A Markov Clustering Topic Model for Mining Behaviour in Video

Timothy Hospedales, Shaogang Gong and Tao Xiang
School of Electronic Engineering and Computer Science
Queen Mary University of London, London E1 4NS, UK
{tmh,sgg,txiang}@dcs.qmul.ac.uk

Abstract

*This paper addresses the problem of fully automated mining of public space video data. A novel Markov Clustering Topic Model (MCTM) is introduced which builds on existing Dynamic Bayesian Network models (*e.g. *HMMs) and Bayesian topic models (*e.g. *Latent Dirichlet Allocation), and overcomes their drawbacks on accuracy, robustness and computational efficiency. Specifically, our model profiles complex dynamic scenes by robustly clustering visual events into activities and these activities into global behaviours, and correlates behaviours over time. A collapsed Gibbs sampler is derived for offline learning with unlabeled training data, and significantly, a new approximation to online Bayesian inference is formulated to enable dynamic scene understanding and behaviour mining in new video data online in real-time. The strength of this model is demonstrated by unsupervised learning of dynamic scene models, mining behaviours and detecting salient events in three complex and crowded public scenes.*

1. Introduction

The proliferation of cameras in modern society is producing an ever increasing volume of video data which is thus far only weakly and inefficiently exploited. Video data is frequently stored passively for record purposes. If the video data is to be actively analyzed, expert knowledge about the scene and laborious manual analysis and labeling of the dataset is required. There has been some effort on developing methods for automatically learning visual behaviour models without human expertise or labour, and using such models to cluster and classify video data, or to screen for interesting events automatically [7, 16, 10, 15]. This is a challenging problem for various reasons. Classes of 'subjectively interesting behaviour' to a user can be defined task-specifically by various factors: the activity of a single object over time (*e.g.* its track), the correlated spatial state of multiple objects (*e.g.* a piece of *abandoned* luggage is defined by separation from its owner) or both spatial and temporal considerations (*e.g.* traffic flow at an intersection might have a particular order dictated by the lights). The spatial or temporal range over which correlations might be important may be short or long. Typical public scenes are crowded, creating difficulties for segmentation or tracking. In this paper we introduce a new model to address the problem of unsupervised mining of multi-object spatio-temporal behaviours in crowded and complex public scenes by discovering underlying spatio-temporal regularities in video so as to detect irregular patterns that can be consistently interpreted as 'salient behaviours' by human users. A system based on our model can answer queries such as: "Give me a summary of the typical activities and scene behaviour in this scene" and "Show me the (ranked) most interesting (irregular) events happened in the past 24 hours".

1.1. Related Work

Recent research on dynamic scene understanding has broadly fallen into object-centric tracking based and non-object-centric statistical approaches. Tracking based approaches [8, 14] clearly represent the spatial state of visual objects over time. This allows them to easily model behaviours like typical flows of traffic, and detect unusual events such as u-turns. Such models only work well if complete tracks can be reliably obtained in training and test data. For improving robustness to track failures, non-parametric representations of track statistics have been exploited [1, 12]. However, a major limitation of tracking based approaches is the difficulty in modeling behaviours characterized by coordinated activity of multiple objects.

To improve robustness and enable multi-object spatio-temporal correlation modeling, statistical methods have been devised to process directly on quantized pixel data [16, 13] or other low level 'event' features in video [15, 9]. These methods typically employ a Dynamic Bayesian Network (DBN) such as a Hidden Markov Model (HMM) [4, 15], or a probabilistic topic model (PTM) [9, 13] such as Latent Dirichlet Allocation (LDA) [3] or extensions. DBNs are natural for modeling dynamics of behaviour, and with hierarchical structure also have the potential to perform clustering of both activities and behaviours simultaneously.

Nevertheless, modeling the temporal order of visual events explicitly is risky, because noise in the event representation can easily propagate through the model, and be falsely detected as salient [9, 13]. To overcome this problem, PTMs were borrowed from text document analysis [9, 13]. These "bag of words" models represent feature co-occurrence, completely ignoring temporal order information. Therefore robustness to noise is at the cost of discarding vital dynamic information about behaviour. PTMs also suffer from ambiguity in determining the temporal window extent for collecting the bag of words. Large windows risk overwhelming behaviours of shorter duration, and small windows risk breaking up behaviours arbitrarily. This is especially damaging since correlation between bags is not modeled.

1.2. Our Approach

In this paper, a novel Markov Clustering Topic Model (MCTM) is introduced which builds on the strength of existing DBNs and PTMs, but crucially is able to overcome their drawbacks on accuracy, robustness and computational efficiency. In particular, the model makes two important novel contributions to LDA: (1) Hierarchical modeling, allowing simple actions to be combined into complex global behaviours; and (2) temporal modeling, enabling the correlation of different behaviours over time to be modeled. By introducing a Markov chain to model behaviour dynamics, this model defines a DBN generalization of LDA. This gains strength in representing temporal information, while being robust to noise due to its bag of words modeling of visual features. Learning from unlabeled training data is performed offline with Gibbs sampling; and a novel Bayesian inference algorithm enables dynamic scene understanding and behaviour mining in new video data online and in real-time where existing approaches fail [13, 9].

2. Spatio-Temporal Video Mining

2.1. Video Representation

We wish to construct a generative model capable of automatic mining and screening irregular spatio-temporal patterns as 'salient behaviours' in video data captured from single fixed cameras monitoring public spaces with people and vehicles at both far and near-field views (see Sec. 4.1). These camera views contain multiple groups of heterogeneous objects, occlusions, and shadows. Local motions are used as low level features. Specifically, a camera view is divided into $C \times C$ pixel-cells, and optical flow computed in each cell. When the magnitude of the flow is greater than a threshold Th_o, the flow is deemed reliable and quantized into one of four cardinal directions. A discrete visual event is defined based on the position of the cell and the motion direction. For a 320×240 video frame and with cell size of 10×10, a total of 3072 different discrete visual events may occur in combination. For visual scenes where objects may remain static for sustained period of time (*e.g.* people waiting for trains at a underground station), we also use background subtraction to generate a fifth – stationary foreground pixel – state for each cell, giving a visual event codebook size of 3840. This illustrates the flexibility of of our approach: it can easily incorporate other kinds of 'meta-data' features that may be relevant in a given scene. The input video is uniformly segmented into one-second clips, and the input to our model at second t is the bag of all visual events occurring in video clip t, denoted as $\mathbf{x}_t$.

2.2. Markov Clustering Topic Model (MCTM)

Standard LDA [3] (see Fig. 1(a)) is an unsupervised learning model of text documents $\mathbf{x}_m$, $m = 1..M$. A document m is represented as a bag of $i = 1..N_m$ unordered words $x_{i,m}$, each of which is distributed according to a multinomial distribution $p(x_{i,m}|\phi_{y_{i,m}})$ indexed by the current topic of discussion $y_{i,m}$. Topics are chosen from a per-document multinomial distribution θ_m. Inference of latent topics $\mathbf{y}$ and parameters θ and ϕ given data $\mathbf{x}_m$ effectively clusters co-occurring words into topics. This statistical topic based representation of text documents can facilitate, e.g., comparison and searching. For mining behaviours in video, we consider that visual events correspond to words, simple actions (co-occurring events) to topics, and complex behaviours (co-occurring actions) to document categories.

We model the occurrence of a sequence of clips/documents $\mathbf{X} = \{\mathbf{x}_t\}$ where $t = 1..T$ as having a three layer latent structure: events, actions and behaviours, as illustrated by the graphical model in Fig. 1(b). The generative model is defined as follows: Suppose the data contains T clips, each of which exhibits a particular category of behaviour, represented by z_t. The behaviour category z_t is assumed to vary systematically over time from clip to clip according to some unknown multinomial distribution, $p(z_t|z_{t-1}, \psi)$ (denoted Multi$(\cdot)$). Within each clip t, N_t simple actions $\{y_{i,t}\}_{i=1}^{N_t}$ are chosen independently based on the clip category, $y_{i,t} \sim p(y_{i,t}|z_t, \theta)$. Finally, each observed visual event $x_{i,t}$ is chosen based on the associated action $y_{i,t}$, $x_{i,t} \sim p(x_{i,t}|y_{i,t}, \phi)$. All the multinomial parameters $\{\phi, \psi, \theta\}$ are treated as unknowns with Dirichlet priors (denoted Dir$(\cdot)$). The complete generative model is specified by:

$$\begin{aligned} p(\psi_z|\gamma) &= \text{Dir}(\psi_z; \gamma), \\ p(\theta_z|\alpha) &= \text{Dir}(\theta_z; \alpha), \\ p(\phi_y|\beta) &= \text{Dir}(\phi_y; \beta), \\ p(z_{t+1}|z_t, \psi) &= \text{Multi}(z_t; \psi_{z_t}), \\ p(y_{i,t}|z_t, \theta) &= \text{Multi}(y_{i,t}; \theta_{z_t}), \\ p(x_{i,t}|y_{i,t}, \phi) &= \text{Multi}(x_{i,t}; \phi_{y_{i,t}}). \end{aligned}$$

The full joint distribution of variables $\{\mathbf{x}_t, \mathbf{y}_t, z_t\}_1^T$ and pa-

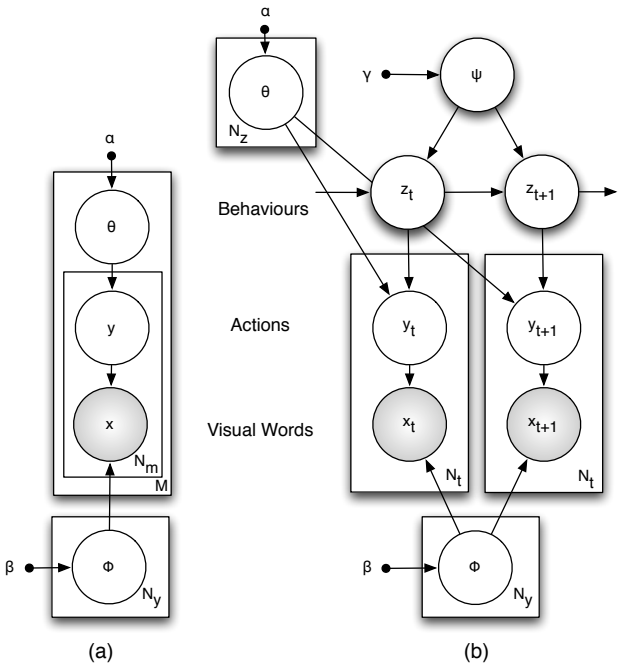

Figure 1. Graphical models representing: (a) Standard LDA model [3], (b) Our MCTM model.

rameters θ, ϕ, ψ given the hyper-parameters α, β, γ is:

$$p(\{\mathbf{x}_t, \mathbf{y}_t, z_t\}_1^T, \phi, \psi, \theta | \alpha, \beta, \gamma) = p(\phi|\beta)p(\psi|\gamma)p(\theta|\alpha) \cdot \prod_t \left(\prod_i p(x_{i,t}|y_{i,t})p(y_{i,t}|z_t) \right) p(z_t|z_{t-1}). \quad (1)$$

2.3. Model Learning

As for LDA, exact inference in our model is intractable, but it is possible to derive a collapsed Gibbs sampler [5] for approximate MCMC learning and inference. The Dirichlet-Multinomial conjugate structure of the model allows the parameters $\{\phi, \theta, \psi\}$ to be integrated out automatically in the Gibbs sampling procedure. The Gibbs sampling update for the action $y_{i,t}$ is derived by integrating out the parameters ϕ and θ in its conditional probability given the other variables:

$$p(y_{i,t}|\mathbf{y}_{\setminus i,t}, \mathbf{z}, \mathbf{x}) \propto \frac{n^-_{x,y} + \beta}{\sum_x n^-_{x,y} + N_x\beta} \frac{n^-_{y,z} + \alpha}{\sum_y n^-_{y,z} + N_y\alpha}. \quad (2)$$

Here $\mathbf{y}_{\setminus i,t}$ denotes all the $\mathbf{y}$ variables excluding $y_{i,t}$; $n^-_{x,y}$ denotes the counts of feature x being associated to action y; $n^-_{y,z}$ denotes the counts of action y being associated to behaviour z. Superscript "$-$" denotes counts over the remaining dataset excluding item (i,t). N_x is the size of the visual event codebook, and N_y the number of simple actions.

The Gibbs sampling update for cluster z_t is derived by integrating out parameters ψ and θ in the conditional $p(z_t|\mathbf{y}, \mathbf{z}_{\setminus t}, \mathbf{x})$, and must account for the possible transitions between z_{t-1} and z_{t+1} along the Markov chain of clusters:

$$p(z_t|\mathbf{y}, \mathbf{z}_{\setminus t}, \mathbf{x}) \propto \frac{\prod_y \Gamma(\alpha + n_{y,z_t})\Gamma(N_y\alpha + n^-_{\cdot,z_t})}{\prod_y \Gamma(\alpha + n^-_{y,z_t})\Gamma(N_y\alpha + n_{\cdot,z_t})} \frac{n^-_{z',z} + \gamma}{n^-_{z',z} + N_z\gamma} \frac{n_{z_{t+1},z_t} + \mathbf{I}(z_{t-1} = z_t)\mathbf{I}(z_t = z_{t+1}) + \gamma}{n_{\cdot,z_t} + \mathbf{I}(z_{t-1} = z_t) + N_z\gamma}. \quad (3)$$

Here $n_{z',z}$ are the counts of behaviour z' following behaviour z, $n_{\cdot,z} \triangleq \sum_{z'} n_{z',z}$, and N_z is the number of clusters. $\mathbf{I}$ is the identity function that returns 1 if its argument is true, and Γ is the gamma function. Note that we do not obtain the simplification of gamma functions as in standard LDA and Eq. (2), because the inclusive and exclusive counts may differ by more than 1, but this is not prohibitively costly, as Eq. (3) is computed only once per clip. Iterations of Eqs.(2) and (3) entail inference by eventually drawing samples from the posterior $p(\{\mathbf{y}_t, z_t\}_1^T | \{\mathbf{x}\}_1^T, \alpha, \beta, \gamma)$. Parameters $\{\phi, \psi, \theta\}$ may be estimated from the expectation of their distribution given any full set of samples [11], e.g.

$$\hat{\phi}_y = \frac{n_{x,y} + \beta}{n_{\cdot,y} + N_x\beta}. \quad (4)$$

3. Online Inference and Saliency Detection

A limitation of the (standard) model learning and inference method described above, also adopted by [9, 13], is that they are offline, batch procedures. For on-the-fly behaviour mining in video, we formulate a new real-time filtered (or smoothed) inference algorithm for our MCTM after an offline batch learning phase.

Given a training dataset of T_{tr} clips, we have generated N_s samples $\{\{\mathbf{y}_t, z_t\}_1^{T_{tr}}, \hat{\phi}, \hat{\psi}, \hat{\theta}\}_{s=1}^{N_s}$ from the posterior distribution of latents in our model $p(\{\mathbf{y}_t, z_t\}_{t=1}^{T_{tr}} | \{\mathbf{x}\}_1^{T_{tr}}, \alpha, \beta, \gamma)$. We assume that no further adaptation of the parameters is necessary, i.e. the training dataset is representative, so $p(\phi, \psi, \theta | \mathbf{x}_{t'>T_{tr}}) = p(\phi, \psi, \theta | \mathbf{x}_{1:T_{tr}})$. We then perform Bayesian filtering in the Markov chain of clusters to infer the current clip's behaviour $p(z_t|\mathbf{x}_{1:t})$ by approximating the required integral over the parameters with sums over their Gibbs samples [11]. Conditioned on each set of (sampled) parameters, the other action $y_{i,t}$ and behaviour z_t variables decorrelate, so efficient recursions can be derived to compute the behaviour category for each clip online:

$$p(\mathbf{z}_{t+1}|\mathbf{x}_{1:t+1}) = \int_{\phi,\theta,\psi,z_t} \frac{p(\mathbf{x}_{t+1}, z_{t+1}|z_t, \phi, \theta, \psi, \mathbf{x}_{1:t})\, p(z_t, \phi, \theta, \psi|\mathbf{x}_{1:t})}{p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t})}, \approx \frac{1}{N_s} \sum_s \frac{p(\mathbf{x}_{t+1}|z_{t+1}, \phi^s, \theta^s)\, p(z_{t+1}|z_t^s, \psi^s)}{p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t})}. \quad (5)$$

Bayesian surprise (saliency, or irregularity), is optimally measured by the marginal likelihood of the new observation given all the others, $p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t})$. This can be determined from the normalization constant of Eq. (5), or explicitly as:

$$p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}) = \int_{\phi,\theta,\psi,z_t} p(\mathbf{x}_{t+1}|\psi,\theta,\phi,\mathbf{x}_{1:t})\, p(z_t,\phi,\psi,\theta|\mathbf{x}_{1:t}),$$
$$\approx \frac{1}{N_s} \sum_{s,z_{t+1}} p(\mathbf{x}_{t+1}, z_{t+1}|\psi^s,\theta^s,\phi^s,z_t^s). \quad (6)$$

Without the iterative sweeps of the Gibbs sampler, even summing over samples s, behaviour inference (or clip categorization) and saliency detection can be performed online and in real-time by Eq. (5) and Eq. (6). Note that in practice Eq. (5) may suffer from label switching [5], so a single sample should be used for interpretable results. Eq. (6) is independent of label switches and should be used with all samples. This online approach has no direct analogy in vanilla LDA [3] (Fig. 1(a)), as the per document parameter θ requires iterative computation to infer. We compare the computational cost of our MCTM, LDA [3], Dual-HDP [13] and HMMs in Sec. 4.4.

The Bayesian measure of saliency $p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t})$ of test point $\mathbf{x}_{t+1}$ given training data $\mathbf{x}_{1:T_{tr}}$ and other previous test data $\mathbf{x}_{t>T_{tr}}$ is used to detect irregularity. $p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t})$ reflects the following salient aspects of the data:

Intrinsic: $\mathbf{x}_t$ rarely occurred in training data $\mathbf{x}_{1:T_{tr}}$.

Actions: $x_{i,t}$s rarely occurred together in the same topic in $\mathbf{x}_{1:T_{tr}}$.

Behaviours: $\mathbf{x}_t$ occurred together in topics, but such topics did not occur together in clusters in $\mathbf{x}_{1:T_{tr}}$.

Dynamics: $\mathbf{x}_t$ occurred together in a cluster z_t, but z_t did not occur following the same cluster z_{t-1} in $\mathbf{x}_{1:T_{tr}}$.

Such detections are made possible because the hierarchical structure of our model represents behaviour at different levels (events, actions, behaviours, behaviour dynamics).

4. Experiments

4.1. Datasets and Settings

Experiments were carried out using video data from three complex and crowded public scenes. **Street Intersection Dataset**: This contained 45 minutes of 25 fps video of a busy street intersection where three traffic flows in different directions are regulated by the traffic lights, in a certain temporal order (see Fig. 3(a)-(e)). The frame size is 360×288. **Pedestrian Crossing Dataset:** This also consists of 45 minutes of 360×288 pixel 25 fps video, and captures a busy street intersection with particularly busy pedestrian activity (see Fig. 3(f)-(i)). Typical behaviours here are pedestrian crossings alternating with two main traffic flows. **Subway Platform Dataset:** A total of 18 minutes of videos from the UK Home Office i-LIDS dataset is selected for the third experiment. Though equally busy, the visual scene in this dataset differs significantly from the other two in that it is indoor and features mainly people and trains (see Fig. 3(j)-(n)). In addition, the camera was mounted much closer to the objects and lower, causing more severe occlusions. Typical behaviours in this scene include people waiting for the train on the platform, and getting on or off the train. The video frame size is 640×480 captured at 25 fps.

We used 5 minutes from each dataset for training, and tested (Eqs. (5) and (6)) on the remaining data. The cell size for both of the two street datasets was 8×8, and 16×16 for the subway dataset. Optical flow computed in each cell is quantized into 4 directions for the two street datasets and 5 for the subway dataset, with the fifth corresponding to stationery foreground objects common in the subway scene. We run the Gibbs sampler (Eqs. (2) and (3)) for a total of 1500 complete sweeps, discarding the first 1000 as burn-in, and then taking 5 samples at a lag of 100 as independent samples of the posterior $p(\{\mathbf{y}_t, z_t\}_1^{T_{tr}}|\mathbf{x}_{1:T_{tr}}, \alpha, \beta, \gamma)$. In each case we selected the number of actions as $N_y = 8$ and the number of behaviour clusters as $N_z = 4$; except for the pedestrian crossing dataset, where we used $N_z = 3$ because there are clearly three traffic flows. We fixed these numbers for ease of illustration. Larger N_y and N_z result in a more fine-grained decomposition of scene behaviour. Dirichlet hyper-parameters were fixed at $\{\alpha = 8, \beta = 0.05, \gamma = 1\}$ for all experiments to encourage composition of specific actions into general topics, but these could be empirically estimated during sampling [6].

4.2. Unsupervised Scene Interpretation

Clustering Visual Events into Actions: The learned topics of our MCTM correspond to actions consisting of co-occurring visual events. These actions are typically associated with patterns of moving objects. Fig. 2 shows some example actions/topics y discovered by way of plotting the visual events $\mathbf{x}$ in the top 50% of the mass of the distribution $p(\mathbf{x}|y, \hat{\phi}_y^s)$ (Eq. 4). Note that each action has a clear semantic meaning. In the street intersection dataset, Figs. 2(a) and (b) represent vertical left lane and horizontal leftwards traffic respectively, while Fig. 2(c) represents the vertical traffic vehicles turning right at the filter. In the pedestrian crossing dataset, Figs. 2(d) and (e) illustrate two independent vertical traffic flows, and Fig. 2(f) represents diagonal traffic flow and pedestrians crossing at the lights while the flows of (d) and (e) have stopped. For the subway dataset, Fig. 2(g) includes people leaving (yellow arrows) from a stopped train (cyan dots on the train). Fig. 2(h) includes people walking up the platform and Fig. 2(i) shows people sitting on the bench waiting.

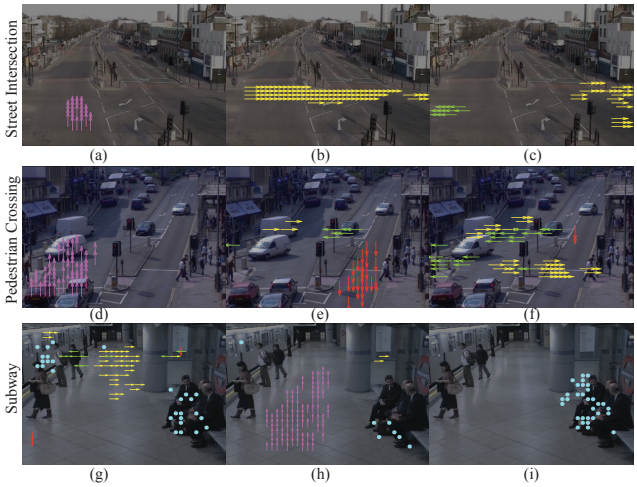

Figure 2. Example topics/actions learned in each of the three scenarios illustrated by the most likely visual events for each $\hat{\phi}_y^s$. Arrow directions and colors represent flow direction of the event.

Discovering Behaviours and their Dynamics: Co-occurring topics are automatically clustered into behaviours z via matrix θ_z (Sec. 2.3), each of which corresponds to a complex behaviour pattern involving multiple interacting objects. Complex behaviour clusters discovered for the three dynamic scenes in the 5 minutes of training data, are depicted in Fig. 3. Specifically, Figs. 3(a) and (b) represent horizontal left and right traffic flows respectively including right turn traffic (compare horizontal only traffic in Fig. 2(b)). Figs. 3(c) and (d) represent vertical traffic flow with and without interleaved turning traffic. The temporal duration and order of each traffic flow is also discovered accurately. For example, the long duration and exclusiveness of the horizontal traffic flows (a) and (b) – and the interleaving of the vertical traffic (c) and vertical turn traffic (d) – are clear from the learned transition distribution $\hat{\psi}^s$ (Fig. 3(e)).

For the pedestrian crossing dataset, three behaviour clusters are learned. Fig. 3(f), diagonal flow of far traffic and downwards vertical traffic flow at the right, excluding the crossing zone where there is pedestrian flow (horizontal yellow arrows). Figs. 3(g) and (h) show outer diagonal and vertical traffic, and inner vertical traffic respectively with no pedestrians crossing. The activity of the pedestrian crossing light is evident by the switching between (f) and (g) in the learned transition distribution (Fig. 3(i), top left).

The four behaviour categories discovered in the subway scene were: People walking towards (red & green arrows) an arriving train (green arrows on train) (Fig 3(j)); People boarding a stopped train (cyan dots on the track) or leaving the station (Fig 3(k)); People leaving the station while the trains wait (Fig 3(l)) (in this dataset, the train usually waited for longer than it took everyone to board; hence this cluster); People waiting for the next train by sitting on the bench (Fig 3(m)). Our model is also able to discover the cycle of

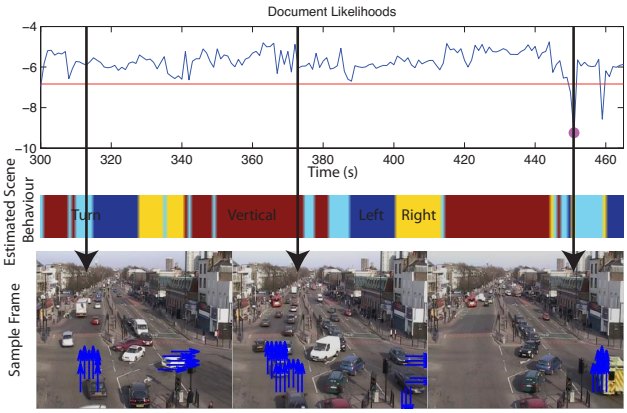

Figure 4. An example of online processing.

behaviour on the platform triggered by arrival and departure of trains (Fig. 3(n)). For example, the long duration of waiting periods (m) between trains, broken primarily by the train arriving state (j), (see Fig. 3(n), fourth column).

4.3. Online Video Screening

The model was learned for each scenario before new video data was screened online. The overall behaviours were identified using Eq. (5), and visual saliency (irregularity) measured using Eq. (6). Fig. 4 shows an example of online processing on test data from the street intersection dataset. The MAP estimated behaviour $\hat{z}_t$ at each time is illustrated by the colored bar, and reports the traffic phase: turning, vertical flow, left flow and right flow. The top graph shows the likelihood $p(\mathbf{x}_t|\mathbf{x}_{1:t-1})$ of each clip as it is processed online. Three examples are shown including two typical clips (turning vertical traffic and flowing vertical traffic categories) and one irregular clip where a vehicle drives in the wrong lane. Each is highlighted with the flow vectors (blue arrows) on which computation is based.

We manually examined the top 1% most surprising clips screened by the model in the test data. Here we discuss some examples of flagged surprises. In Fig. 5(a) and (b), another vehicle drives in the wrong lane. This is surprising, because that region of the scene typically only includes down and leftward flows. This clip is *intrinsically*, (Sec. 3) unlikely, as these events were rare in the training data under any circumstances. In Fig. 5(c) and (d), a police car breaks a red light and turns right through opposing traffic. Here the right flow of the other traffic is a typical action, as is the left flow of the police car. However, their conjunction (forbidden by the lights) is not. Moreover some clips in this multi-second series alternately suggest left and right flows, but such dynamics are unlikely under the learned temporal model (Fig. 3(e)). Therefore this whole series of clips is *behaviorally* and *dynamically* (Sec. 3) unlikely given global and temporal constraints entailed by $p(\mathbf{x}_t|\mathbf{x}_{1:t-1})$.

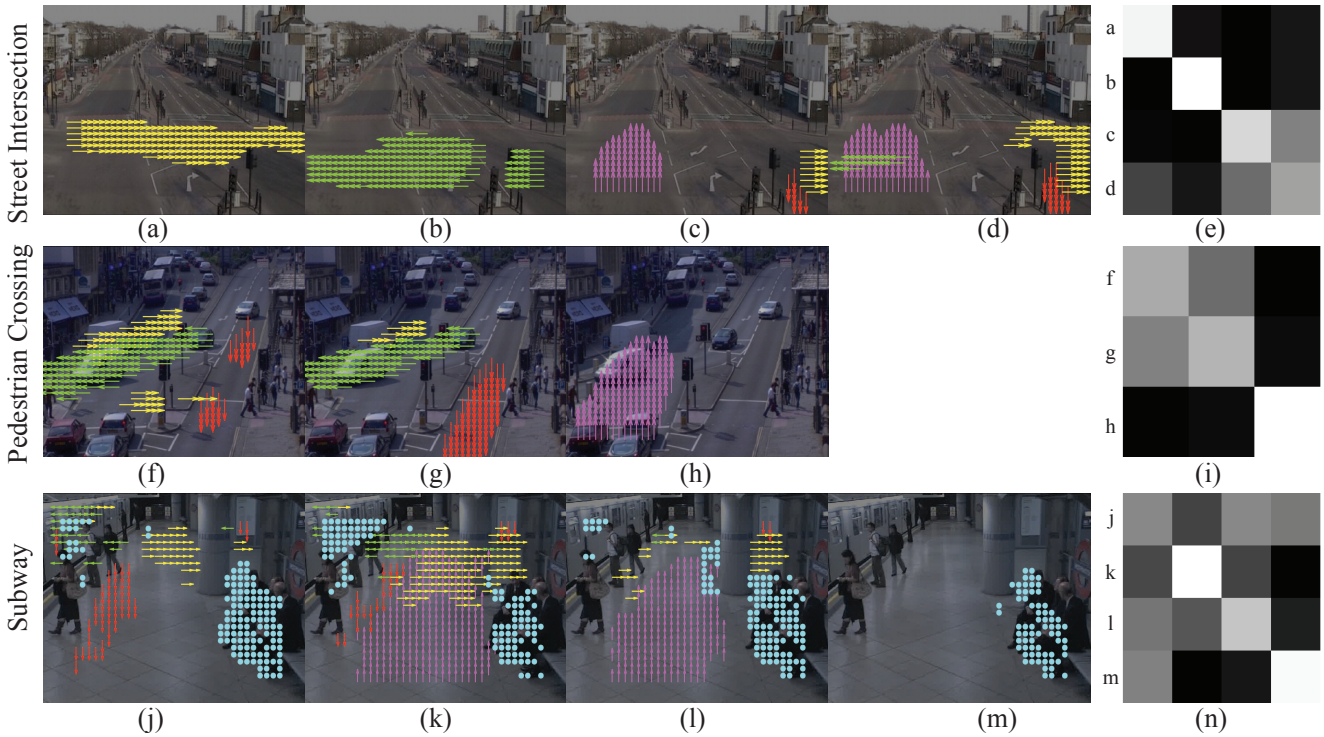

Figure 3. Behaviour and dynamics in each of the three scenarios, illustrated by the most likely visual words/events for each behaviour $\hat{\theta}_z^s$ and the transitions between behaviours $\hat{\psi}_z^s$.

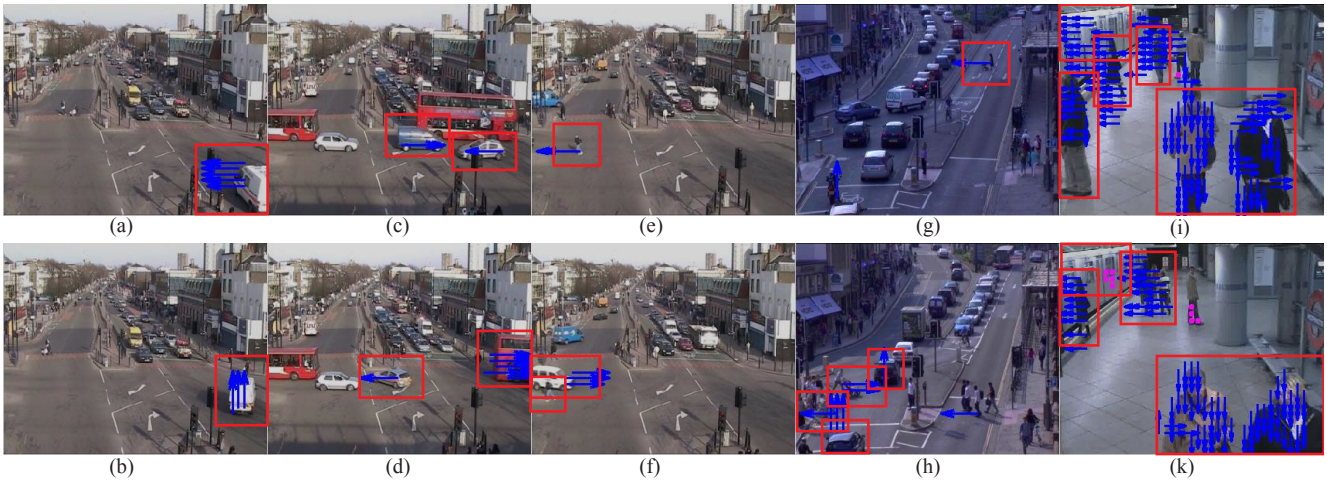

Figure 5. Sample salient clips discovered. Arrows/dots indicate input events and red boxes highlight regions discussed in the text.

Another *behavioral* (action concurrence) surprise to the model is the jay-walker in Fig. 5(e-f). Here a person runs across the intersection to the left, narrowly avoiding the right traffic flow. Both left and right flows are typical, but again their concurrence in a single document, or rapid alteration in time is not. Fig. 5(g) shows the detection of a jaywalker triggered by *intrinsically* unlikely horizontal motion across the street. In contrast, Fig. 5(h) illustrates two plausible pedestrian actions of crossing left and right at the crosswalk, but doing so at the same time as the vertical traffic flow. This is multi-object situation is *behaviorally*, (Sec. 3) irregular. In Fig. 5(i) a train arrives, and three people typically (Fig. 3(j)) walk towards the train for boarding. However, unusually, other people walk away from the train down the platform, a *behaviorally* unlikely concurrence. In Fig. 5(k), the train is now stationary. While most people perform the typical paired action of boarding (Fig. 3(k)), others walk away from the train down the platform, a multi-object

behaviour detected due to low likelihood $p(\mathbf{x}_t|\mathbf{x}_{1:t-1})$.

Figs. 5(c-f) illustrate an important feature of our model that gives a significant advantage over non-temporal LDA based models [9, 13]: Our model is intrinsically less constrained by bag-of-words size, i.e. determining a suitable temporal window (clip) size. With standard LDA, larger bag sizes would increase the chance that vertical and horizontal flows here were captured concurrently and therefore flagged as surprising. However, larger bag sizes also capture much more data, risking loosing interesting events in a mass of normal ones. Our model facilitates the use of a small one second bag size, by providing temporal information so as to penalize unlikely behaviour switches. As a result, our model can discover not only quick events such as Fig. 5(a) and (b) that might be lost in larger bags, but also longer time-scale events such as Fig. 5(c-f) that could be lost in many independently distributed smaller bags.

To demonstrate the breadth of irregular behavioural patterns our model is capable of consistently identifying, some of which are visually subtle and difficult to detect even by human observation, we provide a human interpreted summary of the categories of screened salient clips in Table 1. We compare the results with two alternatives, LDA [3] with N_y topics, and a HMM with N_z states. Clips with no clear salient behaviour were labeled "uninteresting". These were variously due to camera glitches, exposure compensation, birds, very large trucks, and limited training data to accurately profile typical activities. There is no algorithmic way to determine "why" (i.e. action, behaviour, dynamics) events were surprising to the model, so we do not attempt to quantify this. Our MCTM outperforms the other two models especially in the more complex behaviour categories of red-light-breaking, u-turns and jaywalking. In these cases, the saliency of the behaviour is defined by an atypical concurrence of actions and/or sequence of behaviours over time, i.e. a surprise is defined by complex spatio-temporal correlations of actions rather than simple individual actions. In contrast, conventional LDA can infer actions, but cannot reason about their concurrence or temporal sequence simultaneously. HMMs can reason about sequences of behaviours, but with point (EM) learning, and lacking the intermediate action representation, HMMs suffer from severe over-fitting. All the models do fairly well at detecting intrinsically unlikely words which are visually well-defined independently, e.g. wrong way driving.

For the pedestrian crossing dataset, the result is shown in Table 2. Atypical pedestrian behaviours were jaywalking far from the crosswalk (intrinsically unlikely visual events), and crossing at the crosswalk through traffic (unlikely action concurrence; Fig. 3(f) vs (g),(h)). Our MCTM was more adept than both LDA and HMM at detecting the more subtle behaviours. This is due to the same reasons of simultaneous hierarchical and temporal modeling of actions and improved robustness due to Bayesian parameter learning compared to HMMs especially. Finally, for the subway dataset (Table 2) the only interesting behaviours observed were people moving away from the train during clips where typical behaviour was approaching trains and boarding passengers, this was detected by our model and not the others.

Street Intersection	**MCTM**	**LDA**	**HMM**
Break Red Light	3	0	1
Illegal U-Turn	5	2	1
Jaywalking	1	0	0
Drive Wrong Way	12	14	12
Unusual Turns	5	2	4
Uninteresting	29	37	37

Table 1. Summary of human meaningful clip types discovered by different models for the street intersection dataset.

Pedestrian Cross	**MCTM**	**LDA**	**HMM**
Jaywalking	18	15	15
Through Traffic	11	6	5
Uninteresting	33	41	42
Subway Platform	**MCTM**	**LDA**	**HMM**
Contraflow	2	0	0
Uninteresting	3	5	5

Table 2. Summary of human meaningful clip types discovered by different models for crossing and subway platform datasets.

4.4. Computational Cost

The computational cost of MCMC learning in any model is hard to quantify, because assessing convergence is itself an open question [5], as also highlighted by [13]. In training, our model is dominated by the $O(N_T N_y)$ cost of resampling the total number N_T of input features in the dataset per Gibbs sweep, which is the same as [13]. In testing, our model requires $O(N_z^2) + O(N_T N_y N_z)$ time per parameter sample. In practice using Matlab code on a 3GHz CPU, this meant that training on 5 minutes of our data required about 4 hours. Using our model to process one hour of test data online took only 4 seconds in Matlab. Processing the same data with (Variational) LDA in C [3] took about 20 and 8 seconds respectively, while (EM) HMM in Matlab took 64 seconds and 26 seconds. Wang et al.[13] reported that Gibbs sampling in their HDP model required 8 hours to process each hour of data from their quieter (and therefore fewer words, so quicker) dataset; and they do not propose an online testing solution. These numbers should not be compared literally given the differences in implementations and datasets; however the important thing to note is that while our model is competitive in training speed to sophisticated contemporary models [13], it is much faster for online testing. Moreover, it is faster than the simple models which it outperforms in saliency detection.

5. Discussion

We introduced a novel Bayesian topic model for simultaneous hierarchical and temporal clustering of visual events into actions and global behaviours. The model addresses two critical tasks for unsupervised video mining: modeling scene behavioral characteristics under-pinned at different spatial and temporal levels, and online behaviour screening and saliency detection. Our Gibbs learning procedure has proven effective at learning actions, behaviours and temporal correlations in three diverse and challenging datasets. We showed how to use the Gibbs samples for rapid Bayesian inference of clip category and saliency. Evaluating the salient clips returned from our diverse datasets, our MCTM outperforms LDA and HMMs for unsupervised mining and screening of salient behaviours, especially for visually subtle, and temporally extended activity. This was because we model simultaneously temporal evolution of behaviour (unlike LDA), the hierarchical composition of action into behaviours (unlike LDA and HMM) and Bayesian parameter learning (unlike HMM). Compared to object-centric approaches such as [1, 12], our simple and reliable visual features improve robustness to clutter and occlusion.

We have not addressed the issue of determining the optimal number of behaviours and actions in a given dataset, as was done in [13]. For our model, Bayesian model selection can readily be done offline once per scene in a principled if computationally intensive way: maximizing the marginal likelihood $p(\mathbf{x}|N_z, N_y)$ based on the Gibbs output, or Eq. (6). This approach retains the option of subsequent online real-time processing, in contrast to [13] which does not propose an online solution, and whose batch solution is in the order of ten times slower than real time [13].

To put our theoretical modeling contribution in context, it contrasts other hierarchical work which clusters actions, but not over time [13], and other non-hierarchical work which temporally correlates words within (rather than across) documents[6] or provides continuous variation (rather than discrete clustering) of parameters over time[2].

In summary, we have presented a unified model for completely unsupervised learning of scene characteristics, dynamically screening and identifying irregular spatio-temporal patterns as salient behaviour clips that may be of interest to a human user. An important feature of our approach is the breadth of different kinds of behaviours that may be modeled and flagged as salient due to our simultaneous hierarchical topic modeling and temporal correlation globally optimized in a unified model. For example, temporally extended events typically only flagged by object/tracking centric models [12, 1] such as u-turns as well as multi-object events typically only flagged by statistical event models such as jaywalking [13]. Finally, the specific formulation of our model also permits Bayesian saliency discovery of these type of events online in real-time.

Acknowledgment: This research was partially funded by EU FP7 project SAMURAI with grant no. 217899.

References

[1] A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In *CVPR*, pages 1–8, 2008.

[2] D. Blei and J. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[4] T. Duong, H. Bui, D. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *CVPR*, pages 838–845, 2005.

[5] W. Gilks, S. Richardson, and D. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall / CRC, 1995.

[6] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In *NIPS*, pages 537–544, 2007.

[7] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. SMC*, 34(3):334–352, 2004.

[8] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *PAMI*, 28(9):1450–1464, 2006.

[9] J. Li, S. Gong, and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. In *BMVC*, 2008.

[10] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *ICCV*, pages 1–8, 2007.

[11] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, pages 487–494, 2008.

[12] I. Saleemi, K. Shafique, and M. Shah. Probabilistic modeling of scene dynamics for applications in visual surveillance. *PAMI*, 31(8):1472–1485, Aug. 2009.

[13] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception by hierarchical bayesian models. *PAMI*, 31(3):539 – 555, 2009.

[14] X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *ECCV*, 2006.

[15] T. Xiang and S. Gong. Video behavior profiling for anomaly detection. *PAMI*, 30(5):893–908, 2008.

[16] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, pages 819–826, 2004.