Int J Comput Vis (2012) 98:303–323
DOI 10.1007/s11263-011-0510-7

Video Behaviour Mining Using a Dynamic Topic Model

Timothy Hospedales · Shaogang Gong · Tao Xiang

Received: 21 August 2009 / Accepted: 22 November 2011 / Published online: 8 December 2011
© Springer Science+Business Media, LLC 2011

Abstract This paper addresses the problem of fully automated mining of public space video data, a highly desirable capability under contemporary commercial and security considerations. This task is especially challenging due to the complexity of the object behaviors to be profiled, the difficulty of analysis under the visual occlusions and ambiguities common in public space video, and the computational challenge of doing so in real-time. We address these issues by introducing a new dynamic topic model, termed a Markov Clustering Topic Model (MCTM). The MCTM builds on existing dynamic Bayesian network models and Bayesian topic models, and overcomes their drawbacks on sensitivity, robustness and efficiency. Specifically, our model profiles complex dynamic scenes by robustly clustering visual events into activities and these activities into global behaviours with temporal dynamics. A Gibbs sampler is derived for offline learning with unlabeled training data and 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 for four complex and crowded public scenes, and successful mining of behaviors and detection of salient events in each.

T. Hospedales (✉) · S. Gong · T. Xiang
School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK
e-mail: tmh@eecs.qmul.ac.uk

S. Gong
e-mail: sgg@eecs.qmul.ac.uk

T. Xiang
e-mail: txiang@eecs.qmul.ac.uk

Keywords Behaviour profiling · Video behaviour mining · Topic models · Learning for vision · Bayesian methods · Probabilistic modelling

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. Ideally, users would like the ability to mine large volumes of recorded data to extract useful information about behaviour patterns of individuals and groups in the area under surveillance; and to monitor the scene for saliency in real-time in order to provide the potential for immediate response. Learning spatio-temporal behaviour patterns from videos of a public space is frequently of intrinsic commercial or security interest for users to gain more knowledge about activity patterns in public spaces which they are responsible for. For instance, retailers may be interested in shoppers browsing habits, while managers of public infrastructure sites might be interested in understanding typical behaviors. The learned behavior patterns may then be exploited for online analysis of the site state and for detection of salient activity pattern requiring further investigation. The alternative is the expensive and laborious manual analysis of the data and customization of detection software—which is prohibitive for many installations.

In practice, large volumes of recorded video data are frequently only stored passively for record purposes because of the challenges in developing such automatic and robust analysis methods. There are in general three challenges: dealing with the variety of potentially interesting behaviors, achieving sufficient robustness for practical use and real-time online operation.

Behavioral Complexity The nature of behavioural patterns in a given scene, and importantly, the classes of 'subjectively interesting behaviour' to a user with a specific task could be defined by a wide variety of factors: the activity of a single object over time (e.g., its track), the correlated spatial states 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 has a particular order dictated by the lights). In addition, the spatial or temporal range over which correlations might be important may be short or long. Building models general and flexible enough to represent all these aspects of behavior is an open research question.

Robustness & Sensitivity Typical public space surveillance scenarios involve noisy, ambiguous and cluttered input. They may also exhibit extreme lighting variations and a variety of different object classes, poses and dynamics. A robust model is one which copes gracefully with these challenges. However, classical approaches to visual surveillance build upon a segmentation, classification, identification and tracking pipeline which may be brittle under these circumstances.

An important related issue is sensitivity: how well a system can discover behaviors of interest if they are for example, visually subtle, very short in duration or co-occurring with other less interesting behaviors. In practice there is a difficult tradeoff between these two requirements. To be useful, a monitoring system needs to be both sufficiently robust and sensitive.

Computational Tractability To enable prompt response to important or unusual events, any automatic analysis should be performed real-time; thereby constraining the range of potential usable techniques.

In light of all these issues we present in this paper a new model, which we term a "Markov Clustering Topic Model" (MCTM), to address the problem of unsupervised mining of multi-object spatio-temporal behaviours in crowded and complex public scenes. Our approach draws on existing theoretical work on probabilistic topic models (PTMs) and dynamic Bayesian networks (DBNs) to achieve a robust hierarchical model of behaviors and their dynamics. The overall three-layer architecture is as follows: A codebook of simple *visual events* (e.g., foreground pixel presence or moving pixel presence) is learned, so as to generate discrete input features from video. Co-occurring events are automatically composed into *activities* (e.g., a pedestrian crossing the road). Co-occurring activities are automatically composed into complex multi-object *behaviors* (e.g., street intersection interactions), and these behaviors are considered correlated in time. By introducing a Markov chain to model behaviour dynamics, we define a DBN generalization of a static topic model.

Our model is learned offline from unlabeled training data with Gibbs sampling. The hierarchical and temporal structure of our model addresses the challenge of modeling complex dynamic multi-object behaviors. The issues of robustness and sensitivity are addressed by the topic model representation and temporal correlation, and a new inference algorithm permits online real-time operation. The result is a framework which can address the problem of unsupervised profiling and mining of multi-object spatio-temporal behaviours in crowded and complex public scenes by discovering underlying spatio-temporal regularities. The framework can provide domain knowledge about a scene in the form of learned patterns; detect the occurrence of learned activities and behaviors online; and importantly, 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, behaviours and dynamics in this scene", "Estimate how many different activities and behaviors are exhibited in this scene", "Show me a (ranked) list of interesting (irregular) events from the past 24 hours", "Alert me whenever any sufficiently interesting (irregular) event occurs" and "Learn a model of *this* example behaviour, and tell me if it occurs."

The rest of this paper is organized as follows: Sect. 2 gives an overview of related research and highlights the contribution of the work. Section 3 gives a detailed explanation of the theoretical and implementation details of our framework. Section 4 discusses how to exploit our framework for semi-supervised learning. We then evaluate our approach on tasks including learning (Sect. 5.2), unusual activity detection (Sect. 5.3), classification (Sect. 5.4) and semi-supervised classification (Sect. 5.5) using four diverse datasets collected from crowded public spaces. The paper concludes in Sect. 6 with discussion including the limitations of the proposed model and future work.

2 Related Work

Our goals are related to the general field of video mining, in which research has traditionally focused on content based indexing and summarization systems and event detection systems. Indexing and summarization systems try to categorize each segment of a video, often obtained by shot change detection (Meng and Chang 1996) or by content (e.g., presence of a particular object) to permit searching and browsing (Pritch et al. 2008). Event detection systems search for particular defined events of interest (e.g., people falling) in video (Chang et al. 2008; Xie et al. 2008). There has been much recent progress on these problems as evidenced by the successful responses to the TRECvid (3) semantic indexing (e.g., Inoue et al. 2009) and event detection (e.g., Hu

et al. 2009) challenges. Nevertheless, most of these event detection approaches engineer—for each scenario—features that are very high level (e.g., body part detectors) and hard-coded rules to detect specific target events (e.g., embracing, loitering or pointing); or large-vocabulary bags of interest points (1). In this paper we are specifically interested in unsupervised learning of models for complex multiple object behaviors captured in public space surveillance videos of varying view range and composition, etc. In this case, many standard video mining approaches are inappropriate because of the constraint to specific pre-defined events or over-specificity to scene composition.

Recent research on learning behavior models for understanding video has broadly fallen into object-centric detection and tracking approaches (Hu et al. 2004), and non-object-centric approaches. Tracking based approaches (Berclaz et al. 2008; Sillito and Fisher 2008; Hu et al. 2006; Wang et al. 2006; Dee and Hogg 2004; Stauffer and Grimson 2000; Johnson and Hogg 1996) explicitly represent the spatial state of visual objects over time. This allows them to easily model behaviours like typical paths, and detect events readily defined in terms of trajectories such as counter-flow (Berclaz et al. 2008), u-turns (Hu et al. 2006) or goal consistency (Dee and Hogg 2004). However, such models only work well if complete tracks can be reliably obtained in training and test data which as we have discussed, is difficult in crowded public spaces. To improve tracker robustness, scenario specific models of typical dynamics have been learned (Ali and Shah 2008; Berclaz et al. 2008). To improve robustness of activity models to tracking failures, non-parametric representations of track statistics have also been exploited (Basharat et al. 2008; Saleemi et al. 2009). Nevertheless, a major limitation of tracking based approaches in general is the difficulty in modeling behaviours characterized by coordinated activity of multiple objects, which may be the defining characteristic of an interesting behavior in video.

To improve robustness to missed detections and broken tracks, and to enable multi-object spatio-temporal correlation modeling, statistical methods have been devised to process directly on image data (Boiman and Irani 2007), quantized optical flow (Wang et al. 2009; Kim and Grauman 2009; Hospedales et al. 2009) or other low level 'event' features in video (Duong et al. 2005; Xiang and Gong 2008b; Li et al. 2008; Zhong et al. 2004; Benezeth et al. 2009; Hospedales et al. 2009). These methods have typically employed non-parametric indexing (Boiman and Irani 2007), Dynamic Bayesian Networks (DBNs) such as a Hidden Markov Models (HMM) (Duong et al. 2005; Xiang and Gong 2008a, 2008b), Markov Random Fields (MRFs) (Kim and Grauman 2009; Benezeth et al. 2009) or probabilistic topic models (PTMs) (Li et al. 2008) such as Latent Dirichlet Allocation (LDA) (Blei et al. 2003) or extensions (Wang et al. 2009; Hospedales et al. 2009). Non-parametric indexing approaches (Boiman and Irani 2007) are robust behaviour models, but suffer from scalability issues in having to store and search a whole video patch database, and cannot model important and meaningful spatio-temporal correlations between events (e.g., exclusive use of an intersection by traffic flowing in different directions with regulated order). DBNs are natural for modeling dynamics of behaviour (Xiang and Gong 2006, 2008a, 2008b) and, with hierarchical structure, also have the potential to perform clustering of both simple activities and more complex behaviours simultaneously (Duong et al. 2005). Nevertheless, modeling the temporal order of noisy visual events explicitly is risky, because noise in the event representation can easily propagate through the model, and be falsely detected as salient (Li et al. 2008; Wang et al. 2009). Another strategy is to learn a MRF over a space-time volume, which can model permitted local spatio-temporal correlations (Benezeth et al. 2009; Kim and Grauman 2009), and also improve robustness by smoothing salient event detections (Kim and Grauman 2009). However, flat and non-hierarchical models like Benezeth et al. (2009), Kim and Grauman (2009) cannot represent the composition of simpler local actions into more complex behaviors (e.g., traffic intersection modeling). Moreover, because they only correlate a small window of time, they lack the potential for long term temporal behaviour reasoning.

To overcome the problems of robustness and multi-object behavior modeling, PTMs (Blei et al. 2003) were borrowed from text document analysis. These "bag of words" models represent visual event co-occurrence (Li et al. 2008)—potentially hierarchically (Wang et al. 2009)—but completely ignore 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. The hierarchical PTM models developed in Li et al. (2008) and Wang et al. (2009) have also tended to be computationally expensive, precluding the desired usage scenario for real-time monitoring of salient events in a public space.

The most similar work to ours is that of Wang et al. (2009), who use a hierarchical and non-parametric Dirichlet process (DP) topic model of behavior. Our approach differs in three important ways: (i) We model behavior dynamics, while Wang et al. (2009) uses a static topic model. This ensures sensitivity to dynamics (ordering) of behaviours, and also provides some generality in sensitivity to behaviours of longer and shorter time-scale; (ii) Our parametric approach permits semi-supervised learning to detect specific known

interesting behaviors unlike the purely outlier detection approach in Wang et al. (2009) and (iii) MCTM performs real-time online inference, unlike Wang et al. (2009).

The idea of introducing various kinds of dynamics into topic models has been exploited recently for text document analysis. Griffiths et al. (2007), Wallach (2006), Gruber et al. (2007) develop models to temporally correlate words within (rather than across) documents (corresponding to video clips in our case). This makes sense for text where the word tokens come in a one dimensional stream, but is not clearly suitable for video where a (variable sized) set of visual events occurs at each time. Blei and Lafferty (2006) model continuous change (rather than switching) of the parameters correlating words and topics. This models longer term changes in the statistics of the observation model of the corpus and could correspond to adapting to weekly or seasonal trends in video, but not the live dynamics of behaviors which we are interested in.

3 Unsupervised Spatio-Temporal Video Mining

3.1 Video Representation

We wish to construct a generative model capable of automatically 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 Sect. 5.1). These camera views contain multiple groups of heterogeneous objects, occlusions, and shadows, challenging segmentation and tracking based methods. Instead, local motions are used as low level input features. Specifically, a camera view is divided into $C \times C$ pixel cells, and optical flow computed in each cell. When the magnitude of a flow vector is greater than a threshold Th_o, it 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. This is similar to the representations used in Wang et al. (2009), Kim and Grauman (2009), Zhong et al. (2004).

For a 320×240 video frame with cell size of 10×10, a total of $N_x = 32 \times 24 \times 4 = 3072$ different discrete visual events may therefore 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 our approach: it can easily incorporate other kinds of 'metadata' 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 N_t visual events occurring in video clip t, denoted as $\mathbf{x}_t = \{x_{1,t}, \ldots, x_{i,t}, \ldots, x_{N_t,t}\}$.

3.2 Markov Clustering Topic Model (MCTM)

One of the most popular approaches to topic modeling is Latent Dirichlet Allocation (LDA) (Blei et al. 2003), which we first review for reference. LDA (Fig. 1(a)) is a generative 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 discrete distribution $p(x_{i,m}|\phi, y_{i,m})$ with parameter ϕ indexed by the current topic of discussion $y_{i,m}$. Topics are chosen from a per-document Dirichlet 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 topic based representation of text documents can facilitate e.g., querying and similarity matching: by searching for documents containing similar topics to the topics of some query words, or by searching for documents of similar topical content to a query document. Due to the topical representation, similarities between queries and documents can be discovered even with few actual word tokens in common. For mining behaviours in video, we assume that visual events correspond to words, simple actions (co-occurring events) correspond to topics, and complex behaviours (co-occurring actions) correspond to document categories.

We model the occurrence of a sequence of clips (documents) $D = \{\mathbf{x}_t\}$ 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 number of possible actions and behaviors in the dataset are assumed to be known and fixed as N_y and N_z respectively, although this assumption will be relaxed later in Sect. 3.5. The generative model is defined as follows: Suppose each clip t exhibits a particular category of behaviour z_t. The behaviour z_t is assumed to vary systematically over time from clip to clip according to some unknown discrete distribution, $p(z_t|z_{t-1}, \psi)$ (denoted Discr$(\cdot)$) with parameter ψ. Within each clip t, a bag of N_t simple actions $\{y_{i,t}\}_{i=1}^{N_t}$ are chosen (each 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 discrete distribution parameters $\{\phi, \psi, \theta\}$ are treated as Dirichlet distributed unknowns (denoted Dir$(\cdot)$) with symmetric hyper-parameters $\{\alpha, \beta, \gamma\}$. The complete generative model is specified by:

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

The joint distribution of variables $\{\mathbf{x}_t, \mathbf{y}_t, z_t\}_1^T$ and parameters θ, ψ, θ given the hyper-parameters α, β, γ is:

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

The hyper-parameters $\{\alpha, \beta, \gamma\}$ effectively specify an *a priori* belief about how sparsely the visual events, actions and behaviors are distributed, e.g., how much visual events should be shared between topics and actions shared between behaviors. These can be optimized during MCMC learning (Griffiths et al. 2007; Wallach et al. 2009), but as we observed they did not strongly affect our experimental results, we simply fix them as in Griffiths and Steyvers (2004), Rosen-Zvi et al. (2004), Gruber et al. (2007), Wang et al. (2009). In the next sections, we will describe the learning and inference procedures for this model. To learn the typical actions and behaviors in a dataset, we will compute the posterior over all the parameters, $p(\theta, \psi, \phi|\mathbf{x}_{1:t})$. For detecting the occurrence of learned behaviors in the video, we will compute the behavior posterior $p(z_t|\mathbf{x}_{1:t})$; and for detecting salient events, we will compute the predictive likelihood of each new data-point, $p(\mathbf{x}_t|\mathbf{x}_{1:t-1})$. The Bayesian parameter inference will increase robustness to over-fitting, and the intermediate action layer $\mathbf{y}_t$ will increase robustness to occlusion and noise compared to modeling visual events directly. The variety of discoverable behaviour patterns and sensitivity to saliency will be enhanced by the compositional hierarchical structure and Markovian behaviour model.

3.3 Model Inference and Learning

As for LDA, exact inference in our model is intractable, but it is possible to derive a collapsed Gibbs sampler (Gilks et al. 1995) for approximate MCMC learning of the parameters and inference of the latents $p(\mathbf{y}_{1:T}, z_{1:T}|\mathbf{x}_{1:T})$. The Dirichlet-Multinomial conjugate structure of the model allows the parameters $\{\phi, \theta, \psi\}$ to be integrated out automatically in a 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:

$$p(y_{i,t}|\mathbf{y}_{\setminus i,t}, z_{1:T}, \mathbf{x}_{1:T}) \propto \frac{n^-_{x_{i,t},y_{i,t}} + \beta}{\sum_x n^-_{x,y_{i,t}} + N_x\beta} \frac{n^-_{y_{i,t},z_t} + \alpha}{\sum_y n^-_{y,z_t} + N_y\alpha}. \quad (2)$$

Here $\mathbf{y}_{\setminus i,t}$ denotes all the variables $\mathbf{y}_{1:T}$ 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 excluding item (i, t). N_x is the size of the visual event codebook, and N_y the number of simple actions. Detailed derivation of all the learning equations are given in the Appendix.

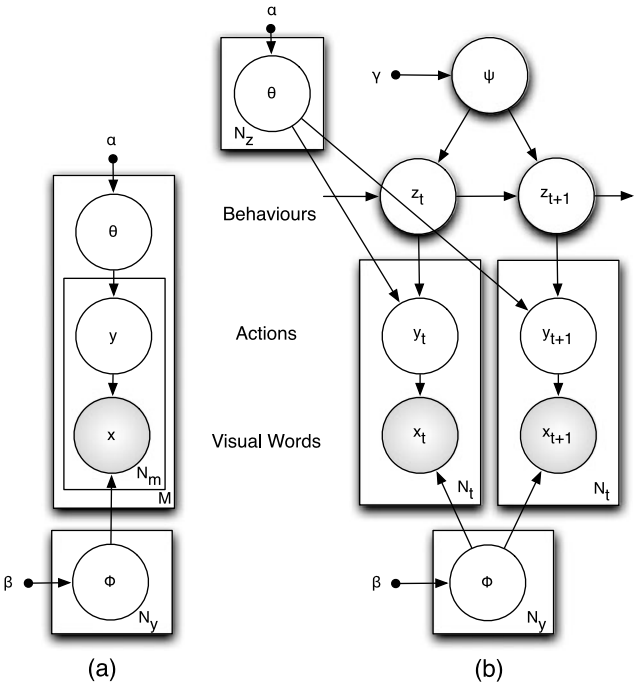


Fig. 1 Graphical models representing: (**a**) Standard LDA model (Blei et al. 2003), (**b**) Our MCTM model

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

$$p(z_t|\mathbf{y}_{1:T}, z_{\setminus t}, \mathbf{x}_{1:T}) \propto \frac{\prod_y \Gamma(n_{y,z_t} + \alpha)\,\Gamma(n^-_{\cdot,z_t} + N_y\alpha)}{\prod_y \Gamma(n^-_{y,z_t} + \alpha)\,\Gamma(n_{\cdot,z_t} + N_y\alpha)} \frac{n^-_{z_t,z_{t-1}} + \gamma}{n^-_{z_t,z_{t-1}} + N_z\gamma} \cdot \frac{n_{z_{t+1},z_t} + \mathrm{I}(z_{t-1} = z_t)\mathrm{I}(z_t = z_{t+1}) + \gamma}{n_{\cdot,z_t} + \mathrm{I}(z_{t-1} = z_t) + N_z\gamma}. \quad (3)$$

Here $n_{z',z}$ represents the counts of behaviour z' following behaviour z, $n_{\cdot,z} \triangleq \sum_{z'} n_{\cdot,z}$, and N_z is the number of clusters. 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 (Griffiths and Steyvers 2004) and (2), because the inclusive and exclusive counts may differ by more than 1, but this is not prohibitively costly, as (3) is computed only once per clip. Iterations of (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 a full

set of samples (Rosen-Zvi et al. 2004; Griffiths and Steyvers 2004):

$$\hat{\phi}_y^s = \frac{n_{x,y} + \beta}{n_{\cdot,y} + N_x\beta}, \tag{4}$$

$$\hat{\theta}_z^s = \frac{n_{y,z} + \alpha}{n_{\cdot,z} + N_y\alpha}, \tag{5}$$

$$\hat{\psi}_z^s = \frac{n_{z',z} + \gamma}{n_{\cdot,z} + N_z\gamma}. \tag{6}$$

3.4 Online Inference and Saliency Detection

A limitation of the sampling approach to learning and inference described above (also adopted by Li et al. 2008; Wang et al. 2009), is that they are offline, batch procedures. This is fine for a calibration step, however as we have seen, a subsequent key goal for real applications is online behavior classification and saliency detection. Given a learned scene profile ((4)–(6)), we therefore formulate a new online filtered inference algorithm for our MCTM which will permit on-the-fly classification of behavior and saliency detection.

Given a training dataset of T_{tr} clips, we have generated N_s samples $\{\{\mathbf{y}_t, z_t\}_{t=1}^{T_{\text{tr}}}, \hat{\phi}^s, \hat{\psi}^s, \hat{\theta}^s\}_{s=1}^{N_s}$ from the posterior distribution of latents in our model $p(\{\mathbf{y}_t, z_t\}_{t=1}^{T_{\text{tr}}} | \{\mathbf{x}\}_1^{T_{\text{tr}}}, \alpha, \beta, \gamma)$. We address online inference (rather than learning), and assume that no further adaptation of the parameters is necessary, i.e. the training dataset is representative, so $p(\phi, \psi, \theta | \mathbf{x}_{t'>T_{\text{tr}}}) = p(\phi, \psi, \theta | \mathbf{x}_{1:T_{\text{tr}}})$. We can 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 (7) with sums over their Gibbs samples (8) (since these are drawn from the required distribution $p(\phi, \psi, \theta|\mathbf{x}_{1:t})$). Conditioned on each set of (sampled) parameters, the other action $y_{i,t}$ and behaviour z_t variables decorrelate, so these can be summed out efficiently by in an online recursion for the behavior category of each clip:

$$p(z_{t+1}|\mathbf{x}_{1:t+1}) = \sum_{z_t} \int_{\phi,\theta,\psi} \frac{p(\mathbf{x}_{t+1}, z_{t+1}|z_t, \phi, \theta, \psi, \mathbf{x}_{1:t}) p(z_t, \phi, \theta, \psi|\mathbf{x}_{1:t})}{p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t})}, \tag{7}$$

$$\approx \frac{1}{N_s} \sum_{s,z_t} \frac{p(\mathbf{x}_{t+1}|z_{t+1}, \phi^s, \theta^s) p(z_{t+1}|z_t, \psi^s) p(z_t|\mathbf{x}_{1:t})}{p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t})}. \tag{8}$$

Bayesian saliency (or irregularity), is 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 (8), or explicitly as:

$$p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}) = \sum_{z_t} \int_{\phi,\theta,\psi} p(\mathbf{x}_{t+1}|z_t, \psi, \theta, \phi, \mathbf{x}_{1:t}) p(z_t, \phi, \psi, \theta|\mathbf{x}_{1:t}),$$

$$\approx \frac{1}{N_s} \sum_{s,z_{t+1},z_t} p(\mathbf{x}_{t+1}, z_{t+1}|\psi^s, \theta^s, \phi^s, z_t) p(z_t|\mathbf{x}_{1:t}). \tag{9}$$

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 the matrix multiplies and sums defined in (8) and (9). Note that in practice (8) may suffer from label switching (Bishop 2006; Gilks et al. 1995), so a single sample should be used for interpretable results (Griffiths and Steyvers 2004). Equation (9) is independent of label switches and should be used with all samples. This incremental approach has no direct analogy in vanilla LDA (Blei et al. 2003) (Fig. 1(a)), as the per document parameter θ requires iterative computation to infer. We compare the computational cost of our MCTM, LDA (Blei et al. 2003), Dual-HDP (Wang et al. 2009) and HMMs in Sect. 5.7. Finally, to account for the fact that clips t contain varying numbers N_t of visual events $x_{i,t}$ (and hence have varying "base" probability), when searching for saliency in the test data, we compute a normalized predictive likelihood π for each clip:

$$\log \pi(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}) = \frac{1}{N_{t+1}} \log p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}). \tag{10}$$

Our measure of saliency (predictive likelihood) $\pi(\mathbf{x}_t|\mathbf{x}_{1:t-1})$ of test clip $\mathbf{x}_t$ given training video data $\mathbf{x}_{1:T_{\text{tr}}}$ and previous test data $\mathbf{x}_{t-1>T_{\text{tr}}}$ indicates irregularity if it is low. $\pi(\mathbf{x}_t|\mathbf{x}_{1:t-1})$ can be low for four different reasons, reflecting the following salient aspects of the data:

Events: Individual $x_{i,t}$ rarely occurred in training data $\mathbf{x}_{1:T_{\text{tr}}}$.
Actions: Events in $\mathbf{x}_t$ rarely occurred together in the same activity in $\mathbf{x}_{1:T_{\text{tr}}}$.
Behaviours: $\mathbf{x}_t$ occurred together in topics, but such topics did not occur together in clusters in $\mathbf{x}_{1:T_{\text{tr}}}$.
Dynamics: $\mathbf{x}_t$ occurred together in a cluster z_t, but z_t did not occur following the same clusters z_{t-1} in $\mathbf{x}_{1:T_{\text{tr}}}$.

Such detections are made possible because the hierarchical structure of our model represents behaviour at different levels (events, actions, behaviours, behaviour dynamics). In Sect. 5.3, we will refer to clips with rare events as *intrinsically* unlikely, those with rare actions and behaviors as *behaviourally* unlikely, and those with rare behavioural dynamics as *dynamically* unlikely. For convenience, Algorithm 1 summarizes the basic learning and inference procedure for our MCTM.

3.5 Model Order Determination

We have discussed methods for learning and inference in our model. The final question a user might ask is what if there is so little prior domain knowledge about a space to surveil that one cannot even estimate the number of typical actions

Algorithm 1 MCTM Algorithm Summary

Learning (Offline)
Input: Visual event detections for every clip, $\{\mathbf{x}_t\}_{t=1}^{T_{\text{tr}}}$.
Initialize $\{z_t, \mathbf{y}_t\}_{t=1}^{T_{\text{tr}}}$ randomly.
Repeat $N_{iter} = 1000$ times:

– For every time t:

1. Resample $p(z_t|\mathbf{y}, z_{\setminus t}, \mathbf{x})$ (3).
2. For every observation i at t:
   – Resample $p(y_{i,t}|\mathbf{y}_{\setminus i,t}, z_{1:T}, \mathbf{x})$ (2).
3. At every 100th iteration:
   – Record independent sample $s = \{z_t, \mathbf{y}_t\}_{t=1}^{T_{\text{tr}}}$.
   – Estimate model parameters ((4), (5) and (6)).

Output: Parameter estimates $\{\hat{\phi}^s, \hat{\psi}^s, \hat{\theta}^s\}_{s=1}^{N_s}$.

Inference for a new clip t (Online)
Input: Parameter samples $\{\hat{\phi}^s, \hat{\psi}^s, \hat{\theta}^s\}_{s=1}^{N_s}$, previous posterior $p(z_{t-1}|\mathbf{x}_{1:t-1})$, visual event detections $\mathbf{x}_t$.

1. Compute behavior profile $p(z_t|x_{1:t})$ (8).
2. Compute saliency $\pi(x_t|x_{1:t-1})$ (10).

Output: Behavior inference $p(z_t|x_{1:t})$, saliency $\pi(x_t|x_{1:t-1})$.

or behaviours exhibited in the scene, and hence cannot specify the model complexity $M = \{N_y, N_z\}$? In this section, we show how these model complexity parameters can be automatically determined from the data.

From a Bayesian modeling perspective (Bishop 2006; Gilks et al. 1995), the ideal model M should be selected as the one which maximizes the marginal likelihood $p(D^{\text{tr}}|M)$ of the training data D^{tr}; or the marginal likelihood of a held-out test dataset D^{te} given the training data $p(D^{\text{te}}|D^{\text{tr}}, M)$.[1] The challenge is that this requires integrating out any latent variables in the model ($\mathbf{y}_{1:T}$ and $z_{1:T}$ in our case). In sampling approaches to learning, a model's marginal likelihood, $p(D|M)$ is often estimated by the harmonic mean of the Gibbs sample likelihoods (Gilks et al. 1995). The harmonic mean approach is commonly used in practice because of its simplicity and efficiency (Griffiths and Steyvers 2004), although it is well known to be highly unstable (Gilks et al. 1995; Wallach et al. 2009). In our case, its extreme variance rendered it useless for model comparison. Instead, we use the test dataset likelihood

$$p(D^{\text{te}}|D^{\text{tr}}, M) = \prod_{t=1}^{T_{\text{te}}} p(\mathbf{x}_t^{\text{te}}|\mathbf{x}_{1:t-1}^{\text{te}}, D^{\text{tr}}, M), \tag{11}$$

[1]Note that the marginal likelihood of each model is affected by the chosen Dirichlet hyper-parameters (α, β, γ), which can also be optimized for each model during sampling using the Gibbs-EM method in Wallach et al. (2009); or fixed and the most likely model under this constraint can be determined (Griffiths and Steyvers 2004).

which is efficiently computable for our model using the predictive likelihood (9) derived in Sect. 3.4. The model to use is then chosen as $M^* = \text{argmax}_M\, p(D^{\text{te}}|D^{\text{tr}}, M)$ with confidence $p(D^{\text{te}}|D^{\text{tr}}, M^*)/p(D^{\text{te}}|D^{\text{tr}}, M')$ where M' is the second most likely model. Since models M in our framework have complexity varying with both N_y and N_z, this requires learning and selecting from a two dimensional grid of models across a large range of N_y and N_z. This is computationally expensive, but need only be done offline and once per dataset.

3.6 An Illustrative Example

To convey intuition into the modeling assumptions, computational mechanisms and capabilities of our model, we now illustrate its behaviour on a simple synthetic dataset. This will also allow us to verify its behaviour in a situation where there is an obviously correct interpretation of the data. We consider data of the flavor that might be found at a traffic intersection. Figure 2(a) illustrates the four true prototype behaviours in this dataset, corresponding to variants of "crossing" and "turning". Importantly, and similarly to some real intersections, only two possible turns are allowed ("bottom left" and "top right" in this case), and the behaviours occur in a particular order (illustrated by the arrows). By sampling from the generative model, we obtain the ordered samples illustrated in Fig. 2(b), which will be the input data $\mathbf{x}_{1:T}$. Note that in the prototype and sample illustrations, brighter cells mean more observations are likely and observed respectively. In the following analysis we assume for simplicity that the model order is known to be $N_y = N_z = 4$, although we have seen how to compute this as well (Sect. 3.5).

Model Learning Our model can concisely explain this dataset (Fig. 2(b)) in terms of an ordered sequence of behavior patterns $z_{1:T}$ each made up of a constrained combination $\mathbf{y}_t$ of prototype actions ϕ. To illustrate this, we generated $T = 1000$ training examples $\mathbf{x}_{1:T}^{\text{tr}}$ including those in Fig. 2(b), and performed unsupervised learning in our model by Gibbs sampling $p(\mathbf{y}_{1:T}, z_{1:T}|\mathbf{x}_{1:T}^{\text{tr}})$ ((2) and (3)). These samples encode distributions over the parameters $\{\phi, \theta, \psi\}$, which can be estimated for visualization by (4)–(6). The learned actions $\hat{\phi}$ correctly represent the correlated activity along each of the "roads" leading to the intersection (Fig. 2(c)), while the learned behaviors $\hat{\theta}$ selectively compose particular actions (arrows) to represent exclusively the four permitted crossing and turning behaviours (Fig. 2(d)). The typical order in which behaviours occur is discovered correctly and encoded in the transition matrix $\hat{\psi}$ (compare arrows indicating probable transitions in Fig. 2(a), (d) and (e)). Note that the $N_z \times N_z$ transition matrix in Fig. 2(e) should be interpreted as columns representing starting behaviors, and the brightness of each row indicating how likely each subsequent behavior is.

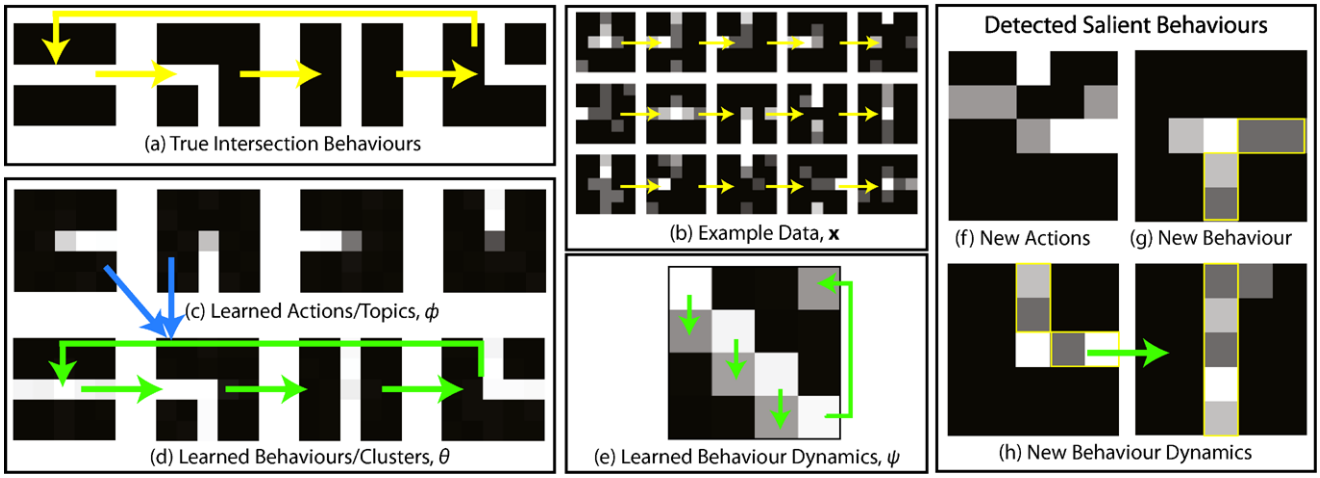


Fig. 2 Illustration of how our model works using a synthetic data example. (**a**) True prototype "intersection" behaviours used to generate (**b**) noisy input data, $\mathbf{x}_{1:T}$. Learned model parameters are (**c**) actions/topics ϕ which are composed into (**d**) behaviours/clusters θ, correlated by (**e**) behaviour dynamics transition matrix ψ. (**f–h**) Salient examples detected by the learned model

Alternative Models We can contrast the MCTM's representation of this data against that of two related models, a hidden Markov model (HMM) (Bishop 2006), and vanilla LDA (Blei et al. 2003). The HMM successfully learns clusters similar to Fig. 2(d) directly without the intermediate representation of actions. In contrast, LDA learns only a topic/action based representation of the data similar to Fig. 2(c), without knowledge of their composition constraints. The HMM learns a behavior dynamics similar to Fig. 2(e), while LDA learns no temporal model.

As we have seen in Sect. 3.5, alternative models should be quantitatively compared by the marginal likelihood of a set of held out test data $\mathbf{x}^{\text{te}}_{1:T}$ under each model after training, $p(\mathbf{x}^{\text{te}}_{1:T}|\mathbf{x}^{\text{tr}}_{1:T}, M)$ ((9) for MCTM). Table 1 summarizes the relative marginal likelihood of the test set for each of the three models—for which a larger value indicates a better representation and improved generalization. To understand why MCTM provides the best representation, consider that by not learning correlation between topics, LDA wastes some of its predictive probability mass on behaviour events which never happen (e.g., "top left turns"). (Note that this moreover means it cannot detect such events as salient.) The HMM learns the clusters and their dynamics directly without the intermediate topic representation. The point-estimation of parameters leads to over-fitting and poor performance by the HMM compared to the MCTM. More subtly, the HMM is not the best representation of even this simple data, because it does not fully exploit the shared structure of the behaviors. For example, the horizontal crossing and bottom right turn behaviours *share* a component: the "right road" activity. Therefore this component of the each behaviour should be represented jointly by the same parameter to minimize over-fitting and maximize parameter learning accuracy. This is exactly what MCTM does and the HMM does not.

Saliency Detection To illustrate salient event detection, we randomly inserted 15 atypical frames into a 1000 sample test set. The model identified salient samples by thresholding $\log p(\mathbf{x}^{\text{te}}_t|\mathbf{x}^{\text{te}}_{1:t-1}, \mathbf{x}^{\text{tr}}_{1:T}, M)$ at the 2% level. Overall, 10 of the 15 atypical frames were in the top 2% most salient frames detected by MCTM. We highlight an illustrative selection of four events in the test dataset which were detected as salient (Fig. 2(f)–(h)). Figure 2(f) illustrates a very noisy sample which included three previously unseen visual events. Figure 2(g) illustrates a sample which contained common events in correlations corresponding to the learned actions (yellow highlights). The correlation in these actions ("bottom right turn"), however, did not correspond to a learned behaviour (Fig. 2(d)). Finally, Fig. 2(h) illustrates two sequential frames which individually contained valid actions and behaviours (yellow highlights), but which were unlikely to occur in sequence (compare Fig. 2(d)). In our experiments, the LDA model did not detect the salient frames illustrated in Figs. 2(g) and (h) because recognizing their saliency necessitates a model of correlation between actions and temporal behavior dynamics respectively—which are not modeled by LDA. This is significant, as the examples in Figs. 2(g) and (h) might correspond to dangerous real-life behaviors such as running a red light.

4 Semi-supervised Learning for Behavior Detection

4.1 Supervised vs. Unsupervised Behavior Categorization

We have considered learning an unsupervised behavior model for a scene (Sect. 3.3), and using this to detect salient

Table 1 Log-likelihood of synthetic test data under each model

Model	Relative Log-likelihood
MCTM	0
HMM	-30 ± 7
LDA	-1704 ± 205

video segments as those with unexpected events, actions or behaviors (Sect. 3.4, (9)); and to classify video segments into one of the common behavior classes according to the observed activity (Sect. 3.4, (8)). Our unsupervised approach to these tasks is appropriate for a new and unknown scene as hands-off deployment of a system to discover typical behaviors and track their statistics as well as find unusual events is desirable.

In a context with more prior knowledge about the scene, a slightly different task can be posed which our model can also address: that of detecting known moderately unusual behaviors based on a few labeled examples. Consider the situation in which there are some particular behavior(s) which are known a-priori to occur from time to time in the monitored public space, and which we are interested to detect future instances of—for example, common driving violations. In this case, saliency is no longer necessarily defined by extreme irregularity and hence low likelihood, so our detection approach in (9) does not apply. At the same time, topics or behaviors specifically representing the situation of interest are unlikely to be allocated be the unsupervised learning framework,[2] so the classification approach in (8) is unlikely to work either.

For these reasons, some kind of human input is therefore necessary to define a known behavior as interesting for detection. Performing fully supervised learning, however requires prohibitive manual effort for two reasons: (i) labeling is time-consuming and expensive, and (ii) extra manual analysis of the scene is required, because examples of all the typical clusters are needed for supervised learning, not just examples of the single salient behavior of special interest. As a good compromise, we will show next how our framework can be used in a semi-supervised manner to learn to detect particular behaviors of interest from a few labeled examples with minimal manual effort.

[2]To see why, consider that the framework as described in Sects. 3.3 and 3.4 is used to sample a likely hierarchical clustering of the data into actions and behaviors according to $p(z_{1:T}, \mathbf{y}_{1:T}|\mathbf{x}_{1:T})$. Given this objective function—which purely tries to find a good density estimate for $\mathbf{x}_{1:T}$—a particular behavior of interest to a human user is unlikely be allocated its own unique behaviour cluster if it is under-represented in the data, or if it is visually very similar to another more common behavior.

4.2 Semi-supervised Learning in MCTM

To perform semi-supervised learning with our model, a few examples of behavior(s) of interest can be labeled in the training set. The only modifications to Algorithm 1 required are that prior to learning, the supervised samples l are initialized to their correct behaviors l_z rather than randomly, and during learning the labeled behavior examples are considered observed, and not updated by the sampler (3). All the action layer variables (2) are still updated in the same way to find a good intermediate action representation under the additional constraint of the specified behavior labels. The model thus samples from $p(z_{\setminus l}, \mathbf{y}_{1:T}|z_l, \mathbf{x}_{1:T})$ rather than $p(z_{1:T}, \mathbf{y}_{1:T}|\mathbf{x}_{1:T})$ as before. It is possible to label some examples of a specific behavior of interest, and allow the model to determine the best clustering of the others under this constraint, or to define all the interesting behaviors and label examples of each.

5 Experiments

5.1 Datasets and Settings

Experiments were carried out using video data from four complex and crowded public scenes. **QMUL 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. 4(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. 4(f)–(i)). Typical behaviours here are pedestrian crossings alternating with two main traffic flows. **Subway Platform Dataset:** A total of 30 minutes of videos from the UK Home Office i-LIDS dataset (2) 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. 4(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. **MIT Traffic Dataset:** This contains 90 minutes of 360×288 pixel 30 fps video of a street corner (previously studied by Wang et al. 2009). Here the traffic flow is less busy but less regulated than the first street intersection dataset, and there are three pedestrian crossings visible.

We used 5 minutes from each dataset for training, and tested ((8) and (10)) on the remaining data. The cell size for both of the three street datasets was 8×8, and 16×16

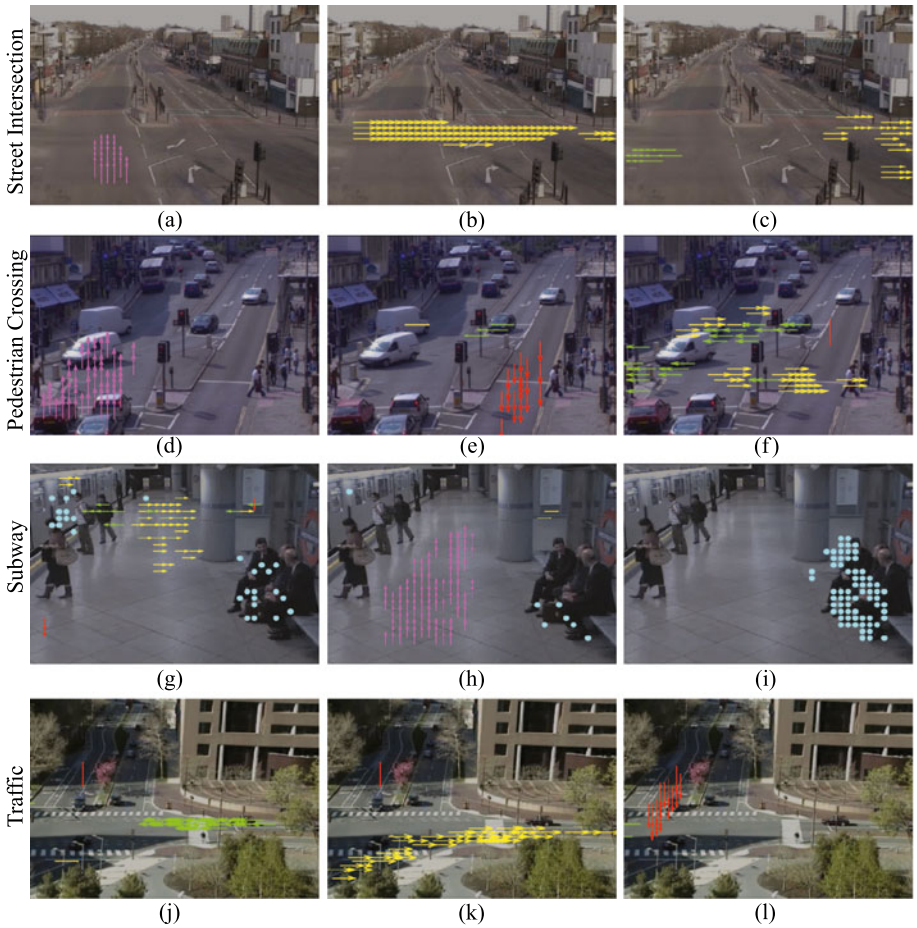


Fig. 3 (Color online) Example topics/actions learned in each of the three scenarios illustrated by the most likely visual events for each $\hat{\phi}_y^s$. *Arrow* directions and *colors* represent flow direction of the event

for the subway dataset due to the bigger frame size. Optical flow computed in each cell was quantized into 4 directions for the street datasets and 5 for the subway dataset, with the fifth corresponding to stationery foreground objects common in the subway scene. The clip length for collecting the bag of words was defined as 1 second for the first three datasets, and 10 seconds for the traffic dataset to facilitate comparison with (Wang et al. 2009). We ran the Gibbs sampler ((2) and (3)) for a total of 1000 sweeps, discarding the first 500 as burn-in, and then taking 5 samples at a lag of 100 as independent samples of the posterior $p(\{\mathbf{y}_t, z_t\}_1^T | \mathbf{x}_{1:T_{\text{tr}}}, \alpha, \beta, \gamma)$.

Although we can learn suitable model complexity $\{N_y, N_Z\}$ (Sect. 3.5), for ease of illustration we specify the number of actions and behaviors in each case. We set the number of actions to $N_y = 8$ and number of behaviours as $N_z = 4$; except for the pedestrian crossing dataset, where we used $N_z = 3$ because there are clearly three traffic flows, and $N_z = 5$ for the traffic dataset to facilitate comparison with Wang et al. (2009). The significance of these complexity parameters is that larger N_y and N_z induce a more fine-grained decomposition of scene behaviour. Dirichlet hyperparameters were also fixed at $\{\alpha = 8, \beta = 0.05, \gamma = 1\}$ for all experiments to encourage composition of specific actions into general topics (Griffiths and Steyvers 2004; Rosen-Zvi et al. 2004), but these can be also be estimated during sampling by the method in Wallach et al. (2009).

5.2 Unsupervised Scene Interpretation

Clustering Visual Events Into Actions The actions learned by our MCTM correspond to co-occurring visual events. These actions are typically associated with patterns of moving objects. Figure 3 shows some example actions y discovered by way of plotting the visual events $\mathbf{x}$ in the top 50% of the mass of the distribution $p(\mathbf{x}|y, \hat{\phi}_y^s)$ (4). Note that each action has a clear semantic meaning. In the street intersection dataset, Figs. 3(a) and (b) represent vertical left lane and horizontal rightwards traffic respectively, while Fig. 3(c) represents the vertical traffic vehicles turning right at the filter. In the pedestrian crossing dataset, Figs. 3(d) and (e) illustrate two independent vertical traffic flows, and Fig. 3(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. 3(g) includes people leaving (yellow arrows) from a stopped train (cyan dots on the train). Figure 3(h) includes people walking up the platform and Fig. 3(i) shows people sitting on the bench waiting. Finally,

for the traffic dataset Fig. 3(j) represents horizontal traffic approaching from the right, and Figs. 3(k) and (l) represent right turns by vertical traffic approaching from below and above respectively.

Discovering behaviours and their dynamics Co-occurring topics are automatically clustered into behaviours z via vector θ_z (5), each of which corresponds to a complex behaviour pattern involving multiple interacting objects. Complex behaviour clusters discovered for the four dynamic scenes in the 5 minutes of training data are depicted in Fig. 4. Specifically, Figs. 4(a) and (b) represent horizontal left and right traffic flows respectively, including right turn traffic (compare horizontal only traffic in Fig. 3(b)). Figures 4(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. 4(e)).

For the pedestrian crossing dataset, three behaviour clusters are learned. Figure 4(f) shows 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). Figures 4(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. 4(i), top left).

In the traffic dataset, behaviours representing straight horizontal flow are learned in Figs. 4(j). Note that left and right flows are together in a single cluster because they occur together, unlike the same flows in the street intersection dataset (Figs. 4(a) and (b)). Horizontal traffic can also turn, which is modeled by the behavior in Fig. 4(k). The remaining three behaviors are vertical traffic flows. Figure 4(l) represents straight vertical flow for both lanes. Figures 4(m) and (n) represent turning right from above and below respectively. Since the first two (horizontal) and second three (vertical) flows in this scene can occur in any sequence, the transition matrix dynamics (Fig. 4(o)) are less clearly structured. One feature of the matrix is the high probability of off-diagonal transitions between the first two behaviors (Fig. 4(o); 'j' and 'k'), indicating that the horizontal traffic is light enough that there is regular alternation between flowing straight and turning traffic, rather than having many turning cars queue up and wait for a break to turn which resulted in the typically longer periods in the intersection dataset (Fig. 4(e)).

The four behaviour categories discovered in the subway scene were: People walking towards (yellow & red arrows) an arriving train (green arrows on train) (Fig. 4(p)); People boarding a stopped train (cyan dots on the track) or leaving the station (Fig. 4(q)); People leaving the station while the trains wait (Fig. 4(r)) (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. 4(s)). Our model is also able to discover the cycle of behaviour on the platform triggered by arrival and departure of trains (Fig. 4(t)). For example, the long duration of waiting periods between trains, broken primarily by the train arriving state (p), (Fig. 4(t), fourth column).

5.3 Online Video Screening

After the model was learned for each scenario, new video data was screened online. The overall behaviours were identified using (8), and visual saliency (irregularity) measured using (10). Figure 5 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 $\pi(\mathbf{x}_t|\mathbf{x}_{1:t-1})$ 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 2% most surprising clips (obtained by ranking $\pi(\mathbf{x}_t^{\text{te}}|\mathbf{x}_{1:t-1}^{\text{te}}, \mathbf{x}_{1:T}^{\text{tr}})$) screened by the model in the test data. Here we discuss some examples of flagged surprises. In Fig. 6(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*, (Sect. 3.4) unlikely, as these events were rare in the training data under any circumstances. In Fig. 6(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. 4(e)). Therefore this whole series of clips is *behaviorally* and *dynamically* unlikely given global and temporal constraints entailed by $\pi(\mathbf{x}_t|\mathbf{x}_{1:t-1})$.

Another *behavioral* (action concurrence) and *dynamic* surprise to the model is the jay-walker in Fig. 6(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. Figure 6(g) shows the detection of a jay-walker triggered by *intrinsically* unlikely horizontal motion across the street. In contrast, Fig. 6(h) illustrates two plausible pedestrian actions of crossing left and

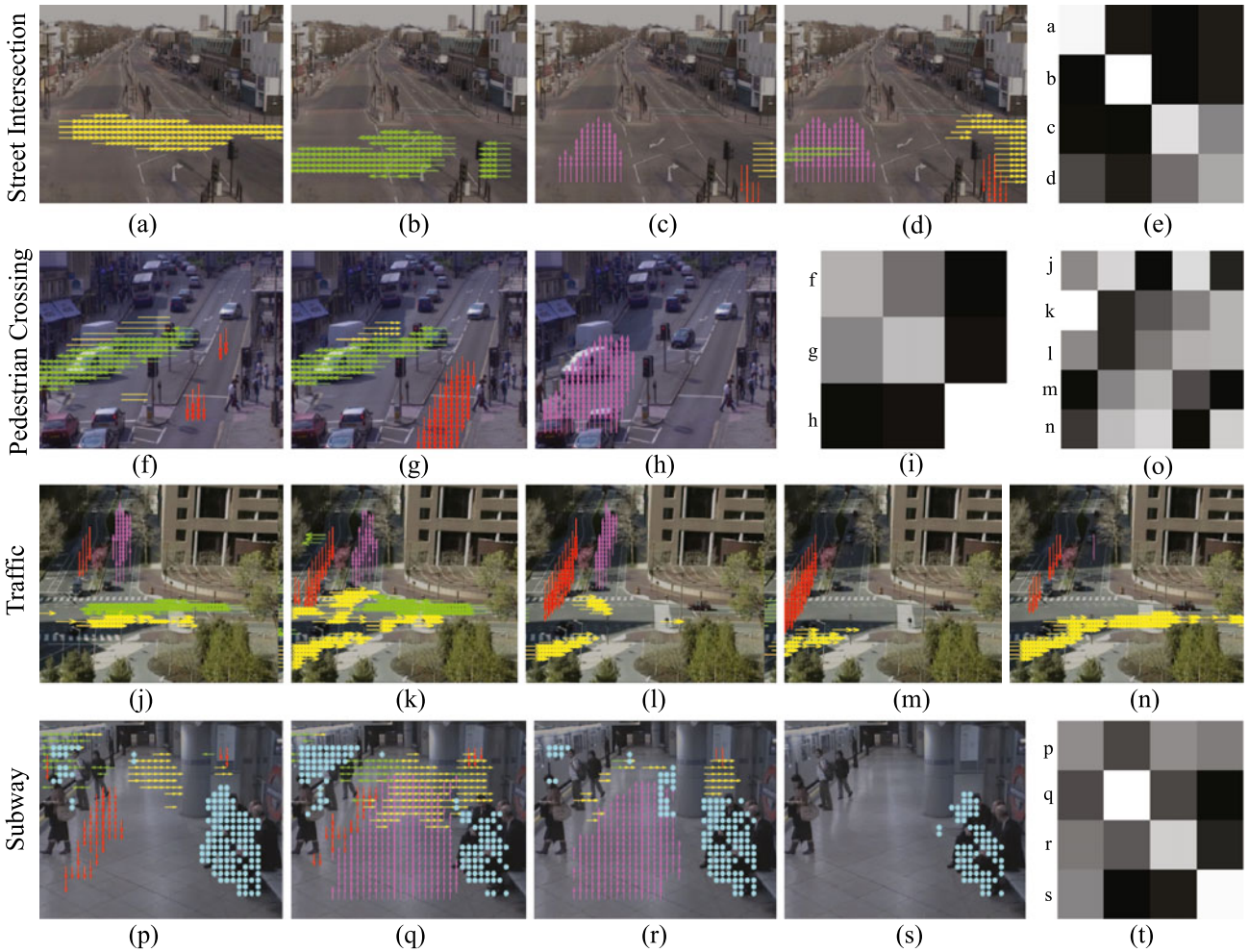
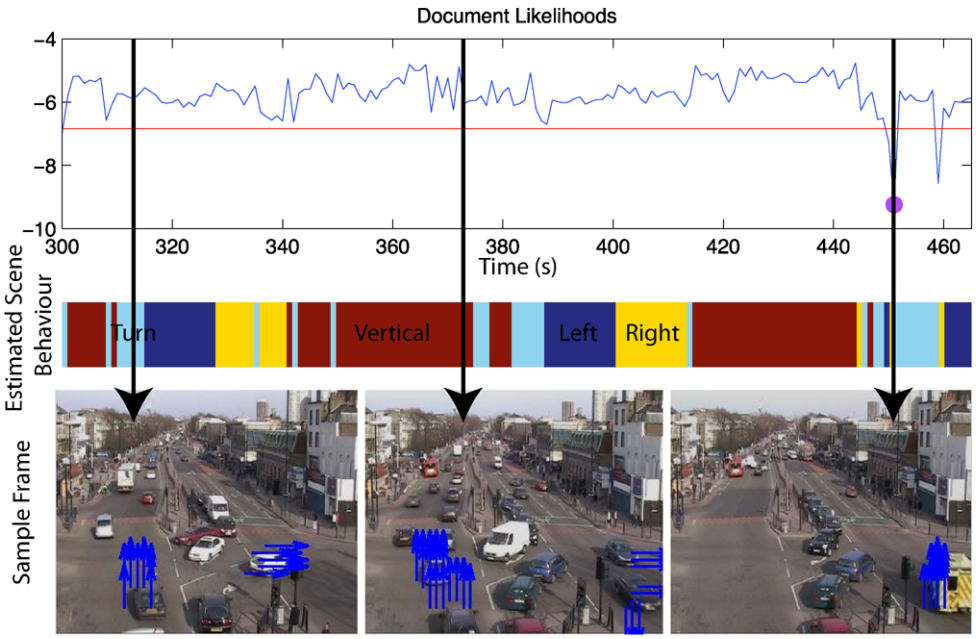


Fig. 4 Behaviour and dynamics learn from each of the four scenarios. The most likely visual words/events for each behaviour $\hat{\theta}_z^s$ are illustrated. Dynamics $\hat{\psi}_z^s$ are illustrated by transition matrices (**e**), (**i**) and (**n**) where *brighter shading* corresponds to greater probability of behavior transition and *row labels* refer to the associated behavior figure

right at the crosswalk, but occurring at the same time as the vertical traffic flow. This is multi-object situation is *behaviorally* irregular. In Fig. 6(i) a train arrives, and three people typically (Fig. 4(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. 6(k), the train is now stationary. While most people perform the typical paired action of boarding (Fig. 4(k)), others walk away from the train down the platform, a multi-object *behaviour* detected due to low likelihood $\pi(\mathbf{x}_t|\mathbf{x}_{1:t-1})$.

Figures 6(c–f) illustrate an important feature of our model that gives a significant advantage over non-temporal LDA based models (Li et al. 2008; Wang et al. 2009): 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. 6(a) and (b) that might be lost in larger bags, but also longer time-scale events such as Fig. 6(c–f) that could be lost in many independently distributed smaller bags.

For the traffic dataset we illustrate the top 5 most unlikely clips discovered by our model for the purposes of direct comparison with Wang et al. (2009) who evaluate their HDP model in this way on this dataset. Intrinsically unlikely events include a car cutting into the other lane in making a left turn (Fig. 7(a)) and a pedestrian jay-walking across the bottom street (Fig. 7(d)). More interestingly, the other three clips are unlikely behavioral concurrences involving near collisions: a pedestrian crossing through traffic start-



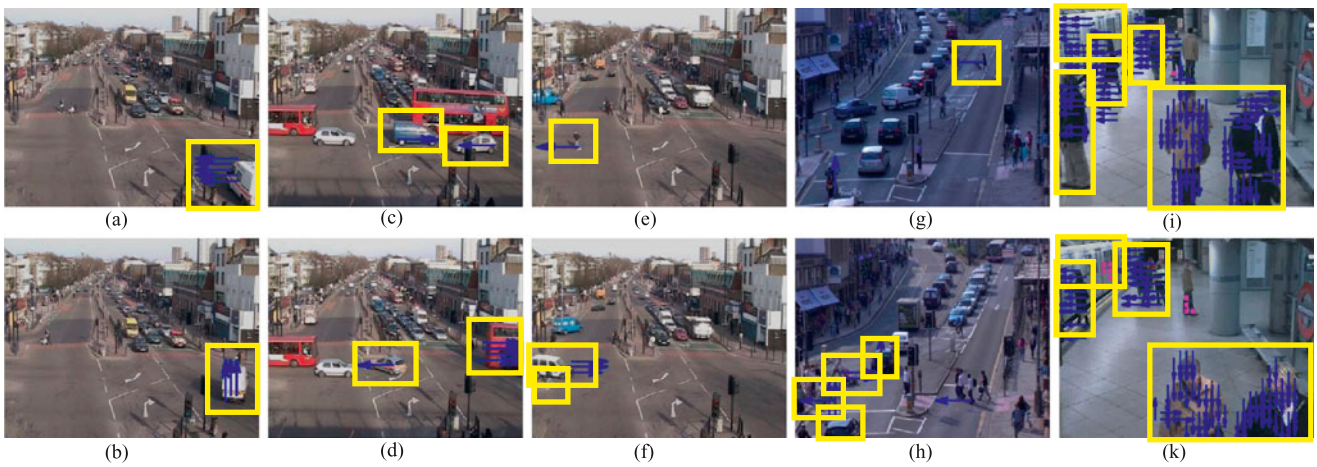


Fig. 5 An example of online processing

Fig. 6 Sample salient clips discovered. *Arrows/dots* indicate input events and *boxes* highlight regions discussed in the text

ing to flow (Fig. 7(b)); and pedestrians on the crossing very near horizontal (Fig. 7(c)) and vertical (Fig. 7(e)) traffic. The reported top 5 most unlikely clips discovered by the HDP model in (Wang et al. 2009) reflected more obvious single-object events of pedestrians and cyclists jay-walking. In contrast, the three near collisions ranked in the top 5 by our MCTM suggesting that it has greater sensitivity to more subtle and interesting multi-object behaviors than (Wang et al. 2009).

Quantitative Evaluation To demonstrate the breadth of irregular behavioral 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 Tables 2–5. We compare the results with two alternatives, LDA (Blei et al. 2003) with N_y topics, and a HMM with N_z states. The LDA model was trained using the visual event counts directly like MCTM, and the HMM learned a N_x dimensional Gaussian distribution over the count histograms for each clip. 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. events, action, behaviour, dynamics) clips were surprising to the model, so we do not attempt to quantify this. Table 2 shows that for the street intersection dataset, our MCTM outperforms the other two models especially in the more complex behaviour categories of red-

(a) (b) (c) (d) (e)

Fig. 7 Top 5 salient clips discovered by MCTM for the traffic dataset. *Arrows* indicate unusual actions or behaviors

Table 2 Summary of meaningful clip types discovered by each model for the street intersection dataset. Overall true positive (interesting) and false positive rates (uninteresting) are also given

Street Intersection	MCTM	LDA	HMM	Total
Break Red Light	4	0	3	13
Illegal U-Turn	5	2	1	15
Jaywalking	1	0	0	1
Drive Wrong Way	12	14	12	15
Unusual Turns	6	1	4	10
Uninteresting	27	38	35	2663
Overall TPR	52%	31%	37%	
Overall FPR	1.0%	1.4%	1.4%	

Table 3 Summary of meaningful clip types and detection rates discovered by different models for the pedestrian crossing dataset

Pedestrian Cross	MCTM	LDA	HMM	Total
Jaywalking	17	11	9	33
Through Traffic	9	5	3	16
Uninteresting	29	39	43	2674
Overall TPR	53%	33%	24%	
Overall FPR	1.1%	1.5%	1.6%	

Table 4 Summary of meaningful clip types and detection rates discovered by different models for the traffic dataset

Traffic	MCTM	LDA	HMM	Total
Jay-walking	4	4	4	20
Out of Lane	1	1	0	1
Near Collision	3	3	2	8
Uninteresting	3	3	5	510
Overall TPR	27%	27%	21%	
Overall FPR	0.6%	0.6%	0.9%	

light-breaking, u-turns and jay-walking. 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 (Fig. 6(a)).

For the pedestrian crossing dataset, the result is shown in Table 3. Atypical pedestrian behaviours were jay-walking far from the crosswalk (intrinsically unlikely visual events), and crossing at the crosswalk but through oncoming traffic (unlikely action concurrence; Fig. 4(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.

The traffic dataset detection results are summarized in Table 4. The most frequently detected unusual events by all models were the rather obvious and intrinsically unlikely pedestrian jay-walking events. Our MCTM and LDA were slightly better than the HMM at detecting the subtler intrinsic event of a car drifting out of lane, and the behavioural concurrence events of near collisions between crossing cars and people on the pedestrian crossings.

Finally, for the subway dataset (Table 5) the only interesting behaviours observed were people moving away from the train during clips where typical behaviour was approaching trains and boarding passengers. These were detected by our model and not by the others. Further results and online processing illustration can be seen at: http://www.eecs.qmul.ac.uk/~tmh/MCTM/.

5.4 Online Behavior Classification

Thus far we have evaluated our MCTM's performance at detecting various kinds of unusual salient behaviors. Users may be also be interested in monitoring the statistics of different common behaviours in the surveilled space online. Our model can provide this information for no further computational cost as classification is essentially performed as a

Table 5 Summary of meaningful clip types and detection rates discovered by different models for the subway platform dataset

Subway Platform	MCTM	LDA	HMM	Total
Contraflow	2	0	0	2
Uninteresting	36	38	38	1155
Overall TPR	100%	0%	0%	
Overall FPR	3.1%	3.3%	3.3%	

byproduct of saliency detection (8). To evaluate the classification ability of our MCTM, we created ground truth for the traffic flow category in the street intersection scene (left, right, vertical, vertical turning). For comparison, we evaluated the ability of the LDA and HMM models (as described in Sect. 5.3) to classify the same states. Note that vanilla LDA does not provide a classification directly, so we post-processed the posterior topic profile of the training data with K-means ($K = N_z = 4$) to learn a complete classifier. For testing we computed the topic profile for each clip, and used the learned K-means topic classifier to estimate the behavior category. Confusion matrices for each model are given in Table 6. MCTM outperforms the other models with 78% average classification accuracy across all the classes compared to 69% for both LDA and HMM.

5.5 Semi-supervised Behavior Learning

To see how semi-supervised learning facilitates behavior detection, we consider an example from the street intersection dataset. In the real intersection there are four regulated light phases: left, right, vertical only and a vertical filter stage—where only turning vertical traffic is permitted. For this data, the fourth cluster learned by the unsupervised model in Sect. 5.2 is a *mixture* of straight vertical traffic and turning vertical traffic (Fig. 4(d)). This is a sensible clustering given the statistics of the data, because the filtering only stage is very brief, and also very similar to the illegal but moderately common behavior of vertical traffic turning *through* the opposing flow. Assume that we are interested in detecting explicitly the known illegal behavior of turning through the oncoming flow. To do this, the behavior needs to be represented explicitly and separately to both the vertical flow and to the filtering only stage, necessitating five behaviours in total.

To show that unsupervised learning is inadequate, we tried to learn an unsupervised model with $N_z = 5$ behaviors, but it does not converge to the desired solution because without prior information, there are various other likely ways to partition this dataset into 5 behaviors. For example, Figs. 8(a)–(e) illustrate an unsupervised clustering of behaviors which represents instead which *lane* vertical traffic used (Fig. 8(c) vs. (d)) or whether the vertical up flow occurred in conjunction with vertical down flow (Fig. 8(d) vs. (e)). Because the filter stage is very brief and relatively under represented in the dataset, and because it is very similar to the illegal—but not entirely unusual—behaviour of turning through oncoming traffic, the unsupervised learner does not cluster behaviors on these grounds.

To solve the task scenario in which we care specifically about detecting this "common but illegal" behavior, we specified ground truth (behaviour cluster z_t) for 18 example clips l (5% of the training data) of the vertical traffic turning *through* the oncoming flow, and re-learned the model. As illustrated in Fig. 8, the semi-supervised approach now builds a model for this behaviour class (Fig. 8(j)), as well as the other most likely classes *under this constraint* (which now sensibly represent vertical only and filtering only stages (Fig. 8(h) and (i)). The model learning is thus guided by domain knowledge towards a maxima in which the interesting behaviour of turning through oncoming traffic, is represented as a cluster, and can be discovered explicitly by evaluating $p(z_t = 5|\mathbf{x}_{1:t})$. Of course, a few examples of all clusters can also be specified to define an exhaustive set of specific behaviours if the statistics of the whole set of behaviours are of interest. Note that this also alleviates the label switching problem in MCMC (Gilks et al. 1995; Griffiths and Steyvers 2004), so that the turn-through behaviour will be represented by the same specified cluster in every sample.

5.5.1 *Known Behavior Classification*

We evaluated the performance of our semi-supervised learning approach by repeating the classification task of Sect. 5.4, but with the five classes of interest. The results are quantified in Table 7. Clearly our semi-supervised approach has the best performance overall (81%) average accuracy. This is because of the crucial improvement of performance in distinguishing the easily confuse-able situations of traffic turning at the lights (Fig. 8(i)) and the interesting dangerous behavior of turning through other oncoming traffic (Fig. 8(j)). In contrast, the other models' optimization criteria of unconstrained likelihood-maximization leads them to cluster the data in ways which do not distinguish the behavior of interest (e.g., Fig. 8(a)–(e), T and VT confusion).

The potential for semi-supervised use of our model is important, because it allows it to be used very flexibly: for fully automatic profiling and saliency detection without prior domain knowledge; or by cheap semi-supervised specification of some or all of the behaviors of interest. This increases the potential sensitivity of the framework by allowing the model to learn and hence flag the occurrence of both "known but interesting behaviors" (by explicitly inferring their occurrence, (8)) as illustrated here, and also "unknown and unusual" behaviors (via their low predictive likelihood (10)) as illustrated in Sect. 5.3.

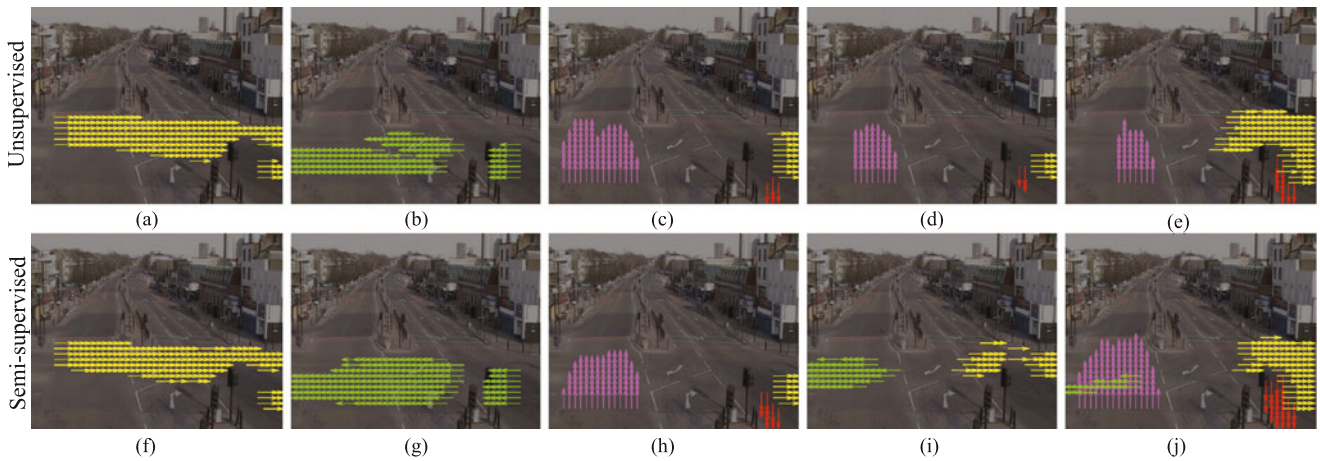


Fig. 8 Different behaviours learned for the street intersection dataset using (**a**)–(**e**) unsupervised and (**f**)–(**j**) semi-supervised learning. Only the semi-supervised approach is able to differentiate turning at the filter light (**i**) and turning through opposing traffic (**j**)

Table 6 Behavior classification performance for four traffic flows in the street intersection dataset

True/Est Class	MCTM				LDA				HMM			
	L	R	V	VT	L	R	V	VT	L	R	V	VT
Left	**.99**	.00	.00	.01	**.49**	.44	.00	.06	**.98**	.00	.01	.01
Right	.00	**.94**	.01	.05	.00	**1.0**	.00	.00	.00	**.92**	.08	.00
Vertical	.00	.00	**.77**	.22	.01	.17	**.82**	.00	.02	.01	**.69**	.28
Vertical-Turn	.31	.05	.20	**.43**	.01	.21	.30	**.46**	.49	.04	.32	**.15**
Average Accuracy	**.78**				**.69**				**.69**			

Table 7 Behavior classification performance for the street intersection dataset. Classes are four traffic flows dataset plus dangerous behavior—turning through oncoming traffic. MCTM-Semisupervised exploits 18 labeled examples of the dangerous behaviour for learning; the other models are unsupervised

True/Est Class	MCTM-Semisupervised					MCTM-Unsupervised				
	L	R	V	T	VT	L	R	V	T	VT
Left	**.98**	.00	.00	.00	.02	**.99**	.01	.00	.00	.00
Right	.00	**.90**	.00	.06	.04	.00	**.90**	.04	.00	.06
Vertical	.00	.00	**.80**	.20	.00	.00	.00	**.95**	.00	.05
Turn	.04	.00	.29	**.51**	.16	.06	.02	.49	**.00**	.43
Vertical+Turn	.05	.00	.00	.10	**.84**	.09	.03	.00	.00	**.88**
Avg. Acc.	**.81**					**.74**				
True/Est Class	LDA					HMM				
	L	R	V	T	VT	L	R	V	T	VT
Left	**.99**	.01	.00	.00	.00	**.65**	.00	.01	.00	.34
Right	.00	**.86**	.00	.00	.14	.00	**.89**	.10	.00	.01
Vertical	.00	.17	**.78**	.00	.04	.00	.02	**.97**	.00	.00
Turn	.19	.06	.61	**.00**	.14	.04	.02	.76	**.00**	.18
Vertical+Turn	.69	.00	.00	.00	**.31**	.02	.09	.20	.00	**.69**
Avg. Acc.	**.59**					**.64**				

Table 8 Summary of model selection results and confidence in terms of log Bayes factor

	Actions N_y	Behaviors N_z	Bayes Factor
Synthetic	4	4	3
Intersection	10	16	11

5.6 Model Complexity Control

Thus far we have used domain knowledge to fix the model complexity $M = \{N_y, N_z\}$ for ease of illustration. Our final experiment relaxes this assumption and tests the effectiveness of our model selection approach (Sect. 3.5) for automatically estimating the number of activities and behaviors in the scene. Specifically, we evaluate the predictive likelihood of the test data $p(D^{\text{te}}|D^{\text{tr}}, N_y, N_z)$ for the synthetic dataset and the street intersection dataset. For the synthetic dataset (Sect. 3.6), we evaluated a grid of 49 models, with N_y and N_z varying from 2 to 8. For the street intersection dataset (Sect. 5.1), we evaluated a grid of 100 models, with N_y and N_z varying from 2 to 20 in increments of 2. Table 8 reports the most probable model determined for each dataset. We also report the confidence in terms of the Bayes factor (Bishop 2006; Gilks et al. 1995), specifically the difference in log probability between the selected model and the next most probable model evaluated.

For the synthetic dataset, we recover the true number of actions and behaviors ($N_y = N_z = 4$, Sect. 3.6) used to generate the data, verifying the procedure's validity. The model estimated by the framework for the street intersection data is more complex than we assumed in Sect. 5.1. Notably, it prefers to break down the data into more clusters—at a finer scale—than previously assumed. That is, the model separates behaviors which were represented singly in Sect. 5.1 into multiple more specific sub-behaviors. For example, Fig. 9 illustrates the prototype flow of two of the 12 learned behaviors for the intersection dataset. Here, the overall rightward traffic phase is now broken down into a flow in which right-flowing cars go straight, and those in which they turn at the intersection. In contrast, both straight and turning traffic were previously encompassed by a single behavior in Fig. 4(a). Because the rightward turning cars were relatively sparse, it is indeed arguably a separate behavior to the only straight rightward flow—just at a slightly finer scale. There is no clearly right answer: a human asked to determine the number of behaviors may or may not break this down into two behaviors depending on their perspective and broader task context. Thus our model selection approach can indeed estimate a plausible number of actions and behaviors; but as there is unlikely to be an obviously right answer for real life data, this should therefore be taken as a suggestion in absence of other information, rather than as a specific correct answer.

Fig. 9 Example of fine grained cluster decomposition discovered: strictly horizontal vs. turning flows

5.7 Computational Cost

The computational cost of MCMC learning in any model is hard to quantify, because assessing convergence is itself an open question (Gilks et al. 1995), as also highlighted by Wang et al. (2009). 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 Wang et al. (2009). 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 3 GHz CPU, this meant that training on 5 minutes of our data required about 15 minutes. 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 (Blei et al. 2003) took about 20 and 8 seconds respectively, while (EM) HMM in Matlab took 64 seconds and 26 seconds. Wang et al. (2009) reported that Gibbs sampling in their HDP model required 8 hours to process each hour of data; and they do not propose an online testing solution. These numbers should not be compared literally given the differences in implementations; however the important thing to note is that our model is competitive in training speed to sophisticated contemporary models (Wang et al. 2009), and it is much faster for online testing. Moreover, it is faster than the simpler models which it outperforms in saliency detection and classification.

6 Conclusions

We introduced a novel dynamic Bayesian topic model for simultaneous hierarchical clustering of visual events into actions and dynamic global behaviours. The model addresses four critical tasks for video mining: unsupervised modeling scene behavioral characteristics under-pinned at different spatial and temporal levels (Sect. 5.2); online behaviour screening and saliency detection (Sect. 5.3); online behavior classification (Sect. 5.4) and semi-supervised detection of specified interesting behaviors (Sect. 5.5). Our approach addresses the initially identified challenges (Sect. 1) of behavioural complexity, robustness & sensitivity and computational tractability in various ways: The local composi-

tional structure of our model and global temporal correlation between behaviours allows complex and dynamic behaviours to be profiled, which moreover enhances sensitivity to salient clips with visually subtle deviations from usual activities. The intermediate action layer and Bayesian parameter learning help to improve robustness. Finally, our new inference algorithm (Sect. 3.4) addresses the issue of computational complexity, enabling real-time operation online.

Mining and Screening Our Gibbs learning procedure has proven effective at learning typical actions, behaviours and temporal correlations in three diverse and challenging test datasets. We showed how to use the Gibbs samples from learning for rapid Bayesian inference of clip category and saliency. Evaluating the salient clips returned from our diverse data sets, our MCTM outperforms LDA and HMMs for unsupervised mining and screening salient behaviours, especially for visually subtle, spatially and temporally extended activity. This is because we model simultaneously temporal evolution of behaviour (unlike LDA), the hierarchical composition of action into behaviours (unlike LDA and HMM) and use Bayesian parameter learning (unlike HMM) to reduce over-fitting. Compared to object-centric approaches such as Basharat et al. (2008), Saleemi et al. (2009), Berclaz et al. (2008), Sillito and Fisher (2008), Hu et al. (2006), Wang et al. (2006), Dee and Hogg (2004), Stauffer and Grimson (2000), Johnson and Hogg (1996), our simple and reliable visual features improve robustness to clutter and occlusion. An important benefit of our approach is the breadth of different kinds of behaviours that may be modeled—and the variety of different irregular salient behaviours that the model is sensitive to—due to our simultaneous hierarchical modeling and temporal correlation globally optimized in a unified model. We have demonstrated that our model can detect temporally extended events typically flagged by object-tracking centric models (Saleemi et al. 2009; Basharat et al. 2008) such as u-turns, as well as multi-object events typically only detected by statistical event models (Wang et al. 2009) such as jay-walking.

Unsupervised and Semi-supervised Classification Our model has proven successful at rapidly classifying on-going behaviors in a scene into one of various learned categories. Building on this ability, we have shown how we can use our framework as a semi-supervised learner to bootstrap the model from a few labeled examples to ensure it represents each behavior of interest (Sect. 5.5) while exploiting the unlabeled data to build a better representation. In combination these different modes of use allow our framework to be flexibly applied in a variety of ways depending on how much domain knowledge is available or affordable in a given situation. With domain knowledge including a complete or partial set of behaviors of interest, we can use our framework in a semi-supervised way to detect specific behaviors. In an intermediate case, with an idea of the complexity or number of usual behaviors in a scenario, but not particular examples of each, we can fix the model complexity and learn a fully unsupervised model to classify new behaviors and detect unusual salient behaviors. Finally, with no domain knowledge at all, we can also learn a suitable number of behaviors & actions offline (Sect. 5.6). Other approaches to introducing supervision to topic models have recently been independently exploited for regression (Blei and McAuliffe 2007) and for action recognition (Wang and Mori 2009), but with the much stronger assumption of full supervision, and without real-time inference for new data.

Input Feature Representation In this study we have used simple quantized optical flow as the input vocabulary. Other more complex features such as bags of 3D SIFT (Scovanner et al. 2007) or space-time interest point (Dollar et al. 2005) descriptors are possible. These have been exploited successfully in near-view action recognition (Niebles et al. 2008). We retain simple motion features however for two reasons. To model multi-object behavior via spatio-temporal co-occurance of events we exploit a grid of cells (Sect. 3.1) which requires that the number of possible events per cell should be minimized. In our case this was 4 motion directions; the 1000s typical for more complex features would be intractible. Secondly, in our far view surveillance data targets are often too small to convey reliable appearance information, which in any case may not be relevant to our high level behavior-based profiling task.

Parametric Versus Non-parametric Models Our procedure for complexity control is in contrast to other related work which uses non-parametric methods (e.g., Wang et al. 2009), and provides two advantages over non-parametrics: efficiency and representation. Firstly, by learning and fixing the model complexity offline once for each scenario, we have less work to do during subsequent processing, thereby enabling our real-time processing. This is in contrast to Wang et al. (2009) which does not propose an online solution, and whose batch solution is in the order of ten times slower than real time (Wang et al. 2009). Secondly, as we have discussed, there may be a specific set of known interesting behaviors that a user desires the model to represent. Our framework allows modeling of these behaviors, which fully unsupervised clustering approaches are likely to miss if they are subtle or under-represented in the data.

Limitations & Future Work There are some outstanding limitations to our approach that are worth mentioning. As for other topic modeling approaches (Wang et al. 2009; Li et al. 2008), our framework does not explicitly allow for individual actions within a clip to be irregular and others normal. This can sometimes have the unfortunate side

effect of allowing an unusual action occurring alongside numerous other common actions to go un-noticed. Another issue is that we only correlate global complex behaviors in time. It would be interesting to be able to correlate visual events and local actions in time for a better model of individual objects (Kim and Grauman 2009; Benezeth et al. 2009). We are investigating extensions to address these issues. Simple features provide robustness to noise for many kinds of behavior modeling, but there is the drawback that our framework is then not ideal for detecting some particular behaviors of potential interest such as abandoning of luggage, because that requires explicit object detection, tracking and association, e.g., Smith et al. (2006).

Our model has been trained and tested on partitions of a batch of standard datasets of approximately an hours length. Developing and testing models capable of dealing with long term (weeks and years) data with changing input statistics is an open question. Additionally, exploiting transfer learning (Pan and Yang 2010) is likely to be advantageous. That is, how to transfer invariant aspects of activities and behaviors learned from one public space to another, avoiding relearning from scratch.

While online inference is fast, our Gibbs learning procedure is slow enough to provide a barrier to learning on truly large and complex datasets. We are investigating faster variational solutions to learning (Bishop 2006; Blei et al. 2003). Finally, an interesting extension to our semi-supervised approach which we are investigating is that of active learning (Kapoor et al. 2007). This approach potentially allows the model to nominate iteratively specific data-points for labeling which will be most helpful for refining its behaviour based clustering of clips, while minimizing the labeling cost required.

Appendix: Deriving Gibbs Learning Updates

Action Variable Updates

It is possible to derive the Gibbs sampling updates for our model either by recourse to cancellation, e.g., $p(y_{i,t}|\mathbf{x}_{1:T},\mathbf{y}_{\setminus i,t},z_{1:T},) \propto \frac{p(\mathbf{x}_{1:T},\mathbf{y}_{1:T},z_{1:T})}{p(\mathbf{x}_{1:T},\mathbf{y}_{\setminus i,t},z_{1:T})}$; or by Bayes' theorem. The update for $p(y_{i,t}|\mathbf{x}_{1:T},\mathbf{y}_{\setminus i,t},z_{1:T})$ (2) using Bayes' theorem is derived as follows:

$$\begin{aligned}
&p(y_{i,t}|\mathbf{x}_{1:T},\mathbf{y}_{\setminus i,t},z_{1:T})\\
&\quad\propto p(x_{i,t}|\mathbf{x}_{\setminus i,t},\mathbf{y}_{1:T},z_{1:T})p(y_{i,t}|\mathbf{x}_{\setminus i,t}\mathbf{y}_{\setminus i,t},z_{1:T}), \qquad (12)\\
&\quad= \int p(x_{i,t}|y_{i,t},\phi)p(\phi|\mathbf{y}_{\setminus i,t},\mathbf{x}_{\setminus i,t})d\phi\\
&\qquad\cdot \int p(y_{i,t}|z_t,\theta)p(\theta|\mathbf{y}_{\setminus i,t},z_{\setminus t})d\theta, \qquad (13)\\
&\quad= \int \phi_{x_{i,t},y_{i,t}}\mathrm{Dir}\left(\phi_{y_{i,t}};n^{-}_{y_{i,t}}+\beta\right)d\phi_{y_{i,t}}\\
&\qquad\cdot \int \theta_{y_{i,t},z_t}\mathrm{Dir}\left(\theta_{z_t};n^{-}_{z_t}+\alpha\right)d\theta_{z_t}. \qquad (14)
\end{aligned}$$

Equation (14) then leads directly to the desired action variable update (2) by the standard formula for the expectation of a Dirichlet distribution.

Behavior Variable Updates

The update for $p(z_t|\mathbf{y},\mathbf{z}_{\setminus t},\mathbf{x})$ (3) is derived as:

$$p(z_t|\mathbf{y}_{1:T},z_{\setminus t},\mathbf{x}_{1:T}) \propto p(\mathbf{y}_t|\mathbf{y}_{\setminus t},z_{\setminus t})p(z_t|,\mathbf{y}_{\setminus t},z_{\setminus t}). \qquad (15)$$

We can recognize the first "likelihood" term in above as a Polya distribution, computed as:

$$\begin{aligned}
&p(\mathbf{y}_t|\mathbf{y}_{\setminus t},z_{1:T})\\
&\quad= \int \mathrm{Multi}(\mathbf{y}_t;z_t,\theta_{z_t})\mathrm{Dir}(\theta_{z_t};\mathbf{y}_{\setminus t},z_{\setminus t})d\theta_{z_t},\\
&\quad= \frac{1}{\Delta(n^{-t}_{z_t}+\alpha)}\int\prod_y \theta^{n^{t}_{y,z_t}}_{y,z_t}\prod_y \theta^{n^{-t}_{y,z_t}+\alpha}_{y,z_t}d\theta_{z_t},\\
&\quad= \frac{\Delta(n^{t}_{z_t}+n^{-t}_{z_t}+\alpha)}{\Delta(n^{-t}_{z_t}+\alpha)},\\
&\quad= \frac{\prod_y\Gamma(n_{y,z_t}+\alpha)}{\prod_y\Gamma(n^{-}_{y,z_t}+\alpha)}\frac{\Gamma(\sum_y n^{-}_{y,z_t}+N_y\alpha)}{\Gamma(\sum_y n_{y,z_t}+N_y\alpha)}, \qquad (16)
\end{aligned}$$

where we write $\Delta(\alpha)\triangleq\frac{\prod_j\Gamma(\alpha_j)}{\Gamma(\sum_j\alpha_j)}$ to indicate the normalizing constant of the Dirichlet distribution $\mathrm{Dir}(\theta;\alpha)$. $n^{t}_{y_t,z_t}$ indicates the action counts solely for the current behavior z_t and n^{-t}_{y,z_t} indicates all the action-behavior counts observed at times other than t. Because there are multiple actions per behavior, there is not a unit difference between n^{-}_{y,z_t} and n_{y,z_t} and (16) does not simplify further unlike LDA and (14) and (2).

Finally, we need the "prior" term in (15), which depends on the values of z_{t-1}, z_t and z_{t+1}.

$$\begin{aligned}
&p(z_t|\mathbf{y}_{\setminus t},z_{\setminus t})\\
&\quad\propto \int p(z_t|z_{t-1},\psi)p(z_{t+1}|z_t,\psi)p(\psi|z_{\setminus t})d\psi \qquad (17)
\end{aligned}$$

Let $z_{t-1}=j$, $z_t=k$ and $z_{t+1}=l$. Then $p(z_t|z_{t-1},\psi)=\psi_{k,j}$ and $p(z_{t+1}|z_t,\psi)=\psi_{l,k}$ and $p(z_t|\mathbf{y}_{\setminus t},z_{\setminus t})$ is determined as:

$$\begin{aligned}
j=k=l:&\ \frac{1}{\Delta(n^{-}_{j}+\gamma)}\int\psi_{k,j}\psi_{l,k}\prod_i\psi^{n^{-}_{i,j}+\gamma-1}_{i,j}d\psi_j,\\
&= \frac{(n^{-}_{k,j}+1+\gamma)(n^{-}_{l,k}+\gamma)}{(n^{-}_{\cdot,j}+1+N_z\gamma)(n^{-}_{\cdot,k}+N_z\gamma)}, \qquad (18)\\
j\neq k\neq l:&\ \frac{1}{\Delta(n^{-}_{j}+\gamma)}\int\psi_{k,j}\prod_i\psi^{n^{-}_{i,j}+\gamma-1}_{i,j}d\psi_j
\end{aligned}$$

$$\cdot \frac{1}{\Delta(n_k^- + \gamma)} \int \psi_{l,k} \prod_i \psi_{i,k}^{n_{i,k}^- + \gamma - 1} d\psi_k,$$

$$= \frac{(n_{k,j}^- + \gamma)}{(n_{\cdot,j}^- + N_z\gamma)} \frac{(n_{l,k}^- + \gamma)}{(n_{\cdot,k}^- + N_z\gamma)}, \tag{19}$$

$$j = k \neq l : \frac{1}{\Delta(n_j^- + \gamma)} \int \psi_{k,j}\psi_{l,k} \prod_i \psi_{i,j}^{n_{i,j}^- + \gamma - 1} d\psi_j,$$

$$= \frac{(n_{k,j}^- + \gamma)}{(n_{\cdot,j}^- + N_z\gamma)} \frac{(n_{l,k}^- + \gamma)}{(n_{\cdot,k}^- + 1 + N_z\gamma)}, \tag{20}$$

where we exploit the identities $\Gamma(z+1)/\Gamma(z) = z$ and $\Gamma(z+2)/\Gamma(z) = (z+1)z$ to simplify the ratios of normalizing constants. Finally, substituting (16) and (18)–(20) into (15), we obtain the desired (3).

References

Ali, S., & Shah, M. (2008). Floor fields for tracking in high density crowd scenes. In *European conference on computer vision*.

Basharat, A., Gritai, A., & Shah, M. (2008). Learning object motion patterns for anomaly detection and improved object detection. In *IEEE conference on computer vision and pattern recognition*.

Benezeth, Y., Jodoin, P.-M., Saligrama, V., & Rosenberger, C. (2009). Abnormal events detection based on spatio-temporal co-occurences. In *IEEE conference on computer vision and pattern recognition*.

Berclaz, J., Fleuret, F., & Fua, P. (2008). Multi-camera tracking and atypical motion detection with behavioral maps. In *European conference on computer vision*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Blei, D., & Lafferty, J. (2006). Dynamic topic models. In *International conference on machine learning*.

Blei, D., & McAuliffe, J. (2007). Supervised topic models. In *Neural information processing systems*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boiman, O., & Irani, M. (2007). Detecting irregularities in images and in video. *International Journal of Computer Vision*, *74*(1), 17–31.

Chang, S. F., Luo, J., Maybank, S., Schonfeld, D., & Xu, D. (2008). An introduction to the special issue on event analysis in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(11), 1469–1472.

Chen, M. y., Li, H., & Hauptmann, A. (2009). Informedia @ trecvid 2009: analyzing video motions. In *Proc TRECvid*.

Dee, H., & Hogg, D. (2004). Detecting inexplicable behaviour. In *British machine vision conference*.

Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *VS-PETS* (pp. 65–72).

Duong, T., Bui, H., Phung, D., & Venkatesh, S. (2005). Activity recognition and abnormality detection with the switching hidden semi-Markov model. In *IEEE conference on computer vision and pattern recognition*.

Gilks, W., Richardson, S., & Spiegelhalter, D. (Eds.) (1995). *Markov chain Monte Carlo in practice*. London/Boca Raton: Chapman & Hall/CRC Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5228–5235.

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2007). Integrating topics and syntax. In *Neural information processing systems*.

Gruber, A., Rosen-Zvi, M., & Weiss, Y. (2007). Hidden topic Markov models. In *Artificial intelligence and statistics*.

HOSDB. Imagery library for intelligent detection systems (i-lids). In *IEEE conf. on crime and security* (2006).

Hospedales, T., Gong, S., & Xiang, T. (2009). A Markov clustering topic model for behaviour mining in video. In *IEEE international conference on computer vision*.

Hu, W., Tan, T., Wang, L., & Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, *34*(3), 334–352.

Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., & Maybank, S. (2006). A system for learning statistical motion patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(9), 1450–1464.

Hu, Z., Ye, G., Jia, G., Chen, X., Hu, Q., Jiang, K., Wang, Y., Qing, L., Tian, Y., Wu, X., & Gaoa, W. (2009). Pku@trecvid2009: Single-actor and pair-activity event detection in surveillance video. In *Proc. TRECvid*.

Inoue, N., Hao, S., Saito, T., & Shinoda, K. (2009). Titgt at trecvid 2009 workshop. In *Proc. TRECvid*.

Johnson, N., & Hogg, D. (1996). Learning the distribution of object trajectories for event recognition. *Image and Vision Computing*, *8*, 609–615.

Kapoor, A., Horvitz, E., & Basu, S. (2007). Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *International joint conference on artificial intelligence*.

Kim, J., & Grauman, K. (2009). Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental update. In *IEEE conference on computer vision and pattern recognition*.

Li, J., Gong, S., & Xiang, T. (2008). Global behaviour inference using probabilistic latent semantic analysis. In *British machine vision conference*.

Meng, J., & Chang, S.-F. (1996). Tools for compressed-domain video indexing and editing. In *SPIE conference on storage and retrieval for image and video databases*.

National institute of standards and technology (NIST): Trec video retrieval evaluation. http://trecvid.nist.gov/.

Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, *79*(3), 299–318.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Pritch, Y., Rav-Acha, A., & Peleg, S. (2008). Nonchronological video synopsis and indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(11), 1971–1984.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Uncertainty in artificial intelligence*.

Saleemi, I., Shafique, K., & Shah, M. (2009). Probabilistic modeling of scene dynamics for applications in visual surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(8), 1472–1485.

Scovanner, P., Ali, S., & Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *ACM international conference on multimedia*.

Sillito, R. R., & Fisher, R. B. (2008). Semi-supervised learning for anomalous trajectory detection. In *British machine vision conference*.

Smith, K., Quelhas, P., & Gatica-Perez, D. (2006). Detecting abandoned luggage items in a public space. In *Performance evaluation of tracking and surveillance (PETS) workshop*.

Stauffer, C., & Grimson, W. (2000). Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 747–757.

Wallach, H. (2006). Topic modeling: beyond bag-of-words. In *International conference on machine learning*.

Wallach, H., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *International conference on machine learning*.

Wang, Y., & Mori, G. (2009). Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(10), 1762–1774.

Wang, X., Tieu, K., & Grimson, E. (2006). Learning semantic scene models by trajectory analysis. In *European conference on computer vision*.

Wang, X., Ma, X., & Grimson, E. (2009). Unsupervised activity perception by hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(3), 539–555.

Xiang, T., & Gong, S. (2006). Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, *61*(1), 21–51.

Xiang, T., & Gong, S. (2008a). Activity based surveillance video content modelling. *Pattern Recognition*, *41*, 2309–2326.

Xiang, T., & Gong, S. (2008b). Video behavior profiling for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(5), 893–908.

Xie, L., Sundaram, H., & Campbell, M. (2008). Event mining in multimedia streams. *Proceedings of the IEEE*, *96*(4), 623–647.

Zhong, H., Shi, J., & Visontai, M. (2004). Detecting unusual activity in video. In *IEEE conference on computer vision and pattern recognition* (pp. 819–826).