

Robust Complex Behaviour Modeling at 90Hz

Xiangyu Kong¹, Yizhou Wang¹, and Tao Xiang²

¹Nat'l Eng. Lab. for Video Technology, Cooperative Medianet Innovation Center
Key Lab. of Machine Perception (MoE), Sch'l of EECS, Peking University, Beijing, 100871, China
{kong, Yizhou.Wang}@pku.edu.cn

²Queen Mary, University of London, London E1 4NS, UK
t.xiang@qmul.ac.uk

Abstract

Modeling complex crowd behaviour for tasks such as rare event detection has received increasing interest. However, existing methods are limited because (1) they are sensitive to noise often resulting in a large number of false alarms; and (2) they rely on elaborate models leading to high computational cost thus unsuitable for processing a large number of video inputs in real-time. In this paper, we overcome these limitations by introducing a novel complex behaviour modeling framework, which consists of a Binarized Cumulative Directional (BCD) feature as representation, novel spatial and temporal context modeling via an iterative correlation maximization, and a set of behaviour models, each being a simple Bernoulli distribution. Despite its simplicity, our experiments on three benchmark datasets show that it significantly outperforms the state-of-the-art for both temporal video segmentation and rare event detection. Importantly, it is extremely efficient – reaches 90Hz on a normal PC platform using MATLAB.

Introduction

The past two decades have witnessed an accelerated expansion of closed-circuit television (CCTV) surveillance for public safety and security applications. This coincides with a growing interest in automatic visual analysis of object behaviours captured by surveillance videos in public spaces, with a particular focus on crowd behaviour modeling for rare/abnormal event detection. Existing approaches fall into two categories depending on what crowd events are of interest. The first category of approaches (Cong, Yuan, and Liu 2011; Lu, Shi, and Jia 2013; Zhao, Fei-Fei, and Xing 2011; Antic and Ommer 2011; Cui et al. 2011; Kratz and Nishino 2009; Kwon and Lee 2012; Mahadevan et al. 2010; Mehran, Oyama, and Shah 2009; Roshtkhari and Levine 2013; Saligrama and Chen 2012) seek to detect well-defined behaviour patterns that stand out in terms of both appearance and motion. For example, Fig. 1(a) shows that a person riding a bicycle is detected as an anomaly due to his distinctive appearance and motion (speed) among the normal walking pedestrians in the scene. In contrast, the second category of approaches (Li, Gong, and Xiang 2012; Hospedales et al.



Figure 1: Examples of simple and salient crowd events (a) vs. complex and subtle events (b).

2011; Hospedales, Gong, and Xiang 2012; Song et al. 2014; Wang and Mori 2009; Wang, Ma, and Grimson 2009; Zhou, Wang, and Tang 2012; Ricci et al. 2013) focus on more subtle and complex behavioural anomalies. An example is shown in Fig. 1(b) where a fire engine interrupts the normal traffic flow at a junction. This type of events are more difficult to detect, yet they are more common in reality. They are thus the focus of this paper.

Modeling complex crowd behaviour for detecting subtle events is challenging because both individual object behaviour and their behavioural context need to be modeled. In particular, the behaviour of an object involved in a complex and subtle event may look perfectly normal. It is abnormal only when put in context – it occurs in the wrong place and/or wrong time in relation to other objects in the scene. For example, when the fire engine in Fig. 1(b) moves horizontally, there are also vertical traffic which should not co-occur for obvious reasons. Therefore to detect such subtle events, one must model both object behaviour and their spatial, temporal and correlation behavioural context.

With dozens of objects to model at any given time as well as their behavioural context, existing approaches to complex crowd behaviour modeling rely on complex models with the hierarchical probabilistic topic models (PTMs) or other forms of graphical models being the most popular choice (Li, Gong, and Xiang 2012; Hospedales et al. 2011; Hospedales, Gong, and Xiang 2012; Wang and Mori 2009; Wang, Ma, and Grimson 2009; Zhou, Wang, and Tang 2012). Other non-parametric Bayesian models such as Dirichlet Process Mixture models (Emonet, Varadarajan, and Odobez 2014) are also considered. Most of these models are generative models; in contrast, (Ricci et al. 2013) presented a prototype learning framework with Earth Mover's Distance (EMD), and (Cheng, Chen, and Fang 2015) proposed to use hierarchical 3D features and Gaussian Process

Regression (GPR). The main limitations of these models are: (1) When a rare event occurs, it often only involves one or two objects with many more other objects in the scene behaving normally. Modeling all objects in a single model thus makes the model insensitive to these rare events, resulting in miss detections (Hospedales, Gong, and Xiang 2012). (2) A complex model with too many parameters can be sensitive to feature noise causing false alarms. (3) Existing models are typically computationally expensive. These limitations make them unsuitable for a practical application scenario whereby multiple video channels need to be processed simultaneously in real time with high detection accuracy and few false alarms.

In this paper, we propose a novel approach to complex crowd behaviour modeling, consisting of a set of particularly simple models which is sensitive to anomalies, robust against noise, and can be computed extremely fast. The approach is designed based on two principles: (1) Instead of using a single elaborate model, a set of simple models is preferred which are more flexible and efficient to compute. (2) Both object behaviour, and spatio-temporal behavioural context are quantized to further reduce model complexity and gain robustness against feature noise. More specifically, a complex dynamic scene is first decomposed into functional regions where spatial context is defined. The temporal behavioural context of the scene are also quantized into discrete states. Each region's behaviour is then represented as a Binarized Cumulative Directional (BCD) feature vector, which is much simpler than the representations used by previous works. Furthermore, we propose to learn spatial and temporal context with a novel joint iterative refinement process by maximizing the motion correlation across the functional regions at each temporal state. Finally each element of the indicator vector (a binary variable) is modeled as a simple Bernoulli distribution (much simpler than even a Gaussian) whose single parameter is dependent on the region (spatial context) and temporal state (temporal context). Extensive experiments are carried out on three benchmark datasets. The results suggest that our approach not only significantly outperforms the state-of-the-arts in terms of accuracy and robustness on rare event detection and temporal segmentation, it is also extremely efficient – it runs at about 90Hz on average on an ordinary PC using MATLAB.

Methodology

Binarized Cumulative Directional (BCD) Feature

An input video sequence is first divided into a sequence of N_T video clips, each of which consists of T frames. Each clip is further divided into a spatial-temporal grid (see Fig. 2), yielding N_c grid cells in each clip. We compute optical flows using the method in (Liu 2009), then quantize all the pixels' optical flow vectors into N_o directions. In each spatial-temporal grid cell, we first compute a Histogram of Optical Flow (HOF) descriptor of the cell by distributing all the pixels into N_b bins according to their N_o motion directions plus one bin to accommodate pixels with the motion magnitude being less than a predefined threshold (i.e. motionless pixels). An example of a 9-bin HOF descriptor is

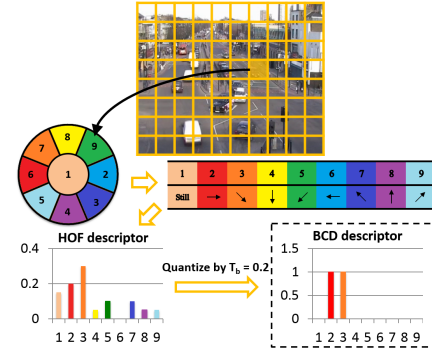


Figure 2: The proposed BCD feature representation shown in Fig. 2. Instead of directly using the continuous HOF descriptor as features, we further quantize the HOF descriptor into a binary motion direction indicator vector by thresholding each bin of the HOF by a threshold T_b , which gives the final Binarized Cumulative Directional (BCD) descriptor (see Fig. 2). By cumulating the quantized optical flows into a histogram representation both spatially and temporally, and further binarize it, the resulting BCD descriptor is extremely simple, which enables the subsequent development of simple and robust behaviour models. Note that our representation is also flexible for extension. For example, if speed is deemed critical for the definition of rare events, additional bins corresponding to flow vector magnitude can be readily augmented to the BCD descriptor.

Initialization of Spatial and Temporal Context

Spatial context initialization The spatial context is learned as semantic regions segmented from the scene (Li, Gong, and Xiang 2012). This dynamic scene decomposition is obtained using the spectral clustering method in (Zelnik-Manor and Perona 2004). More specifically, given N_c cells in each clip, a $N_c \times N_c$ affinity/similarity matrix A is constructed for all the clips in a training dataset with the similarity measure $A_{i,j}$ computed as:

$$A_{ij} = \begin{cases} \exp\left(-\frac{d(x_i, x_j)^2}{\sigma_i \sigma_j} - \frac{d(c_i, c_j)^2}{\sigma_x^2}\right), & \text{if } \|c_i - c_j\| \leq R; \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where c_i , and c_j are the image coordinates of the i -th cell and the j -th cell respectively; x_i , and x_j are obtained by concatenating the N_T BCD descriptors over the N_T clips in the training set; they are thus $N_b \times N_T$ dimensional feature vectors; $d()$ represents the cosine distance; σ_i and σ_j are the scaling factors of the feature vectors representing the i -th cell and the j -th cell; σ_x is the spatial scaling factor; R is the radius of a circle on the image plane within which similarity is computed. The objective of this model is to encourage cells that share similar motion representation and are close spatially to be grouped into the same semantic region. The affinity matrix A is then used as input for spectral clustering to obtain a semantic region segmentation. Note that the number of clusters/regions N_R is determined automatically by the clustering algorithm.

Temporal Context Initialization Temporal context is also initialized by clustering. For that, the Affinity Propaga-

tion algorithm proposed in (Frey and Dueck 2007) is adopted. Specifically, we concatenate the N_c BCD descriptors of all the cells from one frame together as the representation for this frame and then re-quantize it into a N_b dimensional BCD representation with the same threshold T_b . Again the adopted Affinity Propagation clustering method is capable of finding the optimal cluster number/temporal phases N_p automatically.

Context Refinement via Correlation Analysis

The initial spatial and temporal context are obtained independently from each other. In this section, they are further improved by a joint iterative refinement process which utilizes the third type of context, the correlation context. This context models the correlation between the object behaviour in different semantic regions.

Given an initial temporal phase segmentation, in each iteration of the joint iterative refinement process, a greedy search is performed to find an optimal set of phase boundaries within a fixed-size temporal window near each initial phase boundary. Formally, given an initial temporal phase boundary (frame index) B , the refined boundary \hat{B} is obtained by searching through a temporal window of $2W$ frames centred at B , and finding the new boundary as:

$$\hat{B} = \underset{b}{\operatorname{argmax}} \operatorname{Corr}(b) \quad (2)$$

where $B - W \leq b \leq B + W$ and $\operatorname{Corr}(b)$ measures the correlation strength among the behaviours of different semantic regions given a candidate phase boundary b . The correlation strength is computed using Multi-view Canonical Correlation Analysis (MCCA) (Gong et al. 2014). First, for each frame and each region, a regional behaviour feature vector x_r is computed for the r -th region by averaging all the cells' BCD features, normalize it and binarize it again into another N_b dimensional BCD vector using the same threshold T_b . Then for each frame, the MCCA model projects x_r from different regions in to a single embedding space. Finally the average cosine similarity among different regions in the embedding space is computed as the measure of correlation strength between regional behaviours. For robustness, we consider all frames within a window of $2W$ frames centred at b , and take the maximum as the final value of $\operatorname{Corr}(b)$:

$$\operatorname{Corr}(b) = \max_{b-W \leq i \leq b+W} \operatorname{corr}(i) \quad (3)$$

where

$$\operatorname{corr}(i) = 2 / (N_R(N_R - 1)) \sum_{r1=1}^{N_R} \sum_{r2=1}^{N_R} \operatorname{sim}_c(x_{r1}, x_{r2}) \quad (4)$$

where N_R is the number of semantic regions, and sim_c is a function that computes cosine similarity of two regional behaviour vectors x_{r1} and x_{r2} in the multi-view CCA space.

After the initial temporal context is refined, with a set of updated temporal phase boundaries, a new scene decomposition is obtained with the same clustering method as in initialization, but an input without any cross-phase-boundary spatial-temporal cubes. The intuition is that with a more accurate phase boundary, the spatial-temporal cubes within each phase can also be purer and thus enable better scene decomposition. Now with the new spatial context, temporal context refinement is performed following the same greedy

search algorithm but with different values of $\operatorname{Corr}(b)$. This iterative process terminates given no changes in both spatial and temporal context. In our experiments, we found that the refinement process converges after $2 \sim 3$ iterations.

Behaviour Modeling

After spatial and temporal context modeling, we obtain a set of semantic regions and an assignment of each frame to a temporal phase. Each region is now represented as a BCD vector x_r and each dimension of this vector is a binary variable. For behaviour modeling, instead of using a single model for all regional behaviours over different temporal phases, we model each element of each regional behaviour vector under each temporal phase as a simple Bernoulli distribution. Each distribution thus represents how likely the corresponding motion pattern in that direction takes place in each region, under each phase.

Let the e -th element of the BCD descriptor of the r -th region under the p -th phase be $Q_{e,r,p}$, we assume that its value follows a Bernoulli distribution with a parameter of $\mu_{e,r,p}$:

$$Q_{e,r,p} \sim \operatorname{Bern}(\mu_{e,r,p}) \quad (5)$$

Then for a specific dynamic scene, the complex behaviours are modeled as an ensemble of Bernoulli distributions $\{Q_{e,r,p}\}$, $e = 1 \dots N_b$, $r = 1 \dots N_r$, $p = 1 \dots N_p$, where N_p is the number of temporal clusters/phases, N_r is the number of semantic regions, N_b is the number of elements in the BCD representation. There is thus a total of $N_r \times N_p \times N_b$ models, which are typically in the order of hundreds.

Learning Learning each model involves estimating each Bernoulli distribution's parameter $\{\mu_{e,r,p}\}$, which can be estimated with maximum likelihood estimator (MLE)

$$\mu_{e,r,p} = \frac{1}{N} \sum_{n=1}^N Q_{e,r,p,n} \quad (6)$$

where n is the frame index and N is the total number of training frames. It is thus N summations followed by a division and can be computed very efficiently.

Temporal segmentation and rare event detection Once trained, the model can be used to infer the temporal phases for temporal segmentation. Specifically, for the test frames, the spatial context is fixed and the temporal phase is simply assigned to the nearest temporal cluster. For rare event detection, given an input BCD descriptor of region r under phase p , a natural measure of abnormality S_a for that region can be computed as:

$$S_a = \min_{r=1 \dots N_r} \prod_{e=1}^{N_b} \tilde{\mu}_{e,r,p} \quad (7)$$

where

$$\tilde{\mu}_{e,r,p} = \begin{cases} \mu_{e,r,p} & \text{if } Q_{e,r,p} = 1; \\ 1 - \mu_{e,r,p} & \text{otherwise.} \end{cases} \quad (8)$$

The frame/clip is detected as containing rare events if $S_a < T_a$ where T_a is a threshold. It is also very easy to identify which region caused the rare event: all regions with $\prod_{e=1}^{N_b} \tilde{\mu}_{e,r,p} < T_a$ are the regions involved in the rare event.

Experiments

Datasets and Settings

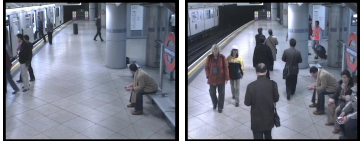
QMUL Junction Dataset (Hospedales, Gong, and Xiang 2012) This dataset consists of a video sequence recorded



(a) Phase 1 (b) Phase 2 (c) Phase 3
Figure 3: Temporal phases for the QMUL dataset



(a) Phase 1 (b) Phase 2
Figure 4: Temporal phases for the MIT dataset



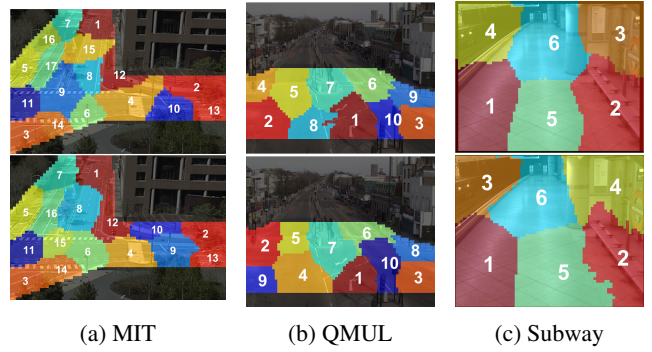
(a) Phase 1 (b) Phase 2
Figure 5: Temporal phases for the Subway dataset

at 25Hz for 1h at a busy urban intersection including vertical, horizontal and turning traffic flows. The frame size in this dataset are 360×288 . The traffic flows are well regulated by traffic lights (temporal context). There are three temporal phases learnt in this video, of which the dominant motion patterns are respectively vertical traffic flow, right-to-left flow and left-to-right flow. Examples of the three temporal phases are shown in Fig. 3.

MIT Traffic Dataset (Wang, Ma, and Grimson 2007) It contains 1.5h of surveillance videos with a frame size of 720×480 and a frame rate of 30Hz. Similar to QMUL Junction, this benchmark was also recorded at a traffic junction, although the cars were driven on different sides of the roads being in the US rather than the UK. However, since the traffic flow is much less busy, the regulation of traffic light for this dataset is much weaker than that of QMUL Junction. Two temporal phases are learnt in this video, which are vertical flow and horizontal flow (see Fig. 4).

Subway Dataset (Hospedales, Gong, and Xiang 2012) This dataset contains a 17-minute sequence selected from the UK Home Office i-LIDS dataset. The frame size is 720×576 . Different from the other two traffic scene datasets, this dataset features a subway platform scene. It is also noted that the behaviours of subway passengers are much diverse compared with that of vehicles at a traffic junction; Although there is no temporal phases of fixed length as in the three traffic scenes, the object behaviours in this scene are still weakly governed by the train arrival and departure events. In particular, this complex scene has two temporal phases: passengers boarding when train arrives and stay on the track, and train departing with passengers entering/leaving the platform or sitting down (see Fig. 5).

Settings Following the setting in (Hospedales, Gong, and Xiang 2012), the two traffic videos were cut to 1-second clips while the subway video was cut to clips of 10 seconds. For the QMUL and MIT datasets, the training set consists of normal clips from the first 40 minutes of the video, and



(a) MIT (b) QMUL (c) Subway
Figure 6: Scene decomposition before (top row) and after (bottom row) refinement

Ours (refined)	Ours (initial)	EMD- L_1 linear	Cas-pLSA	MCTM	DDP-HMM
95.00 %	92.50%	92.31%	89.74%	51.79%	87.18%

Table 1: Temporal segmentation result on QMUL Junction the test set contains the remaining clips for rare event detection. For the Subway dataset, we used the first 5 minutes for training and the rest frames of the video for testing. In all three datasets, optical flow extraction was performed every 4 frames. We also averaged every 4 consecutive optical flow fields to make the optical flow more robust to noise. The grid cell size was set to 5×5 for QMUL and MIT, and 15×15 for Subway since the object sizes are larger in Subway. As the motion patterns of QMUL and MIT are more complex, the optical flow was quantized into $N_o = 8$ directions resulting in a $N_b = 9$ dimensional BCD descriptor. For the simpler motion patterns in Subway, we used a 5 dimensional BCD descriptor. The threshold T_b was set to 0.2 for all three datasets and its effect is analysed later. For computing the affinity matrix for scene decomposition, the radius R (see Eq. (1)) was set to 15 for QMUL and 8 for MIT, and 10 for Subway reflecting the scales of objects in each scene.

Spatial Segmentation

Scene decomposition was performed for learning the spatial context resulting in 17, 10, and 6 semantic functional regions for the MIT, QMUL, and Subway scenes respectively. Fig. 6 compare the scene decomposition results obtained by initial clustering and after joint refinement. It can be seen that the initial segmentation semantic regions are less accurate (e.g. the boundary between Region 7 and 1 in the QMUL scene, and 4 and 6 in the Subway scene). After refinement, the semantic regions are clearly improved with finer and more accurate boundaries. Notice that in the MIT and QMUL scenes, a mask was set manually to eliminate areas irrelevant to the junction zones (e.g. the building in the MIT scene).

Temporal Segmentation

In this experiment, we evaluate the accuracy of the learned temporal context by evaluating the temporal segmentation accuracy. Following the same setting as in (Ricci et al. 2013), temporal segmentation was evaluated on the first 1200 frames of the QMUL Junction dataset with manually labelled ground truth. The result is reported at the clip level, that is, we take the dominant phase label of each clip as the clip’s phase label.

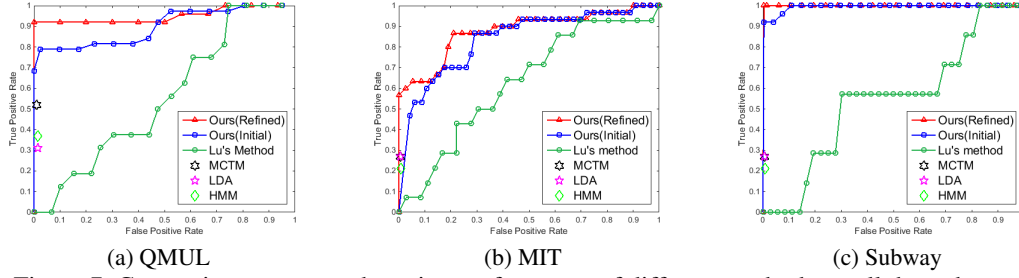


Figure 7: Comparing rare event detection performance of different methods on all three datasets

QMUL Junction	Ours (refined)	Ours (init)	MCTM	Lu's Method	HMM	Total
Brk Red Light	13	12	4	0	3	13
Illegal U-Turn	14	10	5	0	1	15
Jaywalking	1	1	1	0	0	1
Dr Wrong Way	15	15	12	0	12	15
Unusual Turns	6	6	6	0	4	10
Uninteresting	6	11	27	55	35	2663
Overall TPR	89%	81%	52%	0%	37%	
Overall FPR	0.2%	0.4%	1.0%	2.1%	1.4%	

Table 2: The top 2% rarest clip types discovered by each model for the QMUL Junction dataset.

MIT Traffic	Ours (refined)	Ours (init)	MCTM	Lu's Method	HMM	Total
Jay-walking	7	7	4	1	4	20
Out of Lane	1	1	1	0	0	1
Near Collision	3	2	3	0	2	8
Uninteresting	0	1	3	10	5	510
Overall TPR	38%	34%	27%	27%	21%	
Overall FPR	0%	0.2%	0.6%	0.6%	0.9%	

Table 3: The top 2% rarest clip discovered by each model for the MIT Traffic dataset.

Subway	Ours (refined)	Ours (init)	MCTM	Lu's Method	HMM	Total
Contraflow	2	2	2	0	0	2
Uninteresting	36	36	36	38	38	1155
Overall TPR	100%	100%	100%	0%	0%	
Overall FPR	3.1%	3.1%	3.1%	3.3%	3.3%	

Table 4: The top 2% rarest clips discovered by each model for the Subway Platform dataset.

Table 1 gives a comparison between our temporal context model (with and without refinement) and a number of existing methods. Among the compared models, Cas-pLSA (Li, Gong, and Xiang 2012), DDP-HMM (Kuettel et al. 2010), and MCTM (Hospedales, Gong, and Xiang 2012) are two-layer hierarchical generative models, which have shown to give better temporal segmentation results than conventional one-layer models such as HMM, LDA and pLSA (Li, Gong, and Xiang 2012); in contrast, the EMD- L_1 linear model (Ricci et al. 2013) is a discriminative model. Note that, among the compared methods, only Cas-LDA uses spatial segmentation information as we do, but has a much higher

computational cost. It can be seen from Table 1 that, compared with the previous methods, (i) even our simple initial segmentation achieves a better accuracy, which justifies the robustness of the proposed BCD descriptor and the adopted clustering method. (ii) The performance is significantly better than a number of complex hierarchical generative models including MCTM and DDP-HMM. (iii) The joint refinement process does improve the accuracy of temporal segmentation (92.50% increased to 95.00%).

Online Detection of Rare Events

For rare event detection, we followed exactly the same setting as in (Hospedales, Gong, and Xiang 2012) and also used the ground truth provided by the authors. For evaluation metrics, we set different threshold values to the rare event detection score T_a and plot ROC curves. We also plot ROC curves for those compared methods that we have codes. For others, in particular (Hospedales, Gong, and Xiang 2012), we follow their setting and report the rare event detection accuracy, measured by TPR (true positive rate) and FPR (false positive rate), by examining the top 2% most rare clips. This thus corresponds to a single point on the ROC curve.

The comparative results in TPR and FPR are shown in Tables 2, 3 and 4 for the three datasets respectively, which also show the different types of rare events defined in the three datasets (provided by (Hospedales, Gong, and Xiang 2012)). The ROC curves are shown in Fig. 7. The compared methods include MCTM (Hospedales, Gong, and Xiang 2012), the sparse coding model in (Lu, Shi, and Jia 2013) (referred to as "Lu's method" in Tables 2-4 and Fig. 7) and HMM. Note that among them, (Lu, Shi, and Jia 2013) was designed for detecting simple and salient events (e.g. Fig. 1(a)) rather than those complex and subtle ones in the three datasets. The HMM results are from (Hospedales, Gong, and Xiang 2012).

From the results, we can see clearly that our method significantly outperforms all other compared methods on accuracy and robustness. For example, compared to the strongest competitor MCTM, we have a quarter of the FPR (0.2% vs. 1.0%) whilst almost double the detection rate (89% vs. 52%) on the QMUL dataset. On the MIT dataset, the improvement is also large: with 0 false alarm, we obtain a 38% TPR, whilst MCTM can only manage 27% TPR with 0.6% FPR (3 false alarms). On the Subway data, the number of rare events is very small, so both our model and MCTM achieved 100% detection. However, it is noted that these two true positive rare clips are ranked as the top 2 rarest by our model. In other words, our model achieves a perfect rare

event detection on this dataset, as shown by the ROC curve of our method in Fig. 7(c). The result also show that with simple one-layer models such as HMM, the performance is even worse. Furthermore, Fig. 7 shows that the result of (Lu, Shi, and Jia 2013) is close to random guess (which will give a diagonal straight line as its ROC curve), indicating that it is unable of modeling challenging complex crowd behaviours. Tables 2-4 and Fig. 7 also show that, with the joint context refinement (Ours(Refined)), the rare event detection performance is clearly better that of our model without the refinement (Ours(Initial)), demonstrating the effectiveness of context refinement via correlation analysis. Some qualitative results of our model are shown in Fig. 8.

Table 5 shows some other recently reported results on the QMUL Junction dataset. Note that these results are not directly comparable to those in Table 2, because they were obtained under different settings. In particular, for reasons not explicitly detailed, all three compared works, Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) (Jouneau and Carincotte 2011), Integrated Probabilistic Latent Sequential Motifs (IPLSM) (Chockalingam, Emonet, and Odobez 2013), and Gaussian Process Regression (GPR) (Cheng, Chen, and Fang 2015) chose to detect only a subset of the rare events listed in Table 2. Nevertheless, it is obvious that, even when evaluated on detecting a larger variety of rare events, our model yields better performance. In particular, the TPRs of HDP-HMM (58%) and IPLSM (69%) are much lower than ours (89%). Compared with GPR (Cheng, Chen, and Fang 2015) which reported an AUROC (area under ROC) of 80.90%, our method achieves an AUROC of 95.07% (Fig. 7(a)) on all types of events and 94.71% on the two types in (Cheng, Chen, and Fang 2015).

Beyond Complex Events Although our model is primarily designed for modeling complex behaviour for detecting subtle anomalies, this does not restrict it from being applied to simple crowd behaviour for salient anomaly detection. To demonstrate this, our method is tested on the widely used UCSD Ped1 crowd anomaly dataset (Li, Mahadevan, and Vasconcelos 2014) containing simple and salient abnormal events (see Fig. 1(a) for an example frame). Particularly, the anomalies were usually caused by speed deviation from the norm in this dataset (too fast or slow). To capture object speed information, our BCD descriptor is extended by adding 4 more bins to encode flow vector magnitude. Furthermore, no temporal segmentation is performed since there is no obvious temporal phase in this sequence. Our method achieves an AUROC of 88.10% and an EER (Equal Error Rate) of 16.67% at frame level, which is comparable to the best results reported so far in (Lu, Shi, and Jia 2013) (91.8%/15%).

Sensitivity Analysis of Parameter T_b In our experiments, the binarization threshold T_b was set empirically to 0.2 for all datasets. Here we evaluate the effects of different values of T_b on the rare event detection performance of our model. Fig. 9 shows that our model is insensitive to the choice of T_b when its value is between 0.1 and 0.4.

Running Cost Besides detection accuracy, we also evaluated the computational efficiency. The speed of online detection of our method can reach 91.47Hz while the MCTM



Figure 8: Examples of rare event detection by our model with the semantic regions contributed to the detection of the event highlighted. From top to bottom: (a) Near Collision and Jay-walking; (b) Illegal U-Turn and Break Red Light; (c) The two contraflow events.

Abnormality	HDP-HMM		IPLSM		GPR
	GT	Detected	GT	Detected	
U-turn	11	3	10	7	-
Drive Wrong Way	16	11	-	-	-
Jaywalking	13	9	-	-	✓
Disruption*	-	-	6	4	✓
Uninteresting	-	8	-	-	-
AUROC	-	-	-	-	80.90%
TPR	-	58%	-	69%	-

Table 5: Performance of other related works on the QMUL Junction dataset. *The disruption event is equivalent to the “Break Red Light” event in Table 2

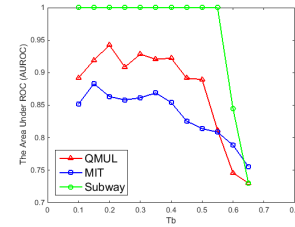


Figure 9: The rare event detection performance given different BCD binarization threshold T_b

model (Hospedales, Gong, and Xiang 2012) gives 9.18Hz and (Lu, Shi, and Jia 2013) reaches 227.37Hz, when they were run on the same/similar platform. Our model is also more than 100 times faster than that of (Cheng, Chen, and Fang 2015). Note that the model in (Lu, Shi, and Jia 2013) is indeed faster; however its detection performance on complex and subtle event detection is too poor to be useful (see Tables 2-4). Note that our model spends almost 99% of the computation on optical flows, whilst the cost of the rest is almost negligible. This suggests that when a GPU based optical flow implementation is used, the speed of our model can be further improved.

Conclusion

We have proposed a novel approach to complex behaviour modeling based on simple features, joint and iterative spatial, temporal and correlation context modeling, and a set of extremely simple Bernoulli-distribution-based behaviour models. Despite its simplicity, our experiments on three benchmark datasets show that it significantly outperforms the state-of-the-arts for both temporal video segmentation and rare event detection.

Acknowledgements

We would like to thank for support from the following research grants 973-2015CB351800, the Okawa Foundation Research Grant, NSFC-61272027, NSFC-61231010, NSFC-61527804, NSFC-61421062, NSFC-61210005.

References

- Antic, B., and Ommer, B. 2011. Video parsing for abnormality detection. In *ICCV*, 2415–2422. IEEE.
- Cheng, K.; Chen, Y.; and Fang, W. 2015. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In *CVPR*, 2909–2917.
- Chockalingam, T.; Emonet, R.; and Odobez, J. 2013. Localized anomaly detection via hierarchical integrated activity discovery. In *AVSS*, 51–56. IEEE.
- Cong, Y.; Yuan, J.; and Liu, J. 2011. Sparse reconstruction cost for abnormal event detection. In *CVPR*, 3449–3456. IEEE.
- Cui, X.; Liu, Q.; Gao, M.; and Metaxas, D. N. 2011. Abnormal detection using interaction energy potentials. In *CVPR*, 3161–3167. IEEE.
- Emonet, R.; Varadarajan, J.; and Odobez, J. 2014. Temporal analysis of motif mixtures using dirichlet processes. *TPAMI*.
- Frey, B. J., and Dueck, D. 2007. Clustering by passing messages between data points. *Science* 315:972–976.
- Gong, Y.; Ke, Q.; Isard, M.; and Lazebnik, S. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV* 106(2):210–233.
- Hospedales, T.; Li, J.; Gong, S.; and Xiang, T. 2011. Identifying rare and subtle behaviors: A weakly supervised joint topic model. *TPAMI* 33(12):2451–2464.
- Hospedales, T.; Gong, S.; and Xiang, T. 2012. Video behaviour mining using a dynamic topic model. *IJCV* 98(3):303–323.
- Jouneau, E., and Carincotte, C. 2011. Particle-based tracking model for automatic anomaly detection. In *ICIP*, 513–516. IEEE.
- Kratz, L., and Nishino, K. 2009. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 1446–1453. IEEE.
- Kuettel, D.; Breitenstein, M. D.; Van Gool, L.; and Ferrari, V. 2010. What’s going on? discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 1951–1958. IEEE.
- Kwon, J., and Lee, K. M. 2012. A unified framework for event summarization and rare event detection. In *CVPR*, 1266–1273.
- Li, J.; Gong, S.; and Xiang, T. 2012. Learning behavioural context. *IJCV* 97(3):276–304.
- Li, W.; Mahadevan, V.; and Vasconcelos, N. 2014. Anomaly detection and localization in crowded scenes. *TPAMI* 36:18–32.
- Liu, C. 2009. *Beyond pixels: exploring new representations and applications for motion analysis*. Ph.D. Dissertation, Citeseer.
- Lu, C.; Shi, J.; and Jia, J. 2013. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2720–2727. IEEE.
- Mahadevan, V.; Li, W.; Bhalodia, V.; and Vasconcelos, N. 2010. Anomaly detection in crowded scenes. In *CVPR*, 1975–1981. IEEE.
- Mehran, R.; Oyama, A.; and Shah, M. 2009. Abnormal crowd behavior detection using social force model. In *CVPR*, 935–942. IEEE.
- Ricci, E.; Zen, G.; Sebe, N.; and Messelodi, S. 2013. A prototype learning framework using emd: Application to complex scenes analysis. *TPAMI* 35(3):513–526.
- Roshtkhari, M. J., and Levine, M. D. 2013. Online dominant and anomalous behavior detection in videos. In *CVPR*, 2611–2618. IEEE.
- Saligrama, V., and Chen, Z. 2012. Video anomaly detection based on local statistical aggregates. In *CVPR*, 2112–2119. IEEE.
- Song, L.; Jiang, F.; Shi, Z.; Molina, R.; and Katsaggelos, A. K. 2014. Toward dynamic scene understanding by hierarchical motion pattern mining. *IEEE Transactions on Intelligent Transportation Systems* 15(3).
- Wang, Y., and Mori, G. 2009. Human action recognition by semilattent topic models. *TPAMI* 31(10):1762–1774.
- Wang, X.; Ma, X.; and Grimson, E. 2007. Unsupervised activity perception by hierarchical bayesian models. In *CVPR*, 1–8. IEEE.
- Wang, X.; Ma, X.; and Grimson, W. E. L. 2009. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *TPAMI* 31(3).
- Zelnik-Manor, L., and Perona, P. 2004. Self-tuning spectral clustering. In *NIPS*, 1601–1608.
- Zhao, B.; Fei-Fei, L.; and Xing, E. P. 2011. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, 3313–3320. IEEE.
- Zhou, B.; Wang, X.; and Tang, X. 2012. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *CVPR*.