Identifying Surprising Events in Videos Using Bayesian Topic Models*

Avishai Hendel, Daphna Weinshall, Shmuel Peleg

School of Computer Science, The Hebrew University, Jerusalem, Israel

Abstract. Automatic processing of video data is essential in order to allow efficient access to large amounts of video content, a crucial point in such applications as video mining and surveillance. In this paper we focus on the problem of identifying interesting parts of the video. Specifically, we seek to identify atypical video events, which are the events a human user is usually looking for. To this end we employ the notion of Bayesian surprise, as defined in [1, 2], in which an event is considered surprising if its occurrence leads to a large change in the probability of the world model. We propose to compute this abstract measure of surprise by first modeling a corpus of video events using the Latent Dirichlet Allocation model. Subsequently, we measure the change in the Dirichlet prior of the LDA model as a result of each video event's occurrence. This change of the Dirichlet prior leads to a closed form expression for an event's level of surprise, which can then be inferred directly from the observed data. We tested our algorithm on a real dataset of video data, taken by a camera observing an urban street intersection. The results demonstrate our ability to detect atypical events, such as a car making a U-turn or a person crossing an intersection diagonally.

1 Introduction

1.1 Motivation

The availability and ubiquity of video from security and monitoring cameras has increased the need for automatic analysis and classification. One urging problem is that the sheer volume of data renders it impossible for human viewers, the ultimate classifiers, to watch and understand all of the displayed content. Consider for example a security officer who may need to browse through the hundreds of cameras positioned in an airport, looking for possible suspicious activities - a laborious task that is error prone, yet may be life critical. In this paper we address the problem of unsupervised video analysis, having applications in various domains, such as the inspection of surveillance videos, examination of 3D medical images, or cataloging and indexing of video libraries.

A common approach to video analysis serves to assist human viewers by making video more accessible to sensible inspection. In this approach the human judgment is maintained, and video analysis is used only to assist viewing.

* In Proceedings of 10th Asian Conf. on Computer Vision (ACCV), Queenstown New Zealand, November 2010

Algorithms have been devised to create a compact version of the video, where only certain activities are displayed [3], or where all activities are displayed using video summarization [4].

We would like to go beyond summarization; starting from raw video input, we seek an automated process that will identify the unusual events in the video, and reduce the load on the human viewer. This process must first extract and analyze activities in the video, followed by establishing a model that characterizes these activities in a manner that permits meaningful inference. A measure to quantify the significance of each activity is needed as a last step.

1.2 Related Work

Boiman and Irani [3] propose to recognize irregular activities in video by considering the complexity required to represent the activity as a composition of codebook video patches. This entails dense sampling of the video and is therefore very time consuming, making it cumbersome to apply this algorithm to real world data. Itti and Baldi [1] present a method for surprise detection that operates in low-level vision, simulating early vision receptors. Their work is directed at the modeling and prediction of human visual attention, and does not address the understanding of high level events.

Other researchers use Bayesian topic models as a basis for the representation of the environment and for the application of inference algorithms. To detect landmark locations Ranganathan and Dellaert [5] employ the surprise measure over an appearance place representation. Their use of only local shape features makes their approach applicable in the field of topological mappings, but not in object and behavior based video analysis. Two closely related models are that of Hospedales et. al. [6] and Wang et. al. [7]. Both models use topic models over low level features to represent the environment. [6] uses Bayesian saliency to recognize irregular patterns in video scenes, while [7] defines abnormal events as events with low likelihood. Both approaches may be prone to the 'white snow paradox' [1], where data that is more informative in the classic Shannon interpretation does not necessarily match human semantic interests.

1.3 Our Approach

We present a generative probabilistic model that accomplishes the tasks outlined above in an unsupervised manner, and test it in a real world setting of a webcam viewing an intersection of city streets.

The preprocessing stage consists of the extraction of video activities of high level objects (such as vehicles and pedestrians) from the long video streams given as input. Specifically, we identify a set of video events (video tubes) in each video sequence, and represent each event with a 'bag of words' model. In previous work words were usually chosen to be local appearance features, such as SIFT [8, 9] or spatio-temporal words [10]. We introduce the concept of 'transition words', which allows for a compact, discrete representation of the dynamics of an object in a video sequence. Despite its simplicity, this representation is

successful in capturing the essence of the input paths. The detected activities are then represented using a latent topic model, a paradigm that has already shown promising results [11, 12, 9, 6].

Next, we examine the video events in a rigorous Bayesian framework, to identify the most interesting events present in the input video. Thus, in order to differentiate intriguing events from the typical commonplace events, we measure the effect of each event on the observer's beliefs about the world, following the approach put forth in [1, 2]. We propose to measure this effect by comparing the prior and posterior parameters of the latent topic model, which is used to represent the overall data. We then show that in the street camera scenario, our model is able to pick out atypical activities, such as vehicle U-turns or people walking in prohibited areas.

The rest of the paper is organized as follows: in Section 2 we describe the basic extraction and representation of activities in input videos. In Section 3 the 'bag of words' model is used to represent the input in a statistical generative manner as explained above. Section 4 and Section 5 introduce the Bayesian framework for identifying atypical events, and in Section 6 the application of this framework to real world data is presented.

2 Activity Representation

2.1 Objects as Space Time Tubes

To recognize unusual activities in input videos, we first need to isolate and localize objects out of the image sequence. The fundamental representation of objects in our model is that of 'video tubes' [13]. A tube is defined by a sequence of object masks carved through the space time volume, assumed to contain a single object of interest (e.g., in the context of street cameras, it may be a vehicle or a pedestrian). This localizes events in both space and time, and enables the association of local visual features with a specific object, rather than an entire video.

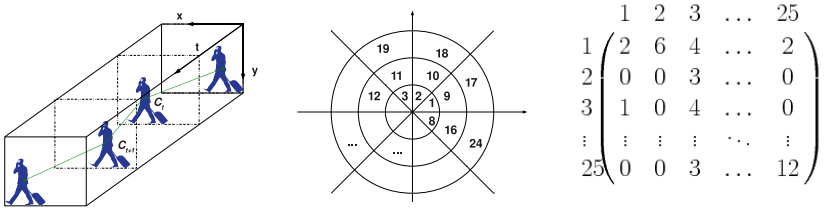
Tubes are extracted by first segmenting each video frame into background and foreground regions, using a modification of the 'Background Cut' method, described in [14]. Foreground blobs from consecutive frames are then matched by spatial proximity to create video tubes that extend through time. A collection of tubes extracted from an input video sequence is the corpus used as the basis for later learning stages.

2.2 Trajectories

An obvious and important characteristic of a video tube is its trajectory, as defined by the sequence of its spatial centroids. Encoding the dynamics of an object is a crucial step for successful subsequent processing. A preferable encoding in our setting should capture the characteristic of the tube's path in a compact and effective way, while considering location, speed and form.

Of the numerous existing approaches, we use a modification of the method suggested in [15]. Denote the displacement vector between two consecutive spatial centroids C_t and C_{t+1} as $D = \overrightarrow{C_t C_{t+1}}$ (Fig. 1a). Since the temporal difference is constant (a single frame interval between centroids) we may ignore it, and assume D has only spatial components $(\Delta x, \Delta y)$. Quantization of possible values of D is obtained through the following procedure: First, the magnitude of all displacement vectors is normalized by the largest displacement found in the trajectory - $\|D\|_{max}$. Then the normalized magnitude is assigned to one of three uniform quantization levels. The orientation component of each displacement vector is binned into one of eight sectors of the unit circle, each sector covering $\pi/4$ radians. The combination of three magnitude scales and eight orientation sectors gives 24 quantization bins (Fig. 1b). Adding another bin to indicate zero displacement, we have a total of 25 displacement bins. After quantizing all of the displacement vectors of a trajectory, we create a transition occurrence matrix (Fig. 1c), indicating the frequency of bin transitions in the tube.

This matrix can be viewed as a histogram of 'transition words', where each word describes the transition between two consecutive quantized displacement vectors. The final representation of a trajectory is this histogram, indicating the relative frequency of the 625 possible transitions.



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ 25 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & \dots & 25 \end{array} \\ \begin{pmatrix} 2 & 6 & 4 & \dots & 2 \\ 0 & 0 & 3 & \dots & 0 \\ 1 & 0 & 4 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 3 & \dots & 12 \end{pmatrix} \end{array}$$

(a) (b) (c)

Fig. 1: Trajectory representation: the three stages of our trajectory representation: (a) compute the displacement of the centroids of the tracked object between frames, (b) quantize each displacement vector into one of 25 quantization bins, and (c) count the number of different quantization bin transitions in the trajectory into a histogram of bin transitions.

3 Modeling of Typical Activities Using LDA

The *Latent Dirichlet Allocation* (LDA) model is a generative probabilistic model, first introduced in the domain of text analysis and classification [16]. As other topic models, it aims to discover latent topics whose mixture is assumed to be the underlying cause of the observed data. Its merits lie in that it is a truly generative

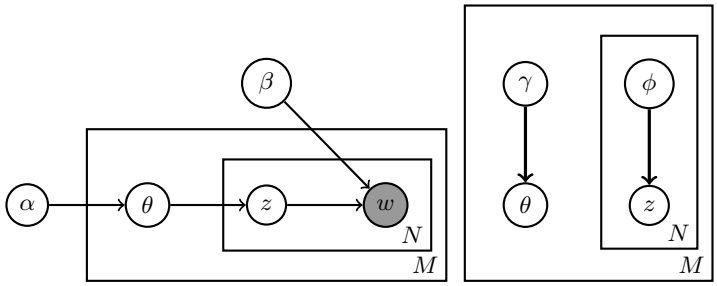


Fig. 2: (a) Graphical model represntation of LDA using plate notation. (b) Simplified model used to approximate the posterior distribution.

model that can be learned in a completely unsupervised manner, it allows the use of priors in a rigorous Bayesian manner, and it does not suffer from over-fitting issues like its closely related pLSA model [17]. It has been successfully applied recently to computer vision tasks, where the text topics have been substituted with scenery topics [9] or human action topics [12].

As is common with models from the 'bag of words' paradigm, the entities in question (video tubes, in our case) are represented as a collection of local, discrete features. The specific mixture of topics of a single video tube determines the observed distribution of these features.

More formally, assume we have gathered a set of video tubes and their trajectories in the corpus $T = \{T_1, T_2, ..., T_m\}$. Each tube is represented as a histogram of transition words taken from the trajectory vocabulary $V = \{w_{1-1}, w_{1-2}, ..., w_{25-24}, w_{25-25}\}, |V| = 625$. Thus the process that generates each trajectory T_j in the corpus is:

1. Choose $N \sim Poisson(\xi)$, the number of feature words (or, in effect, the length of the trajectory).
2. Choose $\theta \sim Dirichlet(\alpha)$, the mixture of latent topics in this tube.
3. For each of the N words w_n, where $1 \leq n \leq N$:
   - Choose a topic $z_n \sim Multinomial(\theta)$
   - Choose a codebook word w_n from the multinomial distribution $p(w_n \mid z_n, \beta)$

In this model, α is a k-dimensional vector that is the parameter for the Dirichlet distribution, k is the predetermined number of hidden topics, and β is a $k \times V$ matrix that characterizes the word distributions conditioned on the selected latent topic. The entry $\beta_{i,j}$ corresponds to the measure $p(w^j = 1 \mid z^i = 1)$. A plate notation representation of the model is shown in Fig. 2a. The joint distribution of the trajectory topic mixture θ, the set of transition words $\mathbf{w}$ and their corresponding topics $\mathbf{z}$ can be summarized as:

$$p(\theta, \mathbf{w}, \mathbf{z} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{1}$$

Once the model has been learned and the values of the vector α and the matrix β are known, we can compute the posterior distribution of the hidden variables of a new unseen tube:

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{w}, \mathbf{z} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)} \tag{2}$$

Although this distribution is computationally intractable, approximate inference algorithms such as Gibbs sampling or variational methods can be used. The basic principle behind the variational approach [18] is to consider a simplified graphical model, where problematic ties between variables are removed. The edges between θ, $\mathbf{z}$, and $\mathbf{w}$ cause the coupling between θ and β, which is the reason for the intractibility of Eq. (2). Dropping these edges and incorporating the free variational parameters γ and ϕ into the simplified model (Fig. 2b), we acquire a family of distributions on the latent variables that is tractable:

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n) \tag{3}$$

where γ approximates the Dirichlet parameter α and ϕ mirrors the multinomial parameters β.

Now an optimization problem can be set up to minimize the difference between the resulting variational distribution and the true (intractable) posterior, yielding the optimizing parameters (γ^*, ϕ^*), which are a function of $\mathbf{w}$. The Dirichlet parameter $\gamma^*(\mathbf{w})$ is the representation of the new trajectory in the simplex spanned by the latent topics. Thus it characterizes the composition of the actual path out of the k basic trajectory topics.

Based on this inference method, Blei [16] suggests an alternating variational EM procedure to estimate the parameters of the LDA model:

1. E-Step: For each tube, find the optimizing values of the variational parameters $\{\gamma_t^*, \phi_t^* : t \in T.\}$
2. M-Step: Maximize the resulting lower bound on the log likelihood of the entire corpus with respect to the model parameters α and β.

The estimation of the model's parameters α and β completes our observer's model of its world. The Dirichlet prior α describes the common topic mixtures that are to be expected in video sequences taken from the same source as the training corpus. A specific mixture θ_t determines the existence of transitions found in the trajectory using the per-topic word distribution matrix β. Crude classification of tubes into one of the learned latent topics can be done simply by choosing the topic that corresponds to the maximal element in the posterior Dirichlet parameter γ_t^*.

4 Surprise Detection

The notion of surprise is, of course, human-centric and not well defined. Surprising events are recognized as such with regard to the domain in question, and background assumptions that can not always be made explicit. Thus, rule based methods that require manual tuning may succeed in a specific setting, but are doomed to failure in less restricted settings. Statistical methods, on the other hand, require no supervision. Instead, they attempt to identify the expected events from the data itself, and use this automatically learned notion of typicality to recognize the extraordinary events.

Such framework is proposed in the work by Itti [1] and Schmidhuber [2]. Dubbed 'Bayesian Surprise', the main conjecture is that a surprising event from the viewpoint of an observer is an event that modifies its current set of beliefs about the environment in a significant manner. Formally, assume an observer has a model M to represent its world. The observer's belief in this model is described by the prior probability of the model $p(M)$ with regard to the entire model space $\mathcal{M}$. Upon observing a new measurement t, the observer's model changes according to Bayes' Law:

$$p(M \mid t) = \frac{p(M)p(t \mid M)}{p(t)} \tag{4}$$

This change in the observer's belief in its current model of the world is defined as the surprise experienced by the observer. Measurements that induce no or minute changes are not surprising, and may be regarded as 'boring' or 'obvious' from the observer's point of view. To quantify this change, we may use the KL divergence between the prior and posterior distributions over the set $\mathcal{M}$ of all models:

$$S(t, M) = KL(p(M), p(M \mid t)) = \int_{\mathcal{M}} p(M) log \frac{p(M)}{p(M \mid t)} dM \tag{5}$$

This definition is intuitive in that surprising events that occur repeatedly will cease to be surprising, as the model is evolving. The average taken over the model space also ensures that events with very low probability will be regarded as surprising only if they induce a meaningful change in the observer's beliefs, thus ignoring noisy incoherent data that may be introduced.

Although the integral in Eq. (5) is over the entire model space, turning this space to a parameter space by assuming a specific family of distributions may allow us to compute the surprise measure analytically. Such is the case with the Dirichlet family of distributions, which has several well known computational advantages: it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution.

5 Bayesian Surprise and the LDA Model

As noted above, the LDA model is ultimately represented by its Dirichlet prior α over topic mixtures. It is a natural extension now to apply the Bayesian surprise framework to domains that are captured by LDA models.

Recall that video tubes in our 'bag of words' model are represented by the posterior optimizing parameter γ^*. Furthermore, new evidence also elicits a new Dirichlet parameter for the world model of the observer, $\widehat{\alpha}$. To obtain $\widehat{\alpha}$, we can simulate one iteration of the variational EM procedure used above in the model's parameters estimation stage, where the word distribution matrix β is kept fixed. This is the Dirichlet prior that would have been calculated had the new tube been appended to the training corpus. The Bayesian Surprise formula when applied to the LDA model can be now written as:

$$S(\alpha, \widehat{\alpha}) = KL_{DIR}(\alpha, \widehat{\alpha}) \tag{6}$$

The Kullback - Leibler divergence of two Dirichlet distributions can be computed as [19]:

$$KL_{DIR}(\alpha, \widehat{\alpha}) = log\frac{\Gamma(\boldsymbol{\alpha})}{\Gamma(\widehat{\boldsymbol{\alpha}})} + \sum_{i=1}^{k} log\frac{\Gamma(\widehat{\alpha}_i)}{\Gamma(\alpha_i)} + \sum_{i=1}^{k}[\alpha_i - \widehat{\alpha}_i][\psi(\alpha_i) - \psi(\boldsymbol{\alpha})] \tag{7}$$

where

$$\boldsymbol{\alpha} = \sum_{i=1}^{k}\alpha_i \quad \text{and} \quad \widehat{\boldsymbol{\alpha}} = \sum_{i=1}^{k}\widehat{\alpha}_i$$

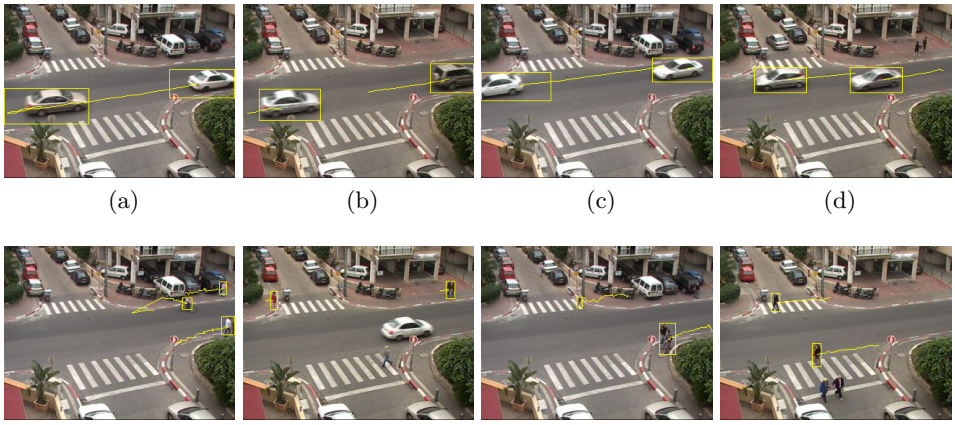
and Γ and ψ are the gamma and digamma functions, respectively.

Thus each video event is assigned a surprise score, which reflects the tube's deviation from the expected topic mixture. In our setting, this deviation may correspond to an unusual trajectory taken by an object, such as 'car doing a U-turn', or 'person running across the road'. To obtain the most surprising events out of a corpus, we can select those tubes that receive a surprise score that is higher than some threshold.

6 Experimental Results

6.1 Dataset

Surveillance videos are a natural choice to test and apply surprise detection algorithms. Millions of cameras stream endless videos that are notoriously hard to monitor, where significant events can be easily overlooked by an overwhelmed human observer. We test our model on data obtained from a real world street camera, overlooking an urban road intersection. This scenario usually exhibits structured events, where pedestrians and vehicles travel and interact in mostly predefined ways, constrained by the road and sidewalk layout. Aside from security measures, intersection monitoring has been investigated and shown to help



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

(e) (f) (g) (h)

Fig. 3: Trajectory classifications: (a,b) cars going left to right, (c,d) cars going right to left, (e,f) people walking left to right, and (g,h) people walking right to left.

in reducing pedestrian and vehicle conflicts, which may result in injuries and crashes [20].

The training input sequence consists of an hour of video footage, where frame resolution is 320x240 and the frame rate is 10fps. The test video was taken in the subsequent hour. The video was taken during the morning, when the number of events is relatively small. Still, each hour contributed about 1000 video tubes. The same intersection at rush hours poses a significant challenge to the tracking algorithm due to multiple simultaneous activities and occlusions, but this tracking is not the focus of this work. Subsequent analysis is agnostic to the mode of tube extraction, and the method we used can be easily replaced by any other method.

6.2 Trajectory Classification

The first step in our algorithm is the construction of a model that recognizes typical trajectories in the input video. We fix k, the number of latent topics to be 8. Fig. 3 shows several examples of classified objects from four of the eight model topics, including examples from both the training and test corpora. Fig. 4 shows the distribution of trajectories into topics, in the train and test corpora.

Note that some of the topics seem to have a semantic meaning. Thus, on the basis of trajectory description alone, our model was able to automatically catalog the video tubes into semantic movement categories such as 'left to right', or 'top to bottom', with further distinction between smooth constant motion (normally cars) and the more erratic path typically exhibited by people. It should be noted, however, that not all latent topics correspond with easily interpretable patterns

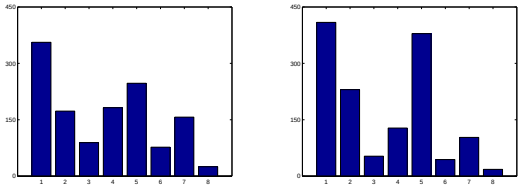
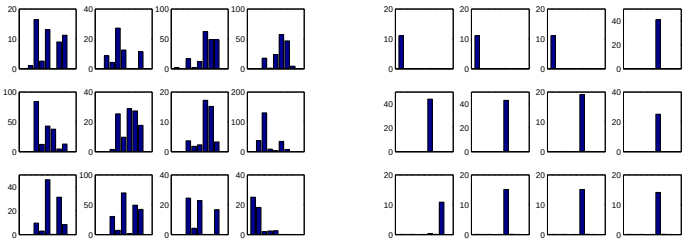


Fig. 4: Number of trajectories assigned to each topic in the train (left) and test (right) corpora. 1306 tubes were extracted from the training sequence, and 1364 from the test sequence.

of motion as depicted in Fig. 3. Other topics seem to capture complicated path forms, where pauses and direction changes occur, with one topic representing 'standing in place' trajectories.

6.3 Surprising Events

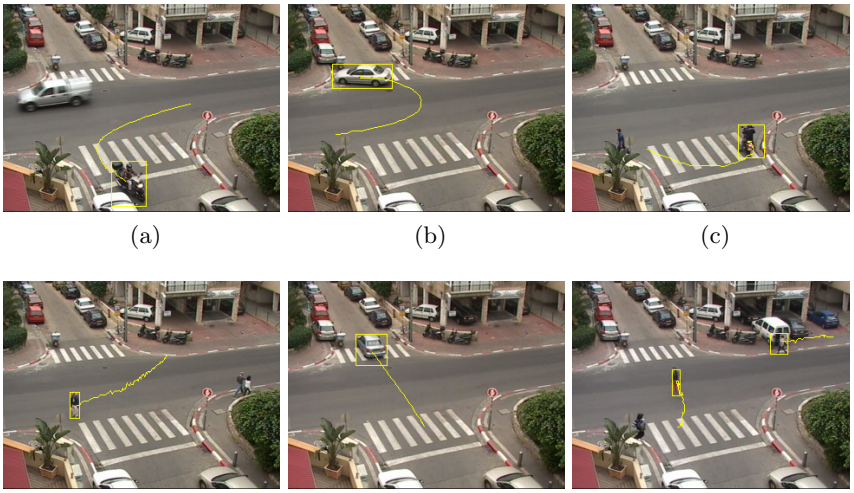


(a) (b)

Fig. 5: Posterior Dirichlet parameters γ^* values for the most surprising (a) and typical (b) events. Each plot shows the values of each of the $k = 8$ latent topics. Note that the different y scales correspond to different trajectory lengths (measured in frames).

To identify the atypical events in the corpus, we look at those tubes which have the highest surprise score. Several example tubes which fall above the 95th percentile are shown in Fig. 6. They include such activities as a vehicle performing a U-turn, or a person walking in a path that is rare in the training corpus, like crossing the intersection diagonally.

In Fig. 5 the γ^* values of the most surprising and typical trajectories are shown. It may be noted that while 'boring' events generally fall into one of



(a) (b) (c)

(d) (e) (f)

Fig. 6: Surprising events: (a) a bike turning into a one-way street from the wrong way, (b) a car performing a U-turn, (c) a bike turning and stalling over pedestrian crossing, (d) a man walking across the road, (e) a car crossing the road from bottom to top, (f) a woman moving from the sidewalk to the middle of the intersection.

the learned latent topics exclusively (Fig. 5b), the topic mixture of surprising events has massive counts in several topics at once (Fig. 5a). This observation is verified by computing the mean entropy measure of the γ^* parameters, after being normalized to a valid probability distribution:

$$\overline{H}(\gamma_{surprising}) = 1.2334, \quad \overline{H}(\gamma_{typical}) = 0.5630$$

7 Conclusions

In this work we presented a novel integration between the generative probabilistic model LDA and the Bayesian surprise framework. We applied this model to real world data of urban scenery, where vehicles and people interact in natural ways. Our model succeeded in automatically obtaining a concept of the normal behaviors expected in the tested environment, and in applying these concepts in a Bayesian manner to recognize those events that are out of the ordinary. Although the features used are fairly simple (the trajectory taken by the object), complex surprising events such as a car stalling in its lane, or backing out of its parking space were correctly identified, judged against the normal paths present in the input.

References

1. Itti, L., Baldi, P.: A principled approach to detecting surprising events in video. In: CVPR (1). (2005) 631–637
2. Schmidhuber, J.: Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In: ABiALS. (2008) 48–76
3. Boiman, O., Irani, M.: Detecting irregularities in images and in video. International Journal of Computer Vision **74** (2007) 17–31
4. Pritch, Y., Rav-Acha, A., Peleg, S.: Nonchronological video synopsis and indexing. IEEE Trans. Pattern Anal. Mach. Intell. **30** (2008) 1971–1984
5. Ranganathan, A., Dellaert, F.: Bayesian surprise and landmark detection. In: ICRA. (2009) 2017–2023
6. Hospedales, T., Gong, S., Xiang, T.: A markov clustering topic model for mining behaviour in video. In: ICCV. (2009)
7. Wang, X., Ma, X., Grimson, E.: Unsupervised activity perception by hierarchical bayesian models. In: CVPR. (2007)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
9. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR (2). (2005) 524–531
10. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: SCVMA. (2004) 91–103
11. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their localization in images. In: ICCV. (2005) 370–377
12. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79** (2008) 299–318
13. Pritch, Y., Ratovitch, S., Hendel, A., Peleg, S.: Clustered synopsis of surveillance video. In: AVSS. (2009) 195–200
14. Sun, J., Zhang, W., Tang, X., Shum, H.Y.: Background cut. In: ECCV (2). (2006) 628–641
15. Sun, J., Wu, X., Yan, S., Cheong, L.F., Chua, T.S., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: CVPR. (2009) 2004–2011
16. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. In: NIPS. (2001) 601–608
17. Hofmann, T.: Probabilistic latent semantic analysis. In: UAI. (1999) 289–296
18. Jordan, M.I., Ghahramani, Z., Jaakkola, T., Saul, L.K.: An introduction to variational methods for graphical models. Machine Learning **37** (1999) 183–233
19. Penny, W.D.: Kullback-liebler divergences of normal, gamma, dirichlet and wishart densities. Technical report, Wellcome Department of Cognitive Neurology (2001)
20. Hughes, R., Huang, H., Zegeer, C., Cynecki, M.: Evaluation of automated pedestrian detection at signalized intersections (2001)