mathbf{y}_{t-j} - \alpha m^{\mu}_{(\phi-\rho j)})^2}{2\sigma_{\mu}(w-1)} \right\} \tag{11}$$

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 a threshold of 30% confidence) are propagated to the next time step according to,

$$\mu_t = \mu_{t-1} \tag{12}$$

$$\phi_t = \phi_{t-1} + \rho_{t-1} + N \tag{13}$$

$$\alpha_t = \alpha_{t-1} + N \tag{14}$$

$$\rho_t = \rho_{t-1} + N \tag{15}$$

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

The propagative dynamics of the cross-correlation windows involve predicting pixel-energy values for the next time step from previously learnt energy models. Hypotheses therefore track a correlation feature space for single pixels, matching new signals against learnt models. However, recognising energy histories for single pixels can be prone to noise and can also suffer from ambiguities arising from spatio-temporal interference effects by textured surfaces. Without resorting to full spatio-temporal motion-energy filters, the effect of synchrony in visual information can be exploited. Global events affect multiple pixels simultaneously and irrespective of the type of textured change occurring at the object level, the pixel-energy information of the involved stream of pixels should exhibit strongly correlated change. Preserving a common time reference for each learnt energy history allows for synchronous cross-propagation of correlation matching hypotheses across pixels. A percentage of the states are reserved for random initialisation and cross-propagation. The propagative dynamics of the samples are upgraded as:

- Given a pixel $\mathbf{x}_t$ at time t, for all the samples $(\mu_t, \phi_t, \alpha_t, \rho_t)$ satisfying the matching confidence threshold, another pixel $\mathbf{y}_t$ with energy-history model $\mu_t^{'}$ and model time index $\phi_t^{'}$ corresponding to ϕ_t is selected. A new sample $(\mu_t^{'}, \phi_t^{'} + \rho_t, \alpha_t, \rho t)$ is cross-propagated into pixel $\mathbf{y}_t$ at the next time step with similar amplitude and temporal scaling factors as the original sample.

The probability of the change in a given pixel at time t matching the pre-learnt normal models is given as the best observation probability over a set of k states,

$$P(\mathbf{y}_t) = max_{i=1}^{k}(P(\mathbf{y}_t|\mathbf{s}_{i,t})) \tag{16}$$

While recognising patterns of pixel change in a new sequence, the technique generates hypotheses of normal(known) energy histories matching energy data from the sequence. Good hypotheses generate cross-hypotheses in other pixels which were synchronously involved in similar fast change during the learning stage. Events can therefore sustain adequate recognition by pixels cross-propagating hypotheses to each other and back. The technique provides a good alternative for learning the binding process of pixels into events without object representations.

4 Experiments

To illustrate how semantics can be incorporated into temporal models based on energy histories and how the learnt models can be used to detect unknown deviant events, we give some preliminary results. The system is trained on a sequence of approximately 1700 frames (from 20 repeated events) containing two people carrying out their normal routine of entering the office from the door on the right, moving to the left for an inspection and leaving by the same door, as in Figure 3.

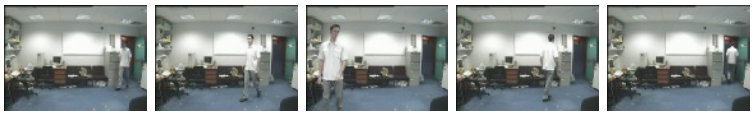

Figure 3: Selected frames from the training sequence.

After training, the system was tested on five sequences of activities performed by three persons, one of whom was not present during training. The testing sequences contain similar events to the training sequences but with differences in the characteristics of performed movement so as to render either part of or the whole activity "abnormal". First, the test subject repeated the movement at (a) slower and (b) faster speeds. A stationary pause (c) and a quick jump (d) were introduced in the middle of the right-left movement. Finally, 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 (e) is not considered as normal.

Table 1: Abnormal event detection results for the test sequences totalling over 1700 frames. This is based on the deviant-event model's (DEM) and GMM model's ability to correctly classify events containing normal and abnormal motion.

Test Events	No. of Event Occurrences	No. of Frames	DEM		GMM	
			Detected	%	Detected	%
Slow Movement	6	615	6	100	6	100
Fast Movement	6	255	6	100	6	100
Stationary pause	6	362	5	83.3	0	0
Jump	6	356	4	66.7	0	0
Falling Box	1	108	1	100	0	0

Table 1 shows the results of deviant-event recognition over the five test categories. For the "Slow Movement" and "Fast Movement" sequences, the deviant parts of the events have been successfully detected as shown in Fig 4. The deviant-event model (DEM) perform better in sequences which involve semantically meaningful deviations from pre-learnt patterns of change, such as "Stationary pause" and "Jump" where only the deviant parts of the events are detected as shown in Fig 5 and Fig 6. As the Gaussian mixture models do not possess any knowledge

of context, they detect all movement as abnormal events. In the "Falling box" sequence, the movement of the person is considered as normal in the particular context of that office environment as the DEM model correctly matches the per-pixel change occurring in the frames with its pre-learnt patterns. Both the DEM model and the Gaussian mixture model do detect the event when the box falls.

Figure 4: Detection results for the "Slow Movement" event. From top row to bottom: Original images, GMM model detection and Deviant-event model detection.

Figure 5: Results from the "Stationary Pause" event.

The results show that general semantics concerning the type of change in individual pixels can be used to differentiate between different classes of events and indeed selectively identify locations and time in the scene where unknown deviant change occurred. Such ability provides additional flexibility over Gaussian Mixture Models for monitoring complex events in dynamic environments.

5 Conclusion and Future Work

Modelling behaviours and recognising events often require object-level representations to interpret visual data. However, object segmentation and trajectory extraction relies on spatial proximity (region-growing) and temporally constrained correlations. However, using such assumptions in busy scenes might not be sufficient. We propose a new low-level representation which can be linked to semantics

Figure 6: Detection results from the "Jump" event.

to understand events and perform abnormal event detection without making any assumptions about objects. Energy-histories provide a condensed variable-length representation of fast temporal change in single pixels. Preliminary results show that they can be used to semantically discriminate between events involving different pace as well as patterns of change. We have also used Gaussian Mixture Models to separately model and recognise slow change such as illumination cycles under a less computationally taxing framework. The ambiguity inherent in viewing a complex world through a single pixel has been addressed by using synchronous change in multiple pixels during events to perform pixel-stream hypotheses.

Although we have used supervised learning to introduce semantics in a low-level framework, unsupervised learning can be used to extract common patterns of change over long periods of time. Currently, the synchronous prediction and recognition of activity in streams of pixels have not been fully investigated. Future work will concentrate on the propagative dynamics for generating hypotheses and augmenting the technique to discriminate more subtle deviant-event recognition.

References

[1] E.H. Adelson and J.R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*. MIT Press, 1991.

[2] M.J. Black and A.D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In *ECCV*, pages 909–924, Freiburg, 1998.

[3] O. Chomat and J. Crowley. A probabilistic sensor for the perception and the recognition of activities. In *ECCV*, pages 487–503, Dublin, 2000.

[4] W.T. Freeman and E.H Adelson. The design and use of steerable filters. *IEEE PAMI*, 13(9):891–906, 1991.

[5] S. Gong, M. Walter, and A. Psarrou. Recognition of temporal structures: Learning prior and propagating observation augmented densities via hidden markov states. In *ICCV*, pages 157–162, Corfu, 1999.

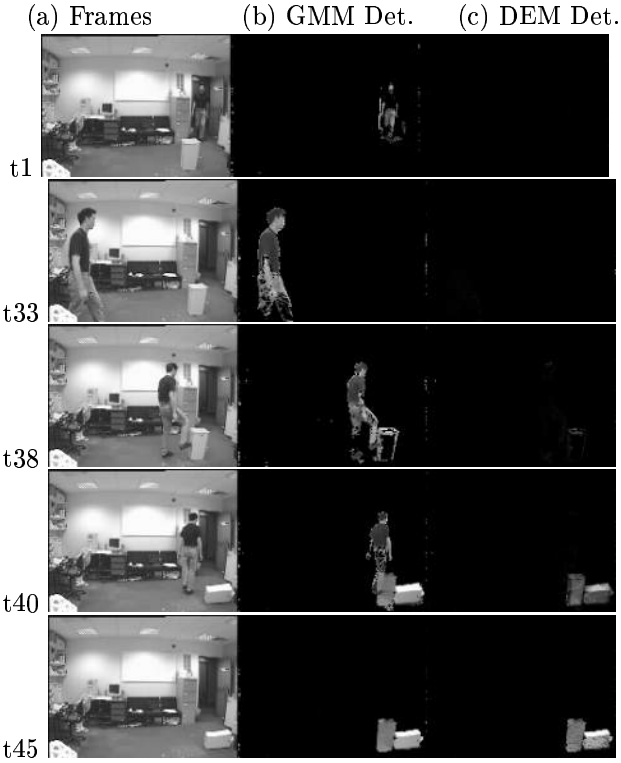

Figure 7: Comparison of detections by a GM-based dynamic model (GMM) and our deviant-event model (DEM).

[6] D.J. Heeger. Optical flow using spatiotemporal filters. *IJCV*, 1:279–302, 1987.

[7] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *ECCV*, pages 343–356, Cambridge, 1996.

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

[9] S.J. McKenna, S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld. Tracking groups of people. *CVIU*, 80(1):42–56, 2000.

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

[11] Y. Raja, S.J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using colour. In *FG*, pages 228–233, Nara, 1998.

[12] 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.

[13] M. Usher and N. Donnelly. Visual synchrony affects binding and segmentation in perception. *Letter to Nature*, 394:179–182, 1998.