Multi-layered Decomposition of Recurrent Scenes

David Russell and Shaogang Gong

Department of Computer Science, Queen Mary, University of London
London E1 4NS, UK [dave|sgg]@dcs.qmul.ac.uk

Abstract. There is considerable interest in techniques capable of identifying anomalies and unusual events in busy outdoor scenes, e.g. road junctions. Many approaches achieve this by exploiting deviations in spatial appearance from some expected norm accumulated by a model over time. In this work we show that much can be gained from explicitly modelling temporal aspects in detail. Specifically, many traffic junctions are regulated by lights controlled by a timing device of considerable precision, and it is in these situations that we advocate a model which learns periodic spatio-temporal patterns with a view to highlighting anomalous events such as broken-down vehicles, traffic accidents, or pedestrians jaywalking. More specifically, by estimating autocovariance of self-similarity, used previously in the context gait recognition, we characterize a scene by identifying a global fundamental period. As our model, we introduce a spatio-temporal grid of histograms built in accordance with some chosen feature. This model is then used to classify objects found in subsequent test data. In particular we demonstrate the effect of such characterization experimentally by monitoring the bounding box aspect ratio and optical flow field of objects detected on a road traffic junction, enabling our model to discriminate between people and cars sufficiently well to provide useful warnings of adverse behaviour in real time.

1 Introduction

Currently countless people are deployed to watch and monitor CCTV screens in the hope of identifying criminal activity, untoward behaviour, and serious but non-malicious situations. A fundamental challenge to computer vision research is to devise algorithms capable of isolating and displaying events of interest in a clear, uncluttered way and with a relatively low false alarm rate. Considerable research effort has produced systems which learn statistical scene content both at the pixel level [1] and from a global perspective [2] with a view to segmenting an image into the usual (background) and unusual (foreground). By relating foreground object size, and possibly shape, to areas within the scene, it becomes possible to identify people and vehicles in the 'wrong' place. However, generally such models are oblivious to relative event timing.

In this paper, with specific reference to road traffic junctions, we wish to extend the definition of 'unusual' to the temporal domain such that the presence of an object is treated explicitly in a spatio-temporal context rather than modelled as a deviation from an accumulated distribution. This approach is aimed specifically at modelling scenarios in which periodic behaviour is present. For example, it should be possible to identify a pedestrian trying to cross a road at a time when cars are moving through the junction, namely this calls for a model possessing a certain *temporal context awareness*.

1.1 Related Work

Considerable work has been published on the biological aspects of perceptual grouping. In terms of the human visual system this amounts to forming relationships between objects in an image. But such grouping also occurs in the temporal dimension, whereby our attention is drawn to objects whose appearances change together, and those whose appearance changes cyclically or periodically. At this point it is important to make the distinction between these two types of variation: Cyclic motion implies events in a certain sequence, whereas Periodic motion involves events associated strictly with a constant time interval.

Within the field of biologically inspired computing, systems using networks of Spiking RBF (Radial Basis Function) Neurons have been used in [3] to characterize and identify spatio-temporal behaviour patterns. Such a neuron generates a pulse of activity when the combination of its inputs reaches a critical threshold. The network of connections from input neurons to output neurons contains groups of parallel paths with varying synaptic delays whose relative weights are learned in a Hebbian fashion such that the delay pattern eventually complements (mirrors) the times between events in training data. By this mechanism, an output neuron can 'learn' to fire when the appropriate events occur with correctly matched time delays, since only under this condition will all spikes reach the nucleus simultaneously, causing its threshold to be breached and hence firing.

This idea is applied to a practical vision system in [4], whereby relations between pixels in the Motion History Image (MHI) over a sequence are learned for a simple shopkeeper/customer scenario. Abnormal behaviour is detected when a customer takes an item of stock but leaves the shop without paying the shopkeeper. Similarly using MHI, [5] discriminates between actions based on movement of the human body by matching against various learned templates. But so far, although these examples identify sequences of learned events occurring at precise times, whereas overall the sequences themselves are asynchronous events - they might happen only once, or repeatedly but at arbitrary times. A model described in [6] forms relations between asynchronous but related scene events by adding links between parallel Hidden Markov Models, making it ideal for many situations where temporal invariance is paramount.

When it comes to periodic motion, [7] describes a method of modelling moving water, flames, and swaying trees as Temporal Textures. An Autoregressive Model is proposed in which a new frame may be synthesized such that each pixel is described by a weighted sum of previous versions of itself and its neighbours, with an added Gaussian noise process. Similar to the Temporal Textures of [7], [8] applies the Wold decomposition to the 1-D temporal signals derived from each image pixel giving rise to deterministic (periodic) and non-deterministic (stochastic) components, permitting distinction between various human and animal gaits, and other types of motion.

On an apparently unrelated problem, much is to be found in the literature concerning gait characterization, modelling and identification. Generally these methods work by analyzing the relative motion of linked body members, which are of course all related by the same fundamental frequency. However, the parallel between this and modelling traffic at a road junction is surprisingly close. Given extracted features, image areas may be likened to body limbs, sharing fundamental frequency, but being of arbitrary phase and harmonic content.

Various forms of periodic human motion are characterized in [9] by tracking candidate objects and forming their 'reference curves'. After evaluating a dominant spectral component if it exists, an appropriate temporal scale may be identified. This idea is developed in [10] which considers periodic self-similarity, Fisher's Test for periodicity and Time Frequency Analysis. The Recurrence Plot described in [11] is a useful tool for visualizing the evolution of a process in state-space, showing specifically when the state revisits a previous location.

Instead of using Fourier analysis directly, [12] employs Phase Locked Loops (PLLs) to discriminate between different gaits, on the basis that it is more efficient. Having identified some fundamental frequency for an object (person), application of a PLL per pixel in the relevant area permits estimation of the magnitude and relative phase of this fundamental component for each pixel in the object. The idea is that the phase 'signature' for every object (person) will be different. The technique is rendered scale and translation invariant by matching these parameters as shapes in the complex plane using the Procrustes mean.

In this work we wish to construct an algorithm to characterize the periodicity of a scene based on its temporal statistics rather than explicit object tracking (therefore avoiding the catch-22 problem of determining appropriate scale vs. saliency). Treating the recovered periodicity as a form of 'temporal background' we aim to discover anomalies in both space and time simultaneously in unseen images. Expanding on a technique employing self-similarity [10], we describe an algorithm for extracting the fundamental period from a video sequence of a scene, and then use this to facilitate a spatio-temporal data-driven model of scene activity. We show experiments in three traffic junctions scenes where we demonstrate the effectiveness and simplicity of such a model in performing anomaly detection.

2 Our Model

Given a video sequence $I_{x,y,t}$ consisting of t_{max} frames each of size $x_{max} \times y_{max}$ pixels in which (x, y) represents spatial pixel location, t the time index, and I the colour triple $\{R, G, B\}$, we split the data into two parts, the first for training and the second for evaluation. Obviously, the first image of the test sequence directly follows the final image from the training sequence - a fact which becomes crucial in ensuring the initialized model is synchronized with the test data. This also enables a natural way for bootstrapping a model from limited initial exposure to the scene. A background model $I^B_{x,y,t}$ is evaluated from and maintained through both the training and test data according to a method detailed in [13]. Our overall algorithm is shown in Figure 1, and described in more detail in the following.

S	Description
1	Derive a background model from training sequence
2	Extract chosen feature from training sequence
3	Quantize samples to a coarser spatio-temporal grid forming linear state data
4	Find dominant fundamental period T_{fund} for the scene using the linear state data
5	'Roll up' Linear State Data using period T_{fund} starting from the end to form average State Cycle estimate
6	Use State Cycle to classify previously unseen frames
7	Synthesize output from background and mis-matched areas in new frames

Fig. 1. Steps in our algorithm

2.1 Feature Selection

A feature which summarizes some local characteristic of the image sequence must be chosen. For modelling the traffic junction we start with selecting the aspect ratio of an object's bounding box, anticipating that pedestrians will always be taller than they are wide, and vehicles will rarely be so under the majority of typical poses. In order to ensure symmetrical treatment of ratios greater and less than unity, we further develop a Log Aspect Ratio (LAR) feature $LAR_{x,y}$ at position (x, y) by taking the natural logarithm and clipping to +/-1, resulting in ratios from $\frac{1}{e}$ to e

$$LAR_{x,y} = \max\left(-1, \min\left(1, \log_e\left(\frac{h_{x,y}}{w_{x,y}}\right)\right)\right) \quad (1)$$

where h and w are box height and width respectively. Bounding boxes are determined after applying morphological operations to a foreground binary mask $M^{fg}_{x,y,t}$ removing shapes below a certain minimum pixel area. The binary mask $M^{fg}_{x,y,t}$ is derived from the difference between the current image and the current background $D_{x,y,t}$ according the the L_1 (Manhattan) norm of the pixel vectors in colour space

$$M^{fg}_{x,y,t} = \begin{cases} 1 \text{ if } D_{x,y,t} > \tau \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

where τ is a constant and

$$D_{x,y,t} = \left\| I_{x,y,t} - I^B_{x,y,t} \right\|_1 \quad (3)$$

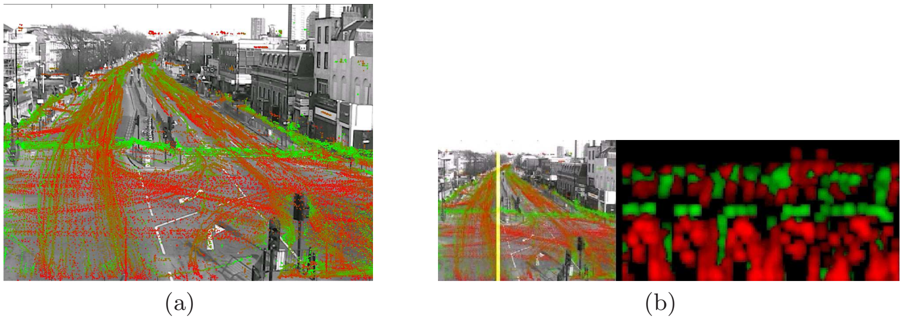
Thus for each frame of video I_t, a (potentially empty) list L_t of valid bounding boxes $B_{t,m}$ is produced governed by the above rules

$$L_t = \{B_{t,1},\ B_{t,2},\ \ldots\ B_{t,m}\} \tag{4}$$

where the mth bounding box is characterized by the quad

$$B_{t,m} = \{x, y, w, h\} \tag{5}$$

in which (x, y) is the bounding box centre, and (w, h) are its size from which the LAR is calculated. The maximum value of m is determined by the number of objects detected in the current image. So the feature we have selected does not exist at every pixel, rather it will exist wherever in the spatio-temporal volume valid objects are detected. Figure 2(a) shows an example of accumulation of LAR over time in the training data, showing how it discriminates between people and vehicles. Meanwhile with a plot of the image y-axis against time, Figure 2(b) illustrates the inherently periodic nature of activity on a road junction.


(a)

(b)

Fig. 2. (a) Bounding Box centres accumulated over time at a road junction scene. Colour represents aspect ratio: green samples have $h > w$ (pedestrians), red samples have $h < w$ (vehicles). The ratio for vehicles becomes unreliable here in the far distance. (b) Y-T cut (right) through the spatio-temporal volume showing periodic behaviour of a road junction scene at the vertical yellow line (left).

2.2 Spatio-temporal Histogram

Thus far the training data is represented by a set of points in a 4-D space (x, y, t, LAR). In order to facilitate comparison of feature occurrence within the spatio-temporal volume, we seek to build a spatio-temporal set of histograms over the feature space. Therefore we split the volume into a grid of $h_{max} \times v_{max}$ equal sized square blocks of pixels spatially and n_{max} equal sized blocks of frames temporally. At each spatio-temporal grid position, consisting of

$$\frac{x_{max}}{h_{max}} \times \frac{y_{max}}{v_{max}} \times \frac{t_{max}}{n_{max}} \tag{6}$$

pixels we construct a histogram $H_{h,v,t}$ of b_{max} equal width bins over feature space. For LAR it is a bounded 1-D set

$$H_{h,v,n}(b) = \{b_1,\ b_2,\ \ldots\ b_{max}\} \tag{7}$$

$$\text{where} \quad b = \lfloor \frac{b_{max}(LAR + 1)}{2} + 1 \rfloor \tag{8}$$

such that the range of the LAR feature ($-1 \leq LAR \leq +1$) is mapped uniformly onto bin number b, where $1 \leq b \leq b_{max}$. The inherent loss of resolution in all dimensions as a result of this down-sampling operation is countered by the advantage of being able to quantify the similarity between any two spatio-temporal regions on the basis of the selected feature purely by comparing histograms. In fact from this point on, the method becomes independent of the chosen feature and thus offers a degree of generality and considerable scope for matching any chosen feature(s).

2.3 The Sparsity Problem

It is quite possible that, given the relatively high dimensionality of the histogram containing the bounding box data points, the density is insufficient to yield meaningful distributions. One potential solution is to decrease the number of blocks in the grid in the dimension(s) causing the deficiency. Alternatively a degree of data smoothing may be applied, both over the bins within each histogram and also between spatio-temporal histograms. It was found that experimental results benefited from convolution of the former with a normalized 1-D Gaussian filter, and of the latter with a 3-D Gaussian kernel having potentially different variance in the spatial and temporal directions. Inevitably there will be some regions which are poorly supported, and steps to mitigate the effects of this may become necessary in later processing.

2.4 Fundamental Period Estimation

To derive an estimate of the fundamental period over which scene changes occur is a non-trivial procedure, and as such it is dealt with separately in Section 3. Suffice to say at this point that a scene may have a number of unrelated fundamental periods (including ‘none’) distributed over various regions (see Figure 3), and optimally distinguishing them is a topic for future research. In this work we consider applications like the traffic junction where it is assumed that there *is* a single dominant effect, for which the period is K_{fund} blocks each of t_{max}/n_{max} frames. Given a frame rate of F per second, the fundamental period is thus

$$T_{fund} = \frac{K_{fund}}{F} \frac{t_{max}}{n_{max}} \quad \text{seconds.} \tag{9}$$

Ideally the training data should be long enough to contain sufficient cycles of the fundamental period that the latter can be distinguished adequately from noise.

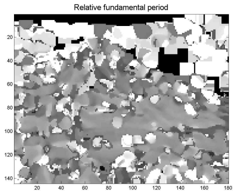


Fig. 3. Relative fundamental period distribution of the scene in Figure 2(a) based on per pixel temporal autocorrelation. Intensity representing period is given by the first significant peak. Much of the junction area is the same shade, indicating shared periodicity.

2.5 State Cycle and Model Initialization

We define the State Cycle $S^k_{h,v} \quad k = \{1 \ldots K_{fund}\}$ of a grid location (h, v) to be a temporal description of how the chosen feature varies throughout a single cycle of its fundamental period of K_{fund} phases. Given that the array $H_{h,v,n}$ contains a number of cycles of this temporal description in succession, we wish to form an 'average histogram' H_{fund} of size $h_{max} \times v_{max} \times K_{fund}$ representing a summary of the scene's typical behaviour over the c most recent cycles of the fundamental period, where $c = \lfloor \frac{n_{max}}{K_{fund}} \rfloor$ cycles. Thus taking the c most recent groups of K_{fund} blocks, the kth element of H_{fund} is the mean of the kth elements of the c groups

$$H_{fund,h,v,k}(b) = \frac{1}{c} \sum_{i=1}^{c} H_{h,v,n_{max}-iK_{fund}+k}\,(b) \tag{10}$$

where $k = \{1,\ 2,\ \ldots\ K_{fund}\}$. Normalization of H_{fund} over b yields an estimate of feature probability P_{fund} which is then our spatio-temporal model of the scene

$$P_{fund,h,v,k}(b) = \frac{H_{fund,h,v,k}(b)}{\sum_{b=1}^{b_{max}} H_{fund,h,v,k}(b)} \tag{11}$$

Assuming that continuous test sequence (e.g. real-time video streamed data) directly follows the initial training sequence, then the state counter k, initialized to 1, may be updated every $\frac{t_{max}}{n_{max}}$ frames according to the relation $k = \text{mod}\,(k, K_{fund}) + 1$ in order to keep track of the learned periodic scene behaviour.

2.6 Output Synthesis

The objective is to provide an output sequence from our algorithm showing only objects in the 'wrong place' at the 'wrong time'. For a query test frame

I^{query} appearing subsequent to model initialization, the foreground mask M^{fg} is obtained as in equation (2), and valid object bounding boxes $B_{t,m}$ derived as in (5). For each candidate bounding box, the LAR is evaluated from width and height using equation (1) and b is given by (8). Values for h and v are calculated using $h = \frac{x \times h_{max}}{x_{max}}$ and $v = \frac{y \times v_{max}}{y_{max}}$. Thus the estimated probability of that particular aspect ratio bounding box at that position is given by the model, and may be compared with a threshold α in order to give a binary decision r as to whether the object is sufficiently rare to be displayed

$$r = \begin{cases} 1 \text{ if } P_{fund,h,v,k}(b) < \alpha \\ 0 \text{ otherwise} \end{cases} \tag{12}$$

On the basis of r being true, for each object in I^{query}, a matting mask M^{matt} is used to re-insert pixels according to the bounding box dimensions from the new frame I^{query} into the background I^B for all objects determined to be anomalous with respect to the current model. The background with insertions forms the output image from the algorithm.

3 Determining the Fundamental Period

The method described in the previous section relies totally on obtaining a robust estimate of the fundamental period of a region or the whole image area using the 3-D spatio-temporal grid of histograms $H_{h,v,n}$ defined in (7). We seek to find the most common lag between instances of temporal self-similarity at times n_1 and n_2 over all possible combinations of n_1 and n_2. As a measure of the similarity between any two histograms, we utilize the general definition of the symmetric Kullback-Leibler Divergence (KLD) between distributions P_1 and P_2 given by

$$D_{KL}(P_1, P_2) = \sum_i (P_{1,i} \log_2 (\frac{P_{1,i}}{P_{2,i}}) + P_{2,i} \log_2 (\frac{P_{2,i}}{P_{1,i}})) \text{ bits} \tag{13}$$

Thus over an arbitrary spatial region R in our grid, we define the 'average Dissimilarity matrix' S between two temporal planes at times n_1 and n_2 as

$$S_{n_1,n_2} = \frac{1}{\|R\|} \sum_{v,h \in R} D_{KL}(P_{n_1}(v,h), P_{n_2}(v,h)) \tag{14}$$

which after simplification yields

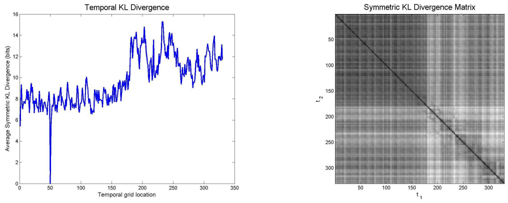
$$S_{n_1,n_2} = \frac{1}{\|R\|} \sum_{v,h \in R} \sum_{i=1}^{b_{max}} (P_{n_1,i} - P_{n_2,i}) \log_2 \left(\frac{P_{n_1,i}}{P_{n_2,i}} \right) \tag{15}$$

An example of the symmetric Divergence relative to a single time is illustrated in Figure 4(a), and between all combinations of times as matrix S in Figure 4(b). Because it is the coincidence of minima in S that we are interested in, we subtract its mean to form S'

$$S'(i,j) = S(i,j) - \frac{1}{i_{max} j_{max}} \sum_{i,j} S(i,j) \tag{16}$$

and construct the normalized 2-D autocovariance matrix A at all possible lags (d_i, d_j) in both directions

$$A(d_i, d_j) = \frac{\sum_{i,j} S'(i,j)\; S'(i+d_i, j+d_j)}{\sqrt{\sum_{i,j} S'(i,j)^2 \cdot \sum_{i,j} S'(i+d_i, j+d_j)^2}} \tag{17}$$



(a) (b)

Fig. 4. (a) Temporal KL Divergence at a single grid position (corresponding to 50 on the x-axis) relative to all other temporal grid positions. Naturally the divergence is zero with respect to itself. (b) Average 'Divergence' matrix between histograms at temporal grid positions n_1, n_2 for all combinations of n_1 and n_2. Using the Symmetric Kullback-Leibler formula, divergence is summed over all spatial grid positions of the scene, as well as over the histogram bins (equation (15)).

As shown in Figure 5(b), matrix A exhibits a regular structure of peaks spaced at the dominant period if it exists. The fundamental interval K_{fund} is identified by exploratory element-wise multiplication of A with a regular matrix of peaks generated by column vector $g(d)$ as shown in Figure 5(a), whereby varying the pitch d yields a peak in the overall temporal scene power observed

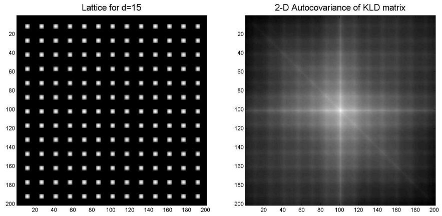
$$K_{fund} = \arg\max_{d}(g(d)^T A\; g(d)) \tag{18}$$

for $d_{min} \leq d \leq d_{max}$ and binary vector g such that

$$g_i(d) = \delta((i - n_{max}) \bmod d) \text{ where } 1 \leq i \leq 2n_{max} - 1 \tag{19}$$

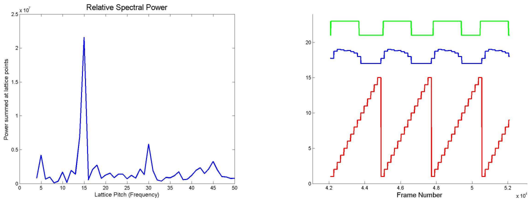
Figure 6(a) shows how the scene's signal power peaks at a given value of d.

For our application the region R represents the entire scene, but this technique could equally well work with subsets of the scene, be they rectangular or square blocks, or even arbitrary shapes. A yet more elaborate scheme for analyzing the autocovariance matrix A is described in [10], in particular explaining that a diagonal equivalent of the matrix in Figure 5(a) is necessary to detect periodicity in scenes in which self-similarity of appearance peaks more than once per cycle (e.g. a swinging pendulum).



(a) (b)

Fig. 5. (a) Lattice for distance $d = 15$ generated by $g(d)g(d)^T$. Multiplying such a lattice by the autocovariance matrix in (b) for a range of d identifies the fundamental period. (b) Autocovariance of the Divergence matrix in Figure 4(b), showing the strong lattice structure corresponding to a dominant fundamental in the video sequence.



(a) (b)

Fig. 6. (a) Relative spectral power of the scene in Figure 2(a) for values of d between 4 and 50. Note the fundamental at $d = 15$, giving a period of $15 \times 7.5s = 112.5s$ corresponding to the cycle time of the junction signals. (b) Timing diagram showing correct synchronization of model throughout test sequence. Top: Pixels from closest green traffic light in scene. Middle: Consensus of light over cycles in training data. Bottom: Internal state counter. Note consistent phase relationship between all three.

4 Experiment

For our experiments we chose three busy city-centre road junctions controlled by traffic lights. Each dataset was made up of 30000 frames of 720×576 pixel colour video at a frame rate of 25Hz, yielding sequences of 20 minutes duration. The data was spatially down-sampled to 360×288 pixels to ease computational load. The short-term background model was obtained as described using the method described in [13], based on blocks of 20 frames taken at 12 second intervals. The L_1 norm of the background-subtracted data was thresholded at a value of 30 given an intensity range of 0-255 per colour channel, and after morphological clean-up, identified object areas were thresholded to reject those below 70 pixels. The Log Aspect Ratio feature range of +1 to -1 was split into 5 histogram bins, and spatio-temporal histogram grid was 8×8 pixels wide spatially, and 180 frames deep temporally, giving $h = 45$, $v = 36$, and $n = 167$. For each sequence, we utilized the entire spatio-temporal matrix to estimate the global fundamental period K_{fund} for the scene using the method described in Section 3. We then allowed $c = 5$ cycles of this fundamental period to be used for training data, leaving the remainder for testing. Figure 6(b) illustrates how the state counter is correctly and consistently aligned with junction activity throughout the test sequence, as measured by the actual brightness of pixels representing the green traffic light at the bottom of the scene.

The results for Scenarios 1,2 and 3 are shown in Figures 7, 8, and 9. Figures 7 and 8 show 3 rows of 5 images, with each row representing an example frame from the algorithm output. The left-most image is the original unprocessed frame, whilst the second image is the short-term 'static' background which we have labelled as 'Layer 0'. The objects detected to be anomalous according to our model are shown inserted into the static background and labelled as 'Layer 2' - the foreground. Similarly, the original image with the background inserted where the object was detected, is shown labelled as 'Layer 1' - the dynamic background.

Finally in the right-hand column, for comparison purposes, we show the result of classification using a non-temporal equivalent model derived from the same training data. To achieve this, bin values of each histogram $P_{h,v,k}(b)$ are marginalized out over the time dimension to yield $P'_{h,v}(b)$. Overall, when analyzing images, the algorithm achieves 3FPS throughput on a 2GHz PC, although initially building the model carries a considerably higher computational cost.

5 Discussion

The results in Figures 7, 8 and 9 demonstrate how, in spite of a background that is non-stationary, our algorithm has managed to split scene activity into 3 distinct layers. This has been achieved partly by being able to make reliable estimates of true background amongst a busy scene, and partly by classifying objects based on a spatio-temporal template learned from the scene during training.

What we term Layer 0 takes on the non-stationary background, permitting detection of less persistently occurring objects such as people and vehicles. Having thus obtained reference to the latter in isolation from the background, our spatio-temporal model classifies them into Layer 1, objects of a suitable aspect ratio for the part of state-space they occupy, and Layer 2, objects which contradict the model. Within this framework, Layer 1 has taken on the role of a 'dynamic background' in relation to what might frequently be referred to as 'foreground' objects. Such a dynamic background has three dimensions, and a match in all of them is required as well as an acceptable value for the feature at those coordinates in order that the object is deemed acceptable as a dynamic background item. Thus we claim that our spatio-temporal model has more discriminative power than a spatial-only 2-D probabilistic model, which is oblivious to time. By marginalizing out the time dimension, one effectively increases the likelihood of an object at times in the cycle when it should be considered rare, and reduces its likelihood at times when it should be considered common. The overall unwanted result is thus a desensitization of the model.

The upshot of this situation is that with no temporal processing (termed 'NTP') too many unimportant objects are detected, whilst use of our scene-synchronized spatio-temporal model reveals far more salient detection amongst ‘higher layers’ of temporal change, associated with interesting and unexpected spatio-temporal events. Furthermore, all this may be achieved without prior knowledge of the size and location of potential triggering objects in the scene.

In particular, among the results are examples of our model detecting objects of interest, whilst the model without temporal processing *fails* to highlight these, but identifies *less* truly interesting objects instead. That this remains so, however one decides to select the detection thresholds for the respective models, strongly supports our claim that the temporal dimension is highly significant.

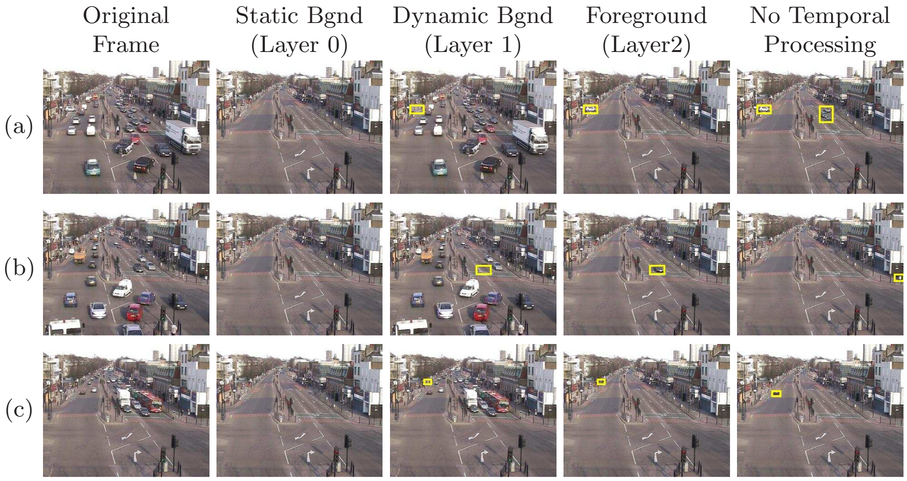


Fig. 7. Examples from Scenario 1 show how the algorithm discovers objects not matching the learned spatio-temporal template, and thus splits the scene into 3 layers on the basis of its dynamic behaviour. Layer 0 is the continuously updated ‘static’ background, Layer 1 normal scene activity - the ‘dynamic background’, and Layer 2 carries ‘novel’ intrusions with respect to the training data. Some objects cannot be separated, regardless of threshold chosen. In (a) L2 correctly shows a car unusually pulling out onto the main road, whereas with No Temporal Processing (NTP), this cannot be distinguished from normal cars on the right. In (b) L2 spots the car over the waiting line, whereas NTP sees only a passing pedestrian. In (c) L2 finds pedestrians waiting at the crossing, whilst NTP wrongly highlights a car.

6 Conclusion and Further Work

We have demonstrated an algorithm capable of automatically learning the global periodicity of scenes, such as that exhibited at junctions controlled by traffic lights. The technique estimates a value for the global fundamental period, and then builds a spatio-temporal model based on this estimate. It has been demonstrated by experiment that the method can be more discriminating with regard to activity of a periodic scene than a model oblivious to repeating temporal trends. As such, we draw the conclusion that the method described has successfully decomposed the scene into separate layers on the basis of its dynamic characteristics. Even using only a restricted feature set, the approach achieves good results. However, as previously alluded to, the histograms defined could readily represent a more diverse range of image features.

In its present form, the model described estimates the period once during training. A practical realization would need to re-evaluate the fundamental period continuously, in order to maintain both frequency and phase lock with respect to current scene activity, especially since many scenes will not be quite periodic in some way. Both short-term phase noise and longer-term frequency drift problems may be soluble using the Phase Locked Loop approach detailed in [12], whilst an on-line solution which augments the current model with additional data as it becomes available would make for a truly adaptive system.

It is clear that many scenes will be composed of more than one harmonically unrelated periodic component. Instead of seeking a single global fundamental, the scene may be searched in a systematic fashion using the estimation technique we have described on smaller regions. If somewhat optimal regions of common periodicity could be found, the 'rolling up' of periodic training data implemented here is equally applicable to different image areas, each with its own K_{fund}.

References

1. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: IEEE CVPR, Colorado (1999) 246–252
2. Oliver, N., Rosario, B., Pentland, A.: A bayesian computer vision system for modelling human interactions. IEEE PAMI **22**(8) (August 2000) 831–843
3. Natschläger, T., Ruf, B.: Spatial and temporal pattern analysis via spiking neurons. Network: Computation in Neural Systems **9**(3) (1998) 319–332
4. Ng, J., Gong, S.: On the binding mechanism of synchronised visual events. In: IEEE Workshop on Motion and Video Computing. (December 2002)
5. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE PAMI **23**(3) (2001) 257–267
6. Xiang, T., Gong, S.: Beyond tracking: Modelling activity and understanding behaviour. IJCV **67**(1) (2006) 21–51
7. Szummer, M.: Temporal texture modeling. Technical Report 346, MIT Media Lab Perceptual Computing (1995)
8. Liu, F., Picard, R.W.: Finding periodicity in space and time. In: ICCV. (1998) 376–383
9. Polana, R., Nelson, R.C.: Detection and recognition of periodic, nonrigid motion. IJCV **23**(3) (1997) 261–282
10. Cutler, R., Davis, L.S.: Robust real-time periodic motion detection, analysis, and applications. IEEE PAMI **22**(8) (2000) 781–796
11. Casdagli, M.: Recurrence plots revisited. Physica D **108** (1997) 12–44
12. Boyd, J.: Synchronization of oscillations for machine perception of gaits. CVIU **96**(1) (October 2004) 35–59
13. Russell, D., Gong, S.: Minimum cuts of a time-varying background. In: BMVC. (Sep 2006) 809–818

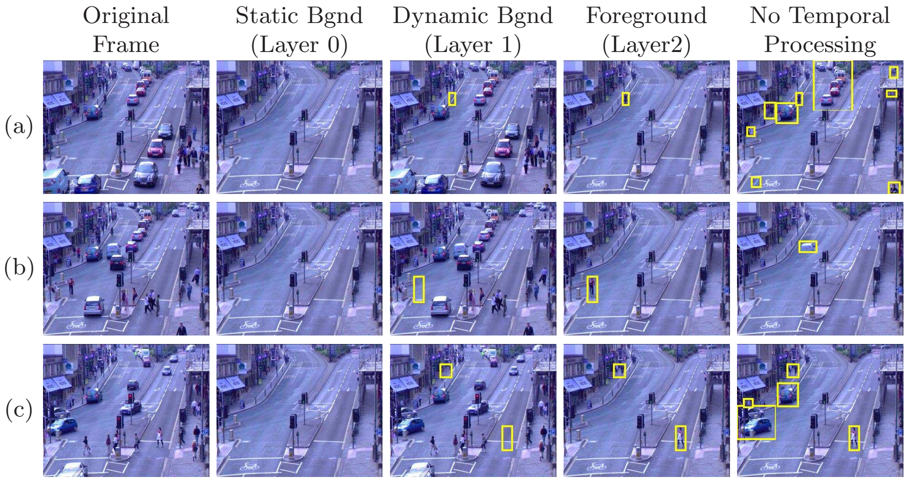


Fig. 8. Examples from Scenario 2, an entirely different traffic junction. From behind, cyclists tend to have an aspect ratio similar to people. Thus in (a) L2 singles out a cyclist close to the pathway, which with No Temporal Processing (NTP), cannot be separated. In (b) L2 has detected a different cyclist, again with the same profile as a person, where there should not be people, whilst NTP sees only part of a car in normal position. In (c) L2 observes a person on the wrong part of the crossing, inseparable from vehicles on the junction with NTP.

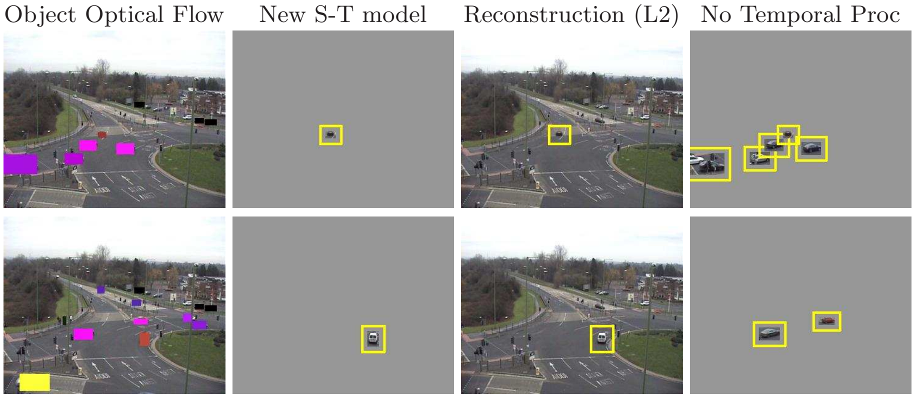


Fig. 9. Scenario 3 with Optical Flow as the feature instead of shape ratio. Top: Spatio-Temporal model correctly highlights errant vehicle crossing normal traffic from left. Bottom: Spatial-only model (NTP) wrongly highlights normal traffic instead of van jumping the red signal.