Exploiting Periodicity in 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 of scene activity in detail. We characterize a scene by identifying the fundamental period of change on a spatial block-by-block basis by estimating autocovariance of self-similarity. As our model, we introduce a spatio-temporal grid of histograms built corresponding to some chosen feature. This model is then used to identify objects found in unexpected spatial and temporal locations in subsequent test data. Employing a Phase-Locked Loop technique, we describe a method of ensuring that the spatio-temporal model maintains *synchronization* with learned scene activity in spite of short-term breakdown in the reliability of acquired data, and long-term change of the mean fundamental period. Results indicate our model to be capable of discrimination between behavioural aspects of cars at a typical road junction sufficiently well to provide useful warnings of adverse activity 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 [14] and from a global perspective [10] with a view to segmenting an image into the usual (background) and unusual (foreground). By relating foreground object size, shape, and direction 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 spatial-only distribution. This approach is aimed specifically at modelling scenarios in which periodic behaviour is present. For example, it should be possible to identify a car crossing a junction when the traffic lights for that direction are red. Namely this calls for a model possessing a certain *temporal contextual awareness*.

BMVC 2008 doi:10.5244/C.22.71

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. 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 [8] 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 it.

This idea is applied to a practical vision system in [9], 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, [2] 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, overall the sequences themselves are asynchronous events - they might happen only once, or repeatedly but at arbitrary times. A model described in [16] forms relations between asynchronous but related scene events by dynamically adding links between parallel Hidden Markov Models, making it ideal for many situations where temporal invariance is paramount.

When it comes to periodic motion, [15] 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, together with an added Gaussian noise process. Similar to the Temporal Textures of [15], [6] 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 all related by the same fundamental. The parallel between this and modelling road junction traffic is surprisingly close. Given extracted features, image areas may be likened to body limbs in that they will likely share fundamental frequency, but be of arbitrary phase and harmonic content.

Various forms of periodic human motion are characterized in [11] 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 [5] which also considers periodic self-similarity, Fisher's Test for periodicity

and Time Frequency Analysis. Meanwhile the Recurrence Plot described in [4] 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, [3] 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), use of a PLL per pixel permits estimation of the magnitude and relative phase of this fundamental component for each pixel making up 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, thus avoiding the catch-22 problem of determining appropriate scale versus saliency. Treating the recovered periodicity as a *temporal background* we aim to discover anomalies in both space and time simultaneously in unseen images. Expanding on a technique employing self-similarity [5], we describe an algorithm for extracting fundamental periods from a video of a scene, and then use these to facilitate a block-based spatio-temporal data-driven model of scene activity. Experiments on two traffic junctions scenes show the effectiveness and simplicity of such a model in performing anomaly detection.

The work presented here relates closely to a method described in [13] whereby a single *global* periodicity of a traffic scene is characterized. The single periodicity represents a severe limitation to the scalability, since most real scenes are much more complicated than this. We thus propose a generalization towards a much more flexible block-wise *multi-periodicity* approach. In addition, we describe a way of ensuring that each per block model remains synchronized to incoming data - a vital aspect if the approach is to be adopted in any practical system.

2 Spatio-Temporal Model

Our model involves building histograms over some chosen feature localized in space and time. Determining the fundamental period at a spatial block location then becomes a problem of finding matches between repeating sets of histograms. The method is thus somewhat decoupled from the particular chosen feature, and as long as it can be expressed by some distribution, any feature suitable for the application at hand may be used.

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. The pixel volume is then split 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

$$\frac{x_{max}}{h_{max}} \times \frac{y_{max}}{v_{max}} \times \frac{t_{max}}{n_{max}} \tag{1}$$

pixels we construct a separate histogram $H_{h,v,n}$ of b_{max} bins over the selected feature space $H_{h,v,n}(b) = \{b_1,\ b_2,\ \ldots\ b_{max}\}$ where h,v and n are spatial and temporal coordinates.

S	Description
1	Choose a suitable feature for the application
2	Extract chosen feature from training sequence
3	Build 3D histogram block by quantizing training data onto a spatio-temporal grid
4	Find dominant fundamental period T_{fund} for each spatial block in scene
5	Form model from Ensemble Avg. of histograms of length T_{fund} in training data
	REPEAT FOREVER
6	Use Ensemble Average model to detect anomalies in unseen frames
7	Form new histogram data from T_{fund} most recent unseen frames
8	Use new data to verify/correct synchronization with model using PLL
	END

Figure 1: Summary of steps in our algorithm

2.1 Feature Selection

For effective analysis of traffic junctions, experiments indicate that coarse histograms in 2D over optical flow, and in 3D over colour component intensity are both useful features. Evidently, a dimensionality 'explosion' occurs if too many features have to be represented at too high a resolution. For optical flow, objects detected by thresholded background subtraction are identified by the coordinates of the centres of their bounding boxes, and a unique flow vector is evaluated for the connected object by the Lucas Kanade method [7]. Each of the x and y flow directions is quantized into only three bins according to whether it is positively or negatively greater than a threshold v_t from zero, or has magnitude less than v_t. Thus the 2D histogram has 3×3 bins, and $b_{max} = 9$. In the case of the colour intensity histogram, the integer range 0-255 for each colour channel is quantized linearly into 4 bins, yielding a $4 \times 4 \times 4$ bin 3D histogram, such that $b_{max} = 64$.

In general, the relatively high potential dimensionality of histograms may lead to a sparsity of data points in each bin. Thus for a given size of training set, a trade-off has to be struck between spatio-temporal resolution and point density, if meaningful distributions are to be achieved. Our algorithm is summarized in Figure 1.

2.2 Fundamental Period Estimation

To estimate the fundamental period over which scene changes occur is non-trivial, and as such it is dealt with in more detail in Section 3. Suffice to say at this point that a scene may exhibit a number of unrelated fundamental periods (including 'none') distributed over various scene regions. For each spatial block (h, v) we define a fundamental period of $K_{h,v}^{fund}$ states, measured in the temporal grid resolution, and relate it to a time

$$T_{h,v}^{fund} = \frac{K_{h,v}^{fund}}{F} \frac{t_{max}}{n_{max}} \quad \text{seconds} \tag{2}$$

given a frame rate of F per second. 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 - normally at least 10 cycles in our experiments.

2.3 State Cycle and Model Initialization

We define the State Cycle $S^k_{h,v}$ where $k_{h,v} = \{1 \ldots K^{fund}_{h,v}\}$ 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 $K^{fund}_{h,v}$ phases. Given that the histogram array $H_{h,v,n}$ contains a number of cycles of this temporal description in succession, we wish to form an Ensemble Average, or 'average histogram' per block $H^{fund}_{h,v}$ of size $K^{fund}_{h,v}$ representing a summary of the scene's typical behaviour at (h,v) over the $c_{h,v}$ most recent cycles of the fundamental, where $c_{h,v} = \lfloor \frac{n_{max}}{K^{fund}_{h,v}} \rfloor$ cycles. Taking the $c_{h,v}$ most recent groups of $K^{fund}_{h,v}$ blocks, the kth element of $H^{fund}_{h,v}$ is the mean of the kth elements of the $c_{h,v}$ groups for each bin b

$$H^{fund}_{h,v,k_{h,v}}(b) = \frac{1}{c_{h,v}} \sum_{i=1}^{c_{h,v}} H_{h,v,n_{max} - iK^{fund}_{h,v} + k_{h,v}}(b) \tag{3}$$

where $k_{h,v} = \{1, 2, \ldots K^{fund}_{h,v}\}$. Normalization of $H^{fund}_{h,v}$ over b yields an estimate of feature probability $P^{fund}_{h,v}$ which is then our spatio-temporal model of the scene

$$P^{fund}_{h,v,k}(b) = \frac{H^{fund}_{h,v,k}(b)}{\sum_{b=1}^{b_{max}} H^{fund}_{h,v,k}(b)} \tag{4}$$

In principle, the state counter $k_{h,v}$, initialized to 1, may be updated every $\frac{t_{max}}{n_{max}}$ frames according to the relation $k_{h,v} = \mathrm{mod}\,(k_{h,v}, K^{fund}_{h,v}) + 1$ in order to keep track of the learned periodic scene behaviour. In practice, the exact update rate is dictated by each block's PLL system to be described in Section 4.

2.4 Output Synthesis

The goal 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} the foreground mask M^{fg} and valid object bounding box for each object are obtained as described in Section 5. Then h and v are calculated using $h = \frac{x \times h_{max}}{x_{max}}$ and $v = \frac{y \times v_{max}}{y_{max}}$. The estimated probability of a particular object being at a position is given by the normalized bin value of the histogram at that location, and may be compared with a threshold α in order to give a binary decision r as to whether the object is considered sufficiently rare to be displayed

$$r = \begin{cases} 1 & \text{if } P^{fund}_{h,v,k}(b) < \alpha \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

On the basis of r being true, for each object in I^{query}, a matting mask M^{matt} is used to re-insert object pixels according to bounding box dimensions from the new frame I^{query} into the background I^B for all objects determined to be anomalous. The background image with matted objects constitutes the useful output 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 block using the 3-D spatio-temporal grid of histograms $H_{h,v,n}$. Following a method detailed in [5], 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 dis-similarity between any two histograms, we utilize the general definition of the symmetric Kullback-Leibler Divergence (KLD) between two discrete distributions P_{n1} and P_{n2} given by

$$D_{KL}(P_{n1},P_{n2}) = \sum_{b=1}^{b_{max}} (P_{n1,b}\log_2(\frac{P_{n1,b}}{P_{n2,b}}) + P_{n2,b}\log_2(\frac{P_{n2,b}}{P_{n1,b}})) \text{ bits} \tag{6}$$

An example of the symmetric Divergence relative to a single time is illustrated in Figure 2(a), and between all combinations of times as matrix S in Figure 2(b), where $S(n1,n2) = D_{KL}(P_{n1},P_{n2})$. Because it is the coincidence of minima in S that we are interested in, we subtract its mean to form S'

$$S'(n1,n2) = S(n1,n2) - \frac{1}{(n_{max})^2}\sum_{n1=1}^{n_{max}}\sum_{n2=1}^{n_{max}} S(n1,n2) \tag{7}$$

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

$$A(d_i,d_j) = \frac{\sum_{i,j} S'(i,j)\, S'(i+d_i,j+d_j)}{\sqrt{\sum_{i,j} S'(i,j)^2 \cdot \sum_{i,j} S'(i+d_i,j+d_j)^2}} \tag{8}$$

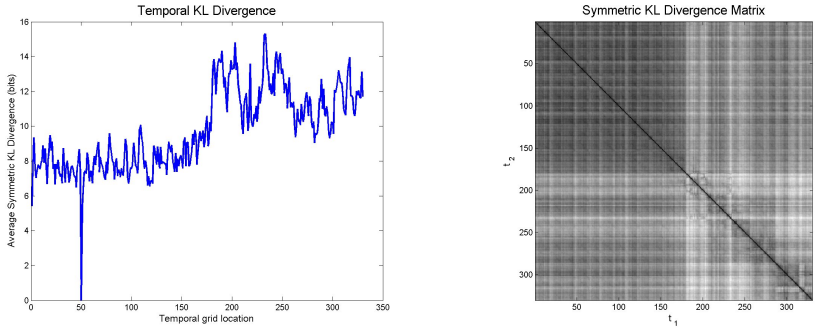

Figure 2: (a) Temporal KL Divergence at a single grid position (50 on the x-axis) relative to all other temporal grid positions. Naturally the divergence is zero with respect to itself. (b) Divergence matrix between histograms at n_1, n_2 for all combinations of n_1, n_2.

As shown in Figure 3(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 3(a), whereby varying the pitch d yields a peak in the overall temporal scene power observed $K^{fund} = \arg\max_d(g(d)^T A\, g(d))$ for $d_{min} \leq d \leq d_{max}$ and binary vector g such that $g_i(d) = \delta((i-n_{max}) \bmod d)$ where $1 \leq i \leq 2n_{max}-1$. Figure 3(c) shows how the scene's signal power peaks at a given value of d.

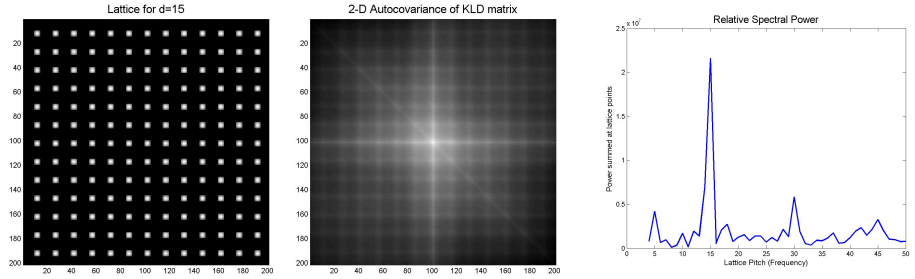

Figure 3: (a) Lattice for $d = 15$ generated by $g(d)g(d)^T$. Multiplying such a lattice by the autocovariance matrix in (b) for a range of d identifies the fundamental period. (b) Autocovariance of the Divergence matrix in Figure 2(b), showing the strong lattice structure corresponding to a dominant fundamental in the video sequence. (c) Relative spectral power of a spatial block in Figure 5 for values of d between 4 and 50. Fundamental at $d = 15$, gives a period of $15 \times 7.5s = 112.5s$ corresponding to the cycle of the junction.

4 Phase-Locked Loop

The spatio-temporal model described so far relies completely on its synchronization with scene activity to provide meaningful results. However, two problems are apparent. Firstly, the initial estimated periodicity of a block from training data may lack precision, and secondly, video data from the scene may be disrupted, corrupted, or some event in the scene may occur to radically alter the phase of the learned dynamic behaviour. In any case, our model may become de-synchronized, and it is highly desirable that it recover automatically from such situations. In [3], a Phase-Locked Loop (PLL) was used to recover the frequency and phase of oscillation in the characterization of human gait.

A PLL is a negative feedback servo mechanism encountered ubiquitously in electronic systems [1]. Implemented in digital or analogue hardware, or software it is usually constructed from the same functional building blocks as shown in Figure 4. It operates by synchronizing a local oscillator in both frequency and phase to a potentially noisy or variable frequency input signal, and is routinely used for demodulation, data recovery and frequency synthesis in communication and data systems. The behaviour of a PLL is largely controlled by the s- or z-plane transfer function of the loop filter.

Here we make use of its 'frequency filtering' property to solve the above mentioned short-comings of our model. By this we mean that short-term frequency variations (jitter) are rejected, such that the output adopts the long-term average of the input frequency and phase. We rely on the fact that although the purpose of our model is to detect unusual events in a scene, on average the behaviour will be largely consistent. The $\div N$ counter in Figure 4 causes the oscillator to run at a multiple of the periodicity. We implement the oscillator in software as a counter or 'phase accumulator', and the higher oscillator frequency yields a finer precision of output rate, and hence block periodicity.

Using the previously described KLD metric, a novel phase detector compares histograms at state l from T^{fund} most recent unseen frames in a circular fashion against the current model at all K^{fund} phases to determine the optimum

$$\Phi^{opt}_{h,v}(l) = \arg\min_j \Big(\sum_k D_{KL}[H_{h,v,k},\, S^{\bmod(j+l+k,\, K^{fund}_{h,v})}_{v,h}] + 1\Big) \qquad j,k \in \{1 \ldots K^{fund}_{h,v}\} \quad (9)$$

The phase exhibiting minimum KLD is considered the optimal target. Although potentially transiently in error, this measurement ensures that on average the model is synchronized to the scene permitting accurate event detection. The loop filter is configured as a Proportional plus Integral controller, the integral term minimizing zero-order phase error.

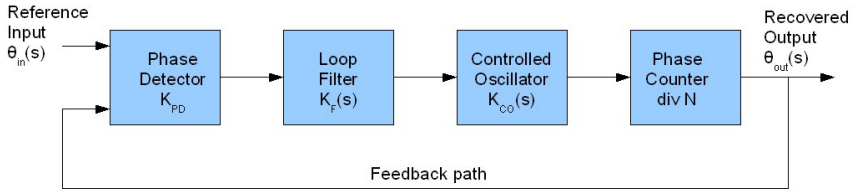

Figure 4: Typical PLL with s-plane transfer function $\frac{\theta_{out}(s)}{\theta_{in}(s)} = \frac{K_{PD}K_F(s)K_{CO}(s)/N}{1+K_{PD}K_F(s)K_{CO}(s)/N}$.

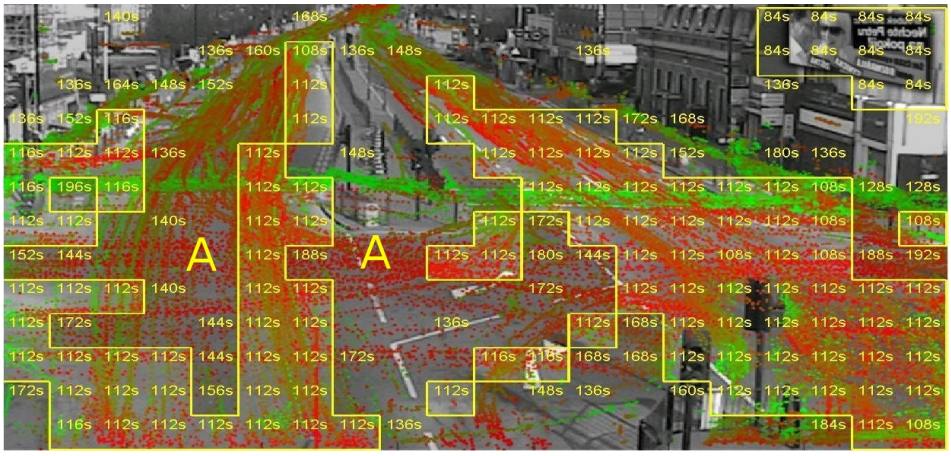

Figure 5: Periodicities from Scenario 1. Much of the junction shares the periodicity of the 112s traffic lights, whilst the rotating advertising board (top right) changes every 84s. Areas marked 'A' are aperiodic. Pedestrians tracks shown in green, vehicles in red.

5 Experiment

For experiments we chose two busy city-centre road junctions controlled by traffic lights. Each dataset had 30000 frames of 720×576 pixel colour video at 25Hz frame rate, yielding sequences of 20 minutes duration. The short-term background model was obtained as described using the method described in [12], 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 to yield the foreground mask. The model learns the 2D optical flow as the feature. The PLL loop filter uses a proportional gain of 3 and an integral gain of 0.02. We compare our method with an equivalent model having no temporal capability, derived by marginalizing out the time dimension of the Spatio-Temporal (S-T) model. As shown in Figure 7, in 4 examples the S-T model correctly identifies cars crossing the junction at completely the wrong

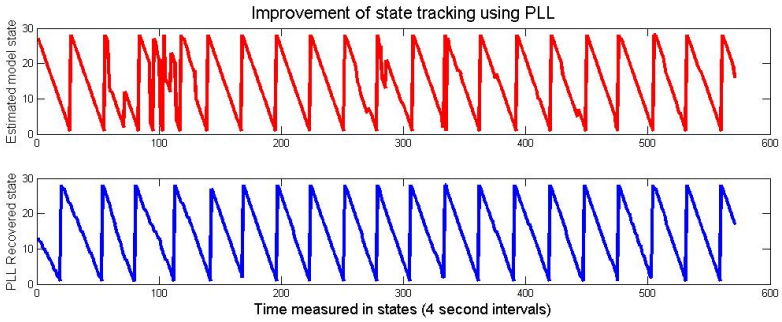

Figure 6: Benefit of PLL on model phase stability. Note its effective acquisition of correct phase (bottom) soon after t=0, and how it maintains a useful output phase even during corrupted input data (top) due to the 'inertia' caused by the loop filter.

time, whilst the model with No Temporal Processing (NTP) is frequently in error. Such marked improvement in detection comes exclusively from exploiting the learned optical flow information over the junction such that the expected instantaneous distribution is tightly coupled to the junction state cycle by the influence of the PLL. Computationally, the algorithm achieves 3FPS on a 2GHz PC after initial model building.

6 Conclusion and Future Work

We have demonstrated an algorithm capable of automatically learning multiple periodicities within a scene, such as exhibited at junctions controlled by traffic lights. It has been demonstrated by experiment that the method can be more discriminating with regard to activity of a periodic scene than a model which is oblivious to repeating temporal trends. The method is not tied to a particular feature, but may be employed wherever a histogram over feature(s) is available. By inclusion of a novel phase detector and control loop, it is possible to maintain model synchronization in the presence of noise. The logical progression of the technique is to permit automatic on-line update of histogram data for a block when its PLL is known to be in the 'locked' condition.

References

[1] R. E. Best. *Phase-Locked Loops: Theory, Design and Applications*. McGraw-Hill, 1993.

[2] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE PAMI*, 23(3):257–267, 2001.

[3] J.E. Boyd. Synchronization of oscillations for machine perception of gaits. *CVIU*, 96(1):35–59, October 2004.

[4] M. Casdagli. Recurrence plots revisited. *Physica D*, 108:12–44, 1997.

[5] Ross Cutler and Larry S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE PAMI*, 22(8):781–796, 2000.

[6] Fang Liu and Rosalind W. Picard. Finding periodicity in space and time. In *ICCV*, pages 376–383, 1998.

[7] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Computer Vision and Image Understanding*, pages 121–130, 1981.

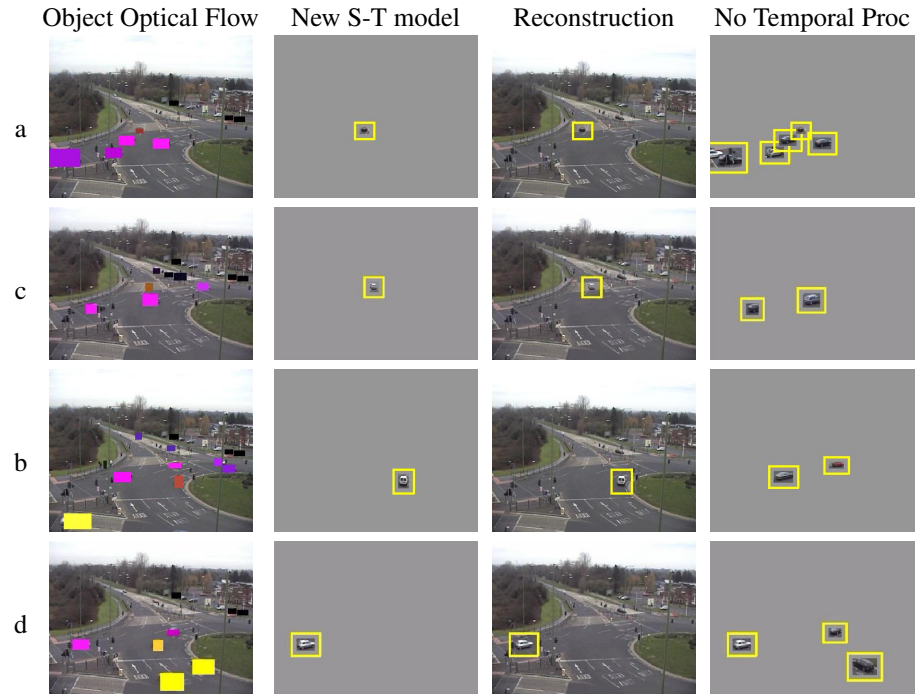

Figure 7: Comparison between new S-T model and one with No Temporal Processing (NTP). From left: Colour-coded optical flow, S-T model output, reconstruction from S-T model, NTP model output. (a,b): Vehicle wrongly crosses lights going from bottom to top, S-T finds it, but NTP only highlights cars behaving normally. (c): Car jumps red light by roundabout from bottom left, S-T model sees it, NTP only finds legal behaviour. (d): Car jumps red light from left, NTP model also highlights two other vehicles erroneously.

[8] T. Natschläger and B. Ruf. Spatial and temporal pattern analysis via spiking neurons. *Network: Computation in Neural Systems*, 9(3):319–332, 1998.

[9] J. Ng and S. Gong. On the binding mechanism of synchronised visual events. In *IEEE Workshop on Motion and Video Computing*, December 2002.

[10] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modelling human interactions. *IEEE PAMI*, 22(8):831–843, August 2000.

[11] Ramprasad Polana and Randal C. Nelson. Detection and recognition of periodic, nonrigid motion. *IJCV*, 23(3):261–282, 1997.

[12] D. Russell and S. Gong. Minimum cuts of a time-varying background. In *BMVC*, pages 809–818, Sep 2006.

[13] D. Russell and S.Gong. Multi-layered decomposition of recurrent scenes. In *ECCV*, Oct 2008.

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

[15] Martin Szummer. Temporal texture modeling. Technical Report 346, MIT Media Lab Perceptual Computing, 1995.

[16] T. Xiang and S. Gong. Beyond tracking: Modelling activity and understanding behaviour. *IJCV*, 67(1):21–51, 2006.