Anomalous Behaviour Detection Using Spatiotemporal Oriented Energies, Subset Inclusion Histogram Comparison and Event-Driven Processing

Andrei Zaharescu[1,2] and Richard Wildes[2]

[1] Aimetis Corporation, Waterloo, Canada
[2] Department of Computer Science and Engineering,
York University, Toronto, Canada
{andreiz,wildes}@cse.yorku.ca

Abstract. This paper proposes a novel approach to anomalous behaviour detection in video. The approach is comprised of three key components. First, distributions of spatiotemporal oriented energy are used to model behaviour. This representation can capture a wide range of naturally occurring visual spacetime patterns and has not previously been applied to anomaly detection. Second, a novel method is proposed for comparing an automatically acquired model of normal behaviour with new observations. The method accounts for situations when only a subset of the model is present in the new observation, as when multiple activities are acceptable in a region yet only one is likely to be encountered at any given instant. Third, event driven processing is employed to automatically mark portions of the video stream that are most likely to contain deviations from the expected and thereby focus computational efforts. The approach has been implemented with real-time performance. Quantitative and qualitative empirical evaluation on a challenging set of natural image videos demonstrates the approach's superior performance relative to various alternatives.

1 Introduction

Detection of anomalous behaviour relative to some model of expected behaviour is a fundamental task in surveillance scenarios. Examples include detection of movement in an area where none should occur (as in a secure storage facility) and detection of "wrong way motion" where movement of objects only should occur in one direction yet are observed in a different direction (as in movement of traffic on a one-way road). In particular, given the increase in video coverage of public and private spaces, an automated ability to monitor the acquired data and signal deviations from expected behaviour would be very useful, as it could serve to alert either human or artificial systems to analyze further the data that is acquired.

A number of challenges must be surmounted for successful detection of anomalous behaviour in surveillance video. In essence, these challenges arise from the need to model a wide range of potentially complicated patterns of normal activity and detect fine deviations from that model, even while being robust to changes that are insignificant. Normal activity can range from simple no temporal change through single and

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part I, LNCS 6311, pp. 563–576, 2010.

multiple motions to complicated situations of dynamic textures (e.g., backgrounds of fluttering vegetation or water waves), including multimodal behaviour. Modeling must be flexible to encompass this entire range. Anomaly detection must be able to register subtle changes of interest (e.g., changes in direction or speed of motion, presence of a coherently moving object against a camouflaging background of texture and dynamic clutter), while not signaling insignificant changes (e.g., naturally occurring illumination changes during the diurnal cycle, differences between objects that are manifest purely in terms of spatial appearance without behaviour differences). It also is desirable to allow for partial matches of observations to the model, as complicated scenarios might encompass multimodal behaviour and observations that correspond to any modeled mode are acceptable, while alternatives are not. Further, in many situations an ability to incorporate deviations that recur over time into the model is desirable, so that they are no longer considered anomalous.

Related Work. One general class of approach to anomaly detection in video is based on explicit tracking of viewed objects [22,31,9,18,5]. Such approaches acquire models of typical trajectories from tracker output over some training period and subsequently signal deviations in observed tracks as anomalies. A significant limitation to this class of approaches is their reliance on (visual) tracking, a still unsolved challenge.

Background subtraction techniques that model typical appearance from a camera view can be applied to detecting behaviour anomalies (see, [27] for review and, e.g., [14,20,36,17] for a sampling of more recent work). The simplest techniques involve unimodal background models of pixelwise image intensity and have limited applicability for complicated backgrounds. Increased sophistication in modeling static background appearance comes through consideration of pixel attributes beyond image intensity (e.g., gradients, edges, texture). More involved techniques account for dynamic backgrounds by acknowledging multimodal intensity distributions, parametric modeling, kernel-based estimation and predictive filtering. An extension of predominantly intensity-based background modeling for video operates by indexing observations relative to a database of normal videos, with failures taken as anomalies [8]. A limitation of appearance-based approaches is their inability to abstract purely dynamic aspects of behaviour, which can lead to overly restrictive, under-generalized models of normal behaviour (e.g. lack of invariance to different actors performing the same activity).

More closely related to the approach proposed in the current paper are efforts that have more explicitly modeled the dynamic behaviour of backgrounds. Typically, such approaches make use of some type of spatiotemporal filtering to define normal local activity with anomalies taken as deviations from the defined model. Along these lines, some work has appealed directly to spatiotemporal gradient measurements [30,25]. Other work has been more restricted to considering only the temporal first derivative (blurred and quantized) [35]. Alternatively, image flow measurements have been used to define local activity models [7,4,23,2,24]. Still other work has abstracted local flow measurements to a simpler consideration of whether or not a pixel typically is in motion to define locally normal behaviour [21]. Direct use of spatiotemporal gradients to define normal activity has a number of limitations, including sensitivity to image contrast and spatial pattern, which lead to lack of robustness to changes in illumination and different appearing actors performing the same activity. Further, reliance on temporal derivative alone leads to an inability to distinguish different motion directions. Alternatively, approaches that rely on (local) flow measurements are limited in the complexity of behaviours they can capture, e.g., multiple motions at a point, temporal flicker and dynamic textures (e.g., water, wind-blown foliage) can be difficult to model, as they violate the underlying assumptions of the flow computation (e.g., brightness conservation) and thereby yield unreliable results in such scenarios.

A number of recent approaches are concerned with modeling of non-local behaviour (but typically building on local measurements) with application to anomaly detection [6,25,24,29,28,32,26]. While such approaches make strides in accounting for non-local activity, they still can be limited by overly restrictive local representations, e.g., spatiotemporal gradient models that are not invariant to spatial appearance and flow models that do not account for activity that is amenable to characterization as a single local flow vector (e.g., multiple motions and more general dynamic textures).

To account for complicated local behaviour, measures of spatiotemporal oriented energy play a prominent role in the approach proposed in the current paper. Previously, such measures have been used in a variety of vision processing tasks, including image enhancement and motion estimation [16], video segmentation [12], pattern categorization [34] and activity recognition (although not generic anomaly detection) [10,13,11].

Contributions. In the light of previous research, the present approach makes four main contributions. 1) 3D, (x, y, t), spatiotemporal oriented energy measurements are used to represent observations. While almost any approach to anomalous behaviour detection must employ spatiotemporal filtering of some type, no previous work has made use of the energy filtering framework proposed here, which enjoys a number of benefits in being able to capture a wide range of image dynamics (both standard motion as well as more complex dynamic patterns, e.g., flickering lights, swaying vegetation and water), even while being robust to irrelevant variations (e.g., overall illumination variation and different appearing individuals engaged in the same behaviour). 2) A novel histogram comparison method is presented to detect anomalous behaviour relative to an acquired model. A key component of this measure is that it accounts for partial matches of new observations to the acquired model. 3) Event-driven processing is used to automatically mark portions of the video stream that are most likely to correspond to activities and thereby focus computational efforts. 4) The proposed approach has been realized in real-time implementations. A detailed empirical evaluation of the implementations is presented, which documents the contributions of its individual components and its strong overall performance relative to alternative approaches.

2 Technical Approach

The developed approach to detecting anomalous behaviour is based on observed deviations from an acquired model of normal behaviour. The model is image-based and thereby indicates expected (normal) observations on a pixelwise basis as recorded from a specific viewpoint.

2.1 Spatiotemporal Energy Representation

In the developed approach, both model and newly acquired video observations are represented in terms of local distributions of 3D, (x, y, t), spatiotemporal oriented energy as derived from input imagery via application of an orientation tuned filter bank. This representation is selected as it captures the local first-order correlation structure of visual spacetime and thereby allows a wide range of dynamic activities to be captured (e.g., both single and multiple motions as well as more general dynamic textures) with robustness to illumination and purely spatial appearance [12]. In particular, the current approach to spatiotemporal orientation for anomaly detection follows closely the previous work [12], where it was used instead for video segmentation.

To extract the orientation measurements, oriented energy filtering is realized in terms of second derivative of 3D Gaussian filters, $G_{2_\theta}(x, y, t)$, and their Hilbert transforms, $H_{2_\theta}(x, y, t)$, where θ represents the direction of the filter's axis of symmetry. These particular filters are selected due to their (moderately) broad tuning, which allows for a wide range of orientations to be captured with a relatively small number of filters. Additionally, these filters admit a steerable and separable formulation [15], which leads to efficient computations. The filters are taken in quadrature, to yield the following local oriented energy measure,

$$E_\theta(x, y, t) = (G_{2_\theta} * I)^2 + (H_{2_\theta} * I)^2, \tag{1}$$

where $I \equiv I(x, y, t)$ denotes the input imagery and $*$ symbolizes convolution.

For the case of dynamic spacetime orientation (e.g., as related to motion phenomena), each of the oriented energy measurements, (1), is confounded with spatial orientation. Correspondingly, the same pattern of activity will yield different responses across an ensemble of oriented energy filters depending on variations in the spatial appearance of the viewed object/event: This is an undesirable state of affairs for dynamic anomaly detection as it would not be possible to build models of normal behaviour that are robust to irrelevant details of purely spatial appearance (e.g., sensitivity to what people are wearing, when the concern is for how they are moving). To remove this difficulty, the spatial orientation component of the oriented energy responses is discounted by marginalizing this attribute via pointwise, linear combination of energy measures, (1), that support a single spacetime orientation, as specified by the unit normal, $\mathbf{n}$, corresponding to its frequency domain plane. (Recall that a pattern exhibiting a single spacetime orientation, e.g., velocity, manifests as a plane through the origin in the frequency domain [33].) In particular, the energy measure, (1), is refined to become

$$\tilde{E}_{\mathbf{n}}(x, y, t) = \sum_{i=0}^{N} E_{\theta_i}(x, y, t), \tag{2}$$

where θ_i represents one of $N + 1$ equal spaced orientation tunings consistent with direction $\mathbf{n}$ and $N = 2$ is the order of the Gaussian derivative filter (1), for details see [12].

The resulting oriented energies are confounded with local contrast. This makes it impossible to determine whether a high response from a particular filter is indicative of a close match with the underlying structure or is instead a low match that yields a high response due to significant contrast in the signal. To arrive at a purer measure of oriented spacetime structure, the energy measures are normalized by the sum of the oriented responses at each point,

$$\hat{E}_{\mathbf{n}_i}(x, y, t) = \frac{\tilde{E}_{\mathbf{n}_i}(x, y, t)}{\sum\limits_{\mathbf{n}_j \in \mathcal{S}} \tilde{E}_{\mathbf{n}_j}(x, y, t) + \epsilon}, \tag{3}$$

where $\mathcal{S}$ denotes the set of (marginalized) oriented energies, (2), with $\mathbf{n}_j$ a particular sample and ϵ a constant, set to 1% of the maximum filter response, introduced as both a noise floor and to avoid instabilities at points where the overall energy is small.

In the currently implemented representation, $K = 6$ different directions, $\mathbf{n}$, are made explicit, that correspond to leftward, rightward, upward and downward motion (each with peak response at 1 pixel/frame movement), static (orientation orthogonal to the image plane) and flicker (orientation orthogonal to the temporal axis); although, due to the broad tuning of the filters employed, responses arise to a wide range of orientations about the peak tunings. By construction, these measures are marginalized for purely spatial appearance and normalized for contrast, which allows for a degree of robustness to unimportant variability in observations. Further, the representation is simply realized by an alternating series of linear (i.e., separable convolution and pointwise addition) and pointwise non-linear operations (i.e., squaring and division); thus, efficient computations are realized.

Finally, it is straightforward to extend the described approach to multiple scales. In particular, the input imagery is brought under a pyramid representation [19] prior to filtering. Subsequently, the oriented filtering, (1), appearance marginalization, (2), and normalization, (3), are performed separately at each pyramid level to realize a multi-scale oriented energy representation. In the current implementation $\sigma = 5$ scales are employed, with factor of $\sqrt{2}$ subsampling between levels and commensurate lowpass filtering prior to subsampling.

2.2 Model Acquisition and Maintenance

The proposed model is given in terms of a histogram of spatiotemporal orientations observed over some period of time. Since behaviours of interest are (by definition) dynamic, only measures of orientation that arise from non-static observations are explicitly represented in the model. In particular, a key component to the method is the concept of accumulating statistics only on interesting events: The information is aggregated at the pixel level only between frames containing dynamic energy. Dynamic energy is captured in terms of a threshold, β, on the static channel E_{Static}: If static energy is greater than β, it is considered that there is no activity at the current pixel. To formalize the notion of event-driven processing, let

$$\psi(x, y, t) = \begin{cases} 1 \text{ if } E_{Static} < \beta \\ 0 \text{ otherwise.} \end{cases} \tag{4}$$

and the model histogram, $\mathbf{m}(x, y)$, be defined as

$$m_{\mathbf{n}}(x, y) = C \sum_{t=1}^{t=T} \psi(x, y, t) \hat{E}_{\mathbf{n}}(x, y, t) \tag{5}$$

where $m_{\mathbf{n}}(x, y)$ is the histogram bin corresponding to orientation $\mathbf{n}$ at location (x, y), C is a normalization factor ensuring the histogram sums to unity and t indexes from an initial to frame T used in building the model. The histogram at a given spatial location over a period of time is built by concatenating the relative energy of the K spacetime orientations, $\mathbf{n}$, at each of the σ scales, thus leading to a $K \times \sigma$ bin histogram.

Similarly, a new observation is made by constructing a histogram, $\mathbf{o}(x, y)$, analogous to the model, except that it is accumulated only over a relatively small number of frames. In particular,

$$o_{\mathbf{n}}(x, y) = C \sum_{t=t_0-\lfloor(k/2)\rfloor}^{t_0+\lfloor(k/2)\rfloor} \psi(x, y, t)\hat{E}_{\mathbf{n}}(x, y, t) \tag{6}$$

where $o_{\mathbf{n}}(x, y)$ is the histogram bin corresponding to spatiotemporal orientation $\mathbf{n}$ at location (x, y), C is a normalization factor ensuring the histogram sums to unity and t indexes across k frames, $k << T$, that are used in accumulating the current observation at time $t = t_0$.

Finally, the model $\mathbf{m}^t(x, y)$ at time t is updated in an ongoing fashion so as to account for the current observation, $\mathbf{o}^t(x, y)$, according to

$$\mathbf{m}^{t+1}(x, y) = [1 - \delta\psi(x, y, t)]\mathbf{m}^t(x, y) + \delta\psi(x, y, t)\mathbf{o}^t(x, y) \tag{7}$$

on a bin-by-bin basis with δ controlling the update rate. Notice that update is only performed when there is an event $\psi(x, y, t)$, (4).

2.3 Comparison of Model and New Observations

Given a model, $\mathbf{m}(x, y)$, and a current observation, $\mathbf{o}(x, y)$, anomalous behaviour is defined in terms of deviations of the observation from the model. Given that both the model and observation are captured as histograms, various standard comparison methods might be invoked (e.g., χ^2 test of independence or Bhattacharyya coefficient). However, such standard methods fail to capture two key points of relevance for anomaly detection. First, the observation might only encompass a subset of modeled activity: This easily can be the case, due to the fact that the model statistics typically are accumulated over a relatively large number of frames, possibly incorporating multiple activities (e.g., left and right motions); whereas, the observation statistics capture relatively shorter time periods that might not encompass all modeled activities (e.g., left motion only). Second, it is desirable to model scenarios where a lack of activity in the current observation histogram should not be considered anomalous, even if the previously acquired model for that particular pixel differs significantly.

To address the noted points, the χ^2 test of independence [3] is taken as a point of departure and modified, as follows. In the current context, the χ^2 measure between $\mathbf{m}(x, y)$ and $\mathbf{o}(x, y)$ is given as

$$\chi^2[\mathbf{m}(x, y), \mathbf{o}(x, y)] = \sum_{\mathbf{n}\in\mathcal{S}} \frac{(m_{\mathbf{n}}(x, y) - o_{\mathbf{n}}(x, y))^2}{m_{\mathbf{n}}(x, y) + o_{\mathbf{n}}(x, y)}. \tag{8}$$

The first point, regarding any particular current observation not encompassing all possibilities captured in the model, can be addressed by introducing a notion of subset inclusion, i.e., the observed behaviour must be a subset of the modeled behaviour; else, it will be taken as anomalous. To indicate such anomalies, a function is needed that selects particular orientations (histogram bins), $\mathbf{n}$, where there is little response in the model (e.g., relative to some threshold, τ_0) even while there is significant response in the observation (e.g., relative to some threshold, τ_1); a corresponding function can be defined as

$$\phi(m_{\mathbf{n}}, o_{\mathbf{n}}) = \begin{cases} 1, \text{ if } (m_{\mathbf{n}} < \tau_0) \text{ and } (o_{\mathbf{n}} - m_{\mathbf{n}} > \tau_1) \\ 0, \text{ otherwise} \end{cases} \tag{9}$$

The second point, regarding a particular observation not encompassing any activity, can be addressed by assigning decreasing weight to an observation as fewer of the frames observed in its construction drive event-based processing as indicated by (4). This notion can be captured by an event ratio, $\rho[\mathbf{o}(x, y)]$, of the number of frames that contributed to the event-based processing, $\gamma[\mathbf{o}(x, y)]$, to the total number of frames observed, $\alpha[\mathbf{o}(x, y)]$, i.e.,

$$\rho[\mathbf{o}(x, y)] = \frac{\gamma[\mathbf{o}(x, y)]}{\alpha[\mathbf{o}(x, y)]}. \tag{10}$$

Combining the original χ^2 formulation, (8), with the formalization of subset inclusion, (9), and event ratio, (10) yields the final measure of distance between a model and observation

$$D[\mathbf{m}(x, y), \mathbf{o}(x, y)] = \rho(\mathbf{o}) \sum_{\mathbf{n} \in \mathcal{S}} \phi(m_{\mathbf{n}}, o_{\mathbf{n}}) \frac{(m_{\mathbf{n}} - o_{\mathbf{n}})^2}{m_{\mathbf{n}} + o_{\mathbf{n}}}, \tag{11}$$

with larger distances taken as increased evidence for behaviour anomaly at (x, y) and final anomaly detection based on a comparison to a threshold, Δ. (Explicit reference to image coordinates, (x, y), is suppressed on the right-hand side of the final distance measure, (11), for the sake of notational compactness.)

3 Empirical Evaluation

Three implementations of the proposed approach to detecting anomalous behaviour have been developed, which differ according to their software and hardware utilization and are documented in Table 1. Algorithmic parameters are the same for all implementations: $\beta = 0.35, \delta = 0.005, \tau_0 = 1.5/h, \tau_1 = 0.15/h$, where h is the number of histogram bins, i.e., $h = 6$ (orientations) $\times 5$ (scales) $= 30$, unless otherwise noted. The reported timings are with respect to processing an image of size 160×120 and attest to the applicability of the approach to real world operational scenarios. Detection results reported below are with respect to the naive ANSI C implementation; although, all implementations yield similar results.

The implementations have been evaluated on a test suite of video sequences, which are documented in Figs. 1 and 2; actual videos are provided in the supplemental material. All sequences are of spatial dimensions 320×240. Each sequence was manually groundtruthed for anomalous behaviours relative to the depicted backgrounds. For the sake of practicality, images were groundtruthed on a coarse spatial grid of cells, shown

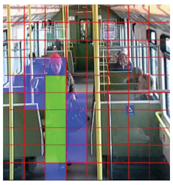
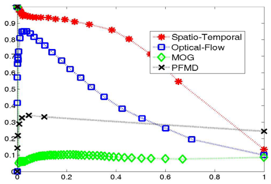
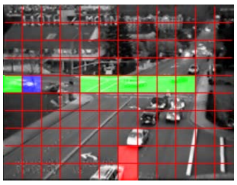
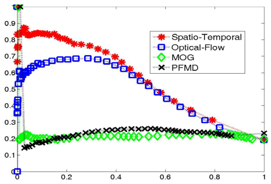
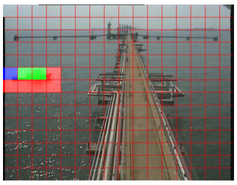
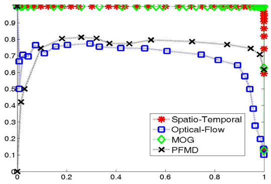
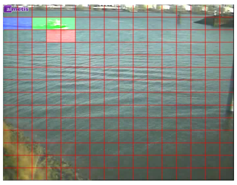
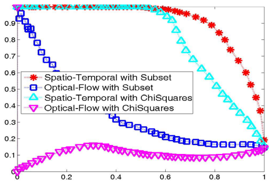
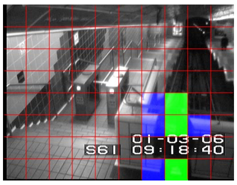
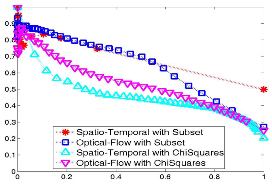
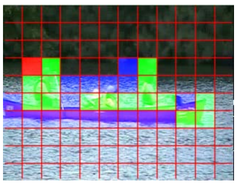
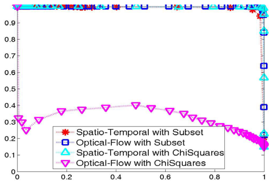
Sample	Recall / Precision	Info
Representation Comparison		
		**Title**: *Train* **Source**: [1] **Desc**: Very challenging train sequence due to drastically varying lighting conditions and camera jitter. Abnormalities: People movement. **Total Frames**: 19218; **Training**: 800
		**Title**: *Belleview* **Source**: [1] **Desc**: Cars moving through an intersection. Model construction during day; testing continuing through night. Abnormalities: Cars entering thoroughfare from left or right. **Total Frames**: 2918; **Training**: 200
		**Title**: *Boat-Sea* **Source**: [1] **Desc**: A sea-boat is passing by (motion on motion). Abnormalities: Boat movement. **Total Frames**: 450; **Training**: 200
Subset Inclusion versus χ^2 Histogram Comparison		
		**Title**: *Boat-River* **Source**: [1] **Desc**: Boat passing by on a river (motion on motion). Abnormalities: Boat movement. **Total Frames**: 250; **Training**: 80
		**Title**: *Subway-Exit* **Source**: [2] **Desc**: Surveillance camera observing pedestrians at a subway exit. Abnormalities: Wrong way motion (leftward and downward). **Total Frames**: 32426; **Training**: 6900
		**Title**: *Canoe* **Source**: [21] **Desc**: A canoe is passing by (motion on motion); also, some wind-blown foliage in background. Abnormalities: Canoe movement. **Total Frames**: 1050; **Training**: 200

Fig. 1. The first column shows a frame during the evaluation of the proposed method, using the manually marked groundtruth information. The Colour coding is: green - true positive; red - false positive; blue - false negative. The second column presents the Precision/Recall curves (abscissa- Recall; ordinate - Precision), with each curve containing 20 measurements. The last column provides additional documentation for each example.

Table 1. Implemented instantiations of the approach for anomalous behaviour detection

Language	Device	Clock	Cores	Time
ANSI C	Intel Core2Duo	2.4GHz	1	80 ms
SSE2	Intel Core2Duo	2.4GHz	1	24 ms
OpenCL	NVIDIA 280GTX	1GHz	120	5 ms

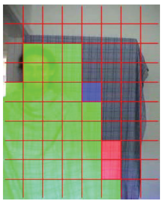
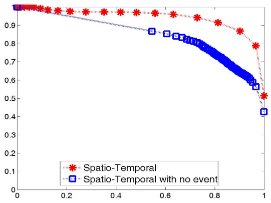
Sample	Recall / Precision	Info
Event versus Non-Event Comparison		
		**Title**: *Camouflage* **Source**: [1] **Desc**: A person in camouflage walking. The right motion is learnt as the normal behaviour. There is a large pause in the middle, to illustrate event based processing. Abnormalities: Left motion. **Total Frames**: 1629; **Training**: 160

Fig. 2. Same formatting as in Figure 1

overlaid on the images. All the videos, the groundtruth data, as well as the groundtruth and the evaluation software are available online [1]. Quantitative evaluation is presented in the form of Precision-Recall (PR) curves by varying the detection threshold, Δ, on (11), where $Recall = \frac{\text{\# True Positives}}{\text{\# Positives in Dataset}}$ and $Precision = \frac{\text{\# True Positives}}{\text{\# True Positives + \# False Positives}}$. In calculating the PR curves, false positive/negative cells adjacent to a true positive cell are discarded.

As detailed in Section 2, the proposed approach to anomaly detection centres around three key ideas: (i) behaviour modeling in terms of a distribution (histogram), (5), of spatiotemporal oriented energy responses, (3), (ii) model and observation comparison via subset inclusion, (11), and (iii) event-based processing, (4). The experiments document how each of these components contribute to the success of the proposed approach.

Experiment 1. The benefits of representation via a **distribution of spatiotemporal oriented energies** are manifested in cases that require robustness to variable illumination and camouflage, even while making fine distinctions between normal and abnormal ac-

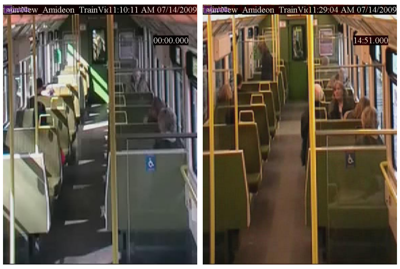

ew Amideon TrainVi11:10:11 AM 07/14/2009
00:00.000
ew Amideon TrainVi11:29:04 AM 07/14/2009
14:51.000

Fig. 3. Example images from *Train* sequence. Extreme background changes are present, as the moving train passes through highly variable exterior lighting conditions.

tivity. A striking example of variable illumination is presented in *Train*, which includes sudden, extreme background changes caused by the moving train passing through tunnels, see Fig. 3 for two dissimilar backgrounds taken shortly apart and the supplemental video. As discussed in Sec. 2.1, the bandpass nature, (1) , and response normalization, (3), of the employed filtering make the representation invariant to additive and multiplicative intensity changes and these properties yield the strong performance in variable illumination shown in Fig. 1. Robustness of the proposed approach to more gradual changes in illumination is illustrated in *Belleview*, as the sequence begins during day and progresses through night. Also of interest in this case is clutter caused by headlights with the onset of dusk.

Spatial camouflage, where novel objects have the same texture patterns as their surround also are not problematic for the proposed approach, as the representation emphasizes distinctions on the basis of dynamics; an example is shown in *Camouflage* where the moving person is covered with the same spatial texture pattern as the background. Dynamic camouflage can come about when normal behaviour is sufficiently erratic to mask novel movement. Representation in terms of a distribution of spatiotemporal orientations allows for such camouflage to be broken, as a wide range of image dynamics can be captured and distinguished: The approach can encompass complicated background dynamics in its model (e.g., motion jitter and rapidly moving shadows/lights in *Train*, and variable waves in *Boat-Sea* and *Canoe*), yet still detect novel moving objects as anomalies (e.g., people, boat and canoe in *Train*, *Boat-Sea* and *Canoe*, resp.). Similarly, since different directions of motion can be distinguished, an observed set of motion directions can be incorporated into the model, while alternative motion directions are marked as anomalous (e.g., wrong-way motion detection of *Belleview*, *Subway* and *Camouflage*).

The benefits of the proposed representation are quantified by the PR curves for *Train*, *Belleview* and *Boat-Sea* in Fig. 1, where a comparison is made to three alternatives. The first is image intensity-based: Capturing behaviour via pixelwise image intensity Mixture of Gaussians (MOG) [27], with a MOG model of normal behaviour acquired during a training period and subsequent intensity observations judged as anomalies based on the joint posterior probability that they belong to any of the modeled modes. The second alternative representation is motion-based: Capturing behaviour via pixelwise Percentage of Fames Motion is Detected (PFMD) [21], with motion detection performed using the opponent spatiotemporal energy magnitude ($|E_{up} - E_{down}|^2 + |E_{left} - E_{right}|^2 > 0.05$), which in preliminary experiments yielded superiour performance to temporal differencing used elsewhere for PFMD modeling [21]. (Notice that opponent spatiotemporal energy magnitude will be relatively large in response to a locally coherent motion [34].) Both of these representations were embedded in the recently proposed *behaviour subtraction* method of anomaly detection [21], as it readily handles both MOG and PFMD models; whereas, the method proposed in the present paper is more specialized for distributed (histogrammed) measurements. The third alternative is based on quantized optical-flow direction and magnitude computed at multiple, 5, scales (e.g., as originally proposed for direction or magnitude [2] and subsequently extended to combine 8 directions with magnitude [24]; the latter is used here, as it was found to provide superiour performance in preliminary experimentation). The quantized optical flow

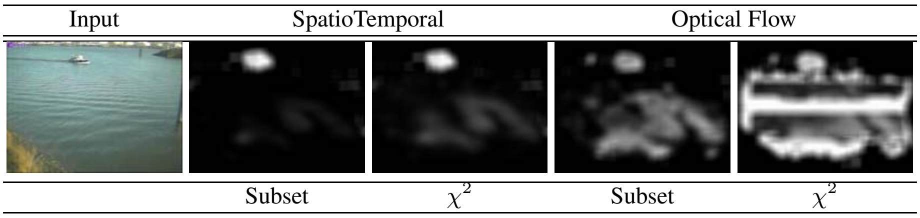


Fig. 4. Comparison of proposed (subset inclusion), (11), vs. χ^2, (8), histogram comparison measures for the Boat-River sequence (frame 161)

defines a histogram [2,24] that is substituted directly into the proposed approach by substituting for oriented energy; thus, a direct comparison is had between oriented energy and optical flow, as all other system components are constant. With one exception, it is seen that the alternative representations yield notably lower PR curves in comparison to the proposed approach, as they are not able to encompass the complicated normal behaviour that is present in the examples. The sole exception is the case of MOG applied to *Boat-Sea* where the appearance of boats (abnormal) are sufficiently different from the acquired mixture that performance is comparable to the spatiotemporal representation. Still, optical flow appears to be second best for the other two cases, *Train* and *Belleview*.

Experiment 2. The main benefit of comparing model, (5), and observation, (6), histograms via **subset inclusion**, (11), is that it allows for partial fits between observations and models. This property is important so that every given observation does not need to encompass the entire range of previously modeled behaviour. To illustrate the practical importance of this consideration, Fig. 4 shows comparative image results of subset-inclusion vs. χ^2 histogram comparison (all other components are exactly the same as those of the proposed method); associated PR curves are shown in Fig. 1. Here, PR curves are shown for both spatiotemporal oriented energy as well as optical flow, as quantized flow can be substituted directly for the energies in the proposed approach (see Exp. 1) to show the benefits of subset inclusion beyond application to energy measurements. Also, flow appeared to be the second best overall performer when comparing representations in Exp. 1. For *Boat-River* and *Subway* using energy as well as flow, it is seen that for a given recall rate, χ^2 has a strong tendency for lower precision relative to subset-inclusion. For *Canoe* spatiotemporal energy already is performing extremely well with just χ^2; however, addition of subset-inclusion allows flow to elevate its level of performance to that of energy. These results are readily explained as χ^2 is not able to accept as normal partial matches to the model; whereas, subset inclusion is with resulting higher precision in its detection, i.e., fewer false positives. The quantitative summaries are supported in the pictorial results, especially for complicated backgrounds (e.g., water in *Boat-River* and water/vegetation in *Canoe*, which encompass a range of motions; whereas, any particular observations show only a subset and such partial matches are reported as anomalies by χ^2, but not by subset-inclusion. Finally, notice that flow leads to similar performance to spatiotemporal oriented energy on *Subway*. This can be accounted for by the fact that both normal and abnormal behaviours

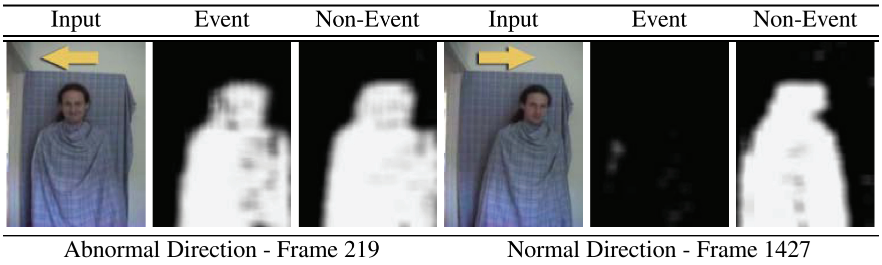


Fig. 5. Comparison of event vs. non-event based update schemes. Without event-based processing, the normal behaviour (right motion) is forgotten after 300 frames of no activity (starting at frame 803) and it is incorrectly detected as abnormal. Event-based processing successfully maintains the model and it does not yield false positives.

(motion of pedestrians) can be captured well by flow (as well as by spatiotemporal oriented energy). Just in this example alone, 10 orientations have been used for spatiotemporal energies by adding 4 directions aligned with motion along diagonals (e.g., up-left, up-right, etc.) to the standard set of 6 (only 4 of which are aligned with motion directions, left, right, up, down), in order to bring its directional discrimination more on par with the optical flow representation, which explicitly encodes motion along diagonals in its histogram bins (as well as left, right, up and down, plus magnitude). Using only 6 orientations for spatiotemporal energy in this example led to performance slightly worse than flow in preliminary experiments, owing to poorer (motion) direction resolution.

Experiment 3. Event-based processing influences construction of models, (5),(7), and observations, (6), to focus computations on portions of the data where behaviour is occurring, as signaled by events, (4). Not only does such processing reduce computational load (e.g., fewer updates are performed), but it also keeps models and observations defined in terms of dynamic behaviour. An interesting benefit of this processing is that it ameliorates problems with forgetting aspects of normal behaviour during model update: Without event-based processing, a modeled event will be discarded from the current model after $1/\delta$ frames by the update, (7). In contrast, by updating only on event frames, the model is prevented from forgetting behaviour due to lack of activity.

Illustrative results are presented in the *Camouflage* example. In this case, after a normal model (rightward motion) is acquired, there is a relatively long period of time when no activity takes place (300 frames); nevertheless, when activity resumes anomalous behaviour still is detected relative to the model acquired prior to the no activity period. The benefit is quantified in the associated PR curve in Fig. 2, which compares the proposed method with the same approach neglecting event-based processing. It is seen that event-based processing yields higher precision at comparable recall for any detection threshold, Δ, as the model is better maintained. Without event-based processing the activity following the period of no activity consistently is misclassified, as shown in Fig. 5; whereas, with event-based processing it consistently is classified correctly. Nevertheless, the approach still allows for the model to encompass newly recurring behaviours (e.g. moving shadows/lights in *Train*), according to the update rule, (7).

4 Discussion

This paper has presented a novel approach to detection of anomalous behaviour in temporal image sequences. The approach centres around three key ideas. First, imagery is represented in terms of distributions of spatiotemporal oriented energy to model normal behaviour as well as record new observations. This representation allows the approach to capture a wide range of naturally occurring behaviours while making fine grained distinctions between model and new observation with robustness to variations in illumination and purely spatial appearance. Second, model and observations are compared via histogram subset inclusion matching. Subset inclusion matching allows for partial matches between model and observation so that not every possible modeled activity must occur at any given time instance to avoid being considered anomalous. Third, event driven processing is employed to allow for focusing of computational effort on portions of the image stream where anomalies might occur. A limitation of the current approach is that it does not explicitly account for non-local phenomena (e.g., interactions between separate local measurements in space and time). Future work will extend the approach to deal with such matters, e.g., by overlaying a MRF on the approach's local observations to abstract interactions.

The entire approach has been instantiated in implementations that show real-time performance. In empirical evaluation, the implementations yield strong performance in being able to model a wide range of potentially complicated patterns of normal activity and detect fine deviations from that model, even while being robust to changes that are insignificant (e.g., illumination and spatial appearance variations). Various compared alternative approaches were not able to yield comparatively strong results.

References

1. `http://www.cse.yorku.ca/vision/research/anomalous-behaviour`
2. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. PAMI 30, 555–560 (2008)
3. Affifi, A., Azen, S.: Statistical Analysis. Academic (1979)
4. Andrade, E., Blunsden, S., Fisher, R.: Modelling crowd scenes for event detection. In: ICPR, pp. 175–178 (2006)
5. Basharat, A., Gritai, A., Shah, M.: Learning object motion patterns for anomaly detection and improved object detection. In: CVPR (2008)
6. Bebezeth, Y., Jodoin, P., Saligrama, V., Rosenberger, C.: Abnormal events detection based on spatio-temporal co-occurences. In: CVPR, pp. 2458–2465 (2009)
7. Black, M.: Explaining optical flow events with parameterized spatio-temporal models. In: CVPR, pp. 326–332 (1999)
8. Boiman, O., Irani, M.: Detecting irregularities in images and in video. IJCV 74, 17–31 (2007)
9. Buxton, H.: Learning and understanding dynamic scene activity: A review. IVC 23 (2003)
10. Chomat, O., Crowley, J.: Probabilistic recognition of activity using local appearance. In: CVPR, pp. 104–109 (September 1999)
11. Derpanis, K., Sizintsev, M., Cannons, K., Wildes, R.: Efficient action spotting based on a spacetime oriented structure representation. In: CVPR (2010)
12. Derpanis, K., Wildes, R.: Early spatiotemporal grouping with a distributed oriented energy representation. In: CVPR (June 2009)

13. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behaviour recognition via sparse spatio-temporal features. In: PETS, pp. 65–72 (2005)
14. Elgammal, A., Durauswami, R., Harwood, D., Davis, L.: Background and foreground modeling using nonparametric kernel density for visual surveillance. Proc. IEEE 90, 1151–1163 (2002)
15. Freeman, W., Adelson, E.: Design and use of steerable filters. PAMI 13, 891–906 (1991)
16. Granlund, G., Knuttson, H.: Signal Processing for Computer Vision. Kluwer, Dordrecht (1995)
17. Heikkila, M., Pietkainin, M.: A texture-based method for modeling the background and detecting moving objects. PAMI 28, 657–662 (2006)
18. Hu, W., Xian, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: System for learning statistical motion patterns. PAMI 28, 1450–1464 (2006)
19. Jahne, B.: Digital Image Processing. Springer, Heidelberg (2005)
20. Javed, O., Shafique, K., Shah, M.: A hierarchical approach to robust background subtraction using color and gradient information. In: Motion Workshop, pp. 22–27 (2003)
21. Jodoin, P.M., Konrad, J., Saligrama, V.: Modeling background activity for behavior subtraction. In: ICDSC, pp. 1–10 (2008)
22. Johnson, N., Hogg, D.: Learning the distribution of object trajectories for event recognition. IVC 14, 609–615 (1996)
23. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: ICCV (2007)
24. Kim, J., Grauman, K.: Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In: CVPR, pp. 2921–2929 (2009)
25. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: CVPR, pp. 1446–1453 (2009)
26. Li, L., Gong, S., Xiang, T.: Global behaviour inference using probabilistic latent semantic analysis. In: BMVC (2008)
27. McIvor, A.: Background subtraction techniques. In: Proc. Vid. And Img. Comp., New Zealand (2000)
28. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using a social force model. In: CVPR (2009)
29. Mittal, A., Monnet, A., Paragios, N.: Scene modeling and change detection in dynamic scences: A subspace approach. CVIU 113, 63–79 (2009)
30. Pless, R.: Spatio-temporal background models for outdoor surveillance. In: EURASIP (2005)
31. Stauffer, C., Grimson, E.: Learning patterns of activity using real-time tracking. PAMI 22, 747–757 (2000)
32. Wang, X., Ma, X., Grimson, E.: Unsupervised activity perception by hierarchical bayesian models. In: CVPR (2007)
33. Watson, B., Ahumada, A.: A look at motion in the frequency domain. In: Motion Workshop. pp. 1–10 (1983)
34. Wildes, R., Bergen, J.: Qualitative spatiotemporal analysis using an oriented energy representation. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 768–784. Springer, Heidelberg (2000)
35. Zhong, H., Shi, J., Visontai, M.: Detecting unusual activity in video. In: CVPR (2004)
36. Zhong, J., Sclaroff, S.: Segmenting foreground objects from a dynamic textured background using a robust Kalman filter. In: ICCV, pp. 44–50 (2003)