

Identifying Rare and Subtle Behaviors: A Weakly Supervised Joint Topic Model

Timothy M. Hospedales, Jian Li, Shaogang Gong, and Tao Xiang

Abstract—One of the most interesting and desired capabilities for automated video behavior analysis is the identification of rarely occurring and subtle behaviors. This is of practical value because dangerous or illegal activities often have few or possibly only one prior example to learn from and are often subtle. Rare and subtle behavior learning is challenging for two reasons: 1) Contemporary modeling approaches require more data and supervision than may be available and 2) the most interesting and potentially critical rare behaviors are often visually subtle—occurring among more obvious typical behaviors or being defined by only small spatio-temporal deviations from typical behaviors. In this paper, we introduce a novel weakly supervised joint topic model which addresses these issues. Specifically, we introduce a multiclass topic model with partially shared latent structure and associated learning and inference algorithms. These contributions will permit modeling of behaviors from as few as one example, even without localization by the user and when occurring in clutter, and subsequent classification and localization of such behaviors online and in real time. We extensively validate our approach on two standard public-space data sets, where it clearly outperforms a batch of contemporary alternatives.

Index Terms—Probabilistic model, behavior analysis, imbalanced learning, weakly supervised learning, classification, visual surveillance, topic model, Gibbs sampling.

1 INTRODUCTION

THE general objective of computer vision-based analysis of behavior in busy public spaces has been much studied in the last decade, both because of the tremendous associated research challenges and strong application demand for algorithms which can work on real-world data. One important problem without a good existing solution is that of learning to detect and classify behaviors of semantic interest in busy public spaces which may be both *rare* and *subtle*. The relevance of this problem is clear as, for most surveillance scenarios, the behaviors of the greatest semantic interest for detection are often rare (for example, civil disobedience, shoplifting, driving offenses) and (possibly intentionally) visually subtle compared to more obvious ongoing behavior in a busy public space. These are also the reasons why this problem is challenging and unsolved: Rare behaviors by definition have few examples to learn from and, moreover, the most interesting rare behaviors are visually subtle and hard to identify. Consider, for example, the scene in Fig. 1; the (rare) traffic violations illustrated here are simple, but hard to pick out among the numerous ongoing typical behaviors. This also highlights the need for effective classification, as different rare behaviors may indicate situations of different severity (e.g., a turn violation versus a collision) requiring different responses.

Our approach is motivated by the modes of failure of existing methods in meeting the identified challenge of rare

behavior classification in busy spaces. Supervised learning methods can potentially learn to classify behaviors [1], [2], but perform poorly in our case where the target class has few examples. Moreover, the manual effort required to label training data by localizing rare behaviors in space and time may be prohibitively costly. For this reason, much recent work has focused on unsupervised density estimation methods [3], [4], [5], [6] which learn generative models of normal behavior and can thereby potentially detect rare behaviors as outliers. However, there are serious limitations: 1) Their performance is suboptimal due to not exploiting the few positive examples that may be available; 2) as outlier detectors they are not able to *categorize* different types of behavior; and 3) they fail dramatically in cases where the target behavior is nonseparable in feature space. That is, if observation or preprocessing limitations mean that a rare behavior is indistinguishable in the chosen feature space from some typical behavior, then outlier detectors will not be able to detect it without a prohibitive cost in false positives.

In this study, we first consider learning behavior models from rare and subtle examples in busy scenes. By rare behavior, we mean as few as one example, i.e., one-shot learning. By subtle we mean little visual evidence: There may be few video pixels associated with the behavior and/or few pixels differentiating a rare behavior from a typical one. We moreover eliminate the prohibitive labeling cost of traditional supervised methods by performing this task in a weakly supervised context—in which the user need not explicitly locate the target behaviors in training video. Second, we consider classification and localization of learned behaviors in test video. To address these problems we introduce a new weakly supervised joint topic model (WS-JTM) and associated learning algorithm which jointly learns a model for all the classes using a partially shared common basis. The intuition behind this approach is that well-learned common behaviors implicitly highlight the

- The authors are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK. E-mail: {tmh, jianli, sgg, txiang}@eecs.qmul.ac.uk.

Manuscript received 15 Oct. 2010; revised 24 Feb. 2011; accepted 3 Mar. 2011; published online 19 Apr. 2011.

Recommended for acceptance by G. Mori.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-10-0791.

Digital Object Identifier no. 10.1109/TPAMI.2011.81.

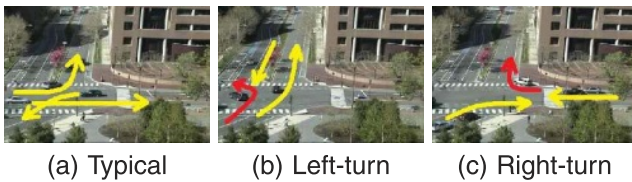


Fig. 1. Surveillance video usually contains numerous examples of (a) typical behaviors and sparse examples of rare behaviors (b) and (c). Rare behaviors (red) also usually co-occur with other typical behaviors (yellow).

few features relevant to the rare and subtle behaviors of interest, thereby permitting them to be learned without explicit supervision. Moreover, the shared common basis helps to alleviate the statistical insufficiency problem in modeling rare behaviors. Importantly, we also introduce a fast inference algorithm for WS-JTM. These innovations allow, for the first time, learning behaviors which are both rare, subtle, and not explicitly localized in the training data, and subsequent real-time classification and localization of rare behaviors in test video.

Terminology. Before continuing, we summarize our terminology as some terms are used in multiple ways by the related literature. *Visual words* refer to extracted pixel-level features used as input to the model. A *behavior* is of semantic significance to a human and may be represented by one or more *activities* or *topics* in the model, each of which corresponds to a learned set of visual words. *Clips* or *documents* refer to short segments of video. Finally, *class* is a clip-level attribute which indicates whether the clip includes a particular *behavior*.

2 RELATED WORK

Computer vision-based behavior analysis approaches vary along three broad axes: input representation, behavior model, and learning approach. Input representations are typically either object-centric—in the form of tracks [7], [8], [9]—or pixel-centric—in the form of low-level pixel [10], texture [11], or optical flow [12], [13], [3], [14], [15] data. Trajectory-based representations allow behaviors such as typical paths to be cleanly modeled by simple clustering [16] and events readily defined in terms of individual trajectories such as counterflow [17] or u-turns [16] to be detected. These models, however, depend crucially on the reliability of the underlying tracker, which can be compromised in many realistic situations of interest including crowded scenes with many targets, interobject occlusion, low video resolution, and low frame rates (discontinuous movement).

To improve robustness to missed detections and broken tracks, many recent studies have processed low-level image data directly [18], [19], [4], [10], [5], [3], [15]. To deal with the relatively impoverished input features and to model more complex multi-object behaviors, these studies have focused on developing more sophisticated statistical models than the relatively simple clustering techniques [16] used for track data. Typical approaches include Gaussian mixture models (GMMs) [15], Dynamic Bayesian Networks (DBNs) such as a Hidden Markov Models (HMM) [18], [19], [20], or probabilistic topic models (PTMs) [4] such as Latent Dirichlet Allocation (LDA) [21] or extensions [3]. DBNs

are natural for modeling dynamics of behavior [1], [20], [19], [18]. However, explicit DBN models of multiple object behaviors are often exponentially costly in the number of objects, rendering them intractable for the busy scenes of interest. To overcome the problems of computational complexity and robustness in DBNs, PTMs [21] were borrowed from text document analysis. In the text domain, these models represent documents as a bag of words via a unique mixture of intermediate topics, each of which defines a distribution over words. Applied to behavior analysis, PTMs represent video clips as a unique mixture of activities, each of which defines a distribution over visual words [3], [5], [4], [12]. There are two drawbacks, however: Inference in many PTMs is computationally expensive (preventing the desired usage for real-time monitoring of events in a public space) and they are unsupervised—limiting their accuracy and precluding classification of behaviors. Our proposed WS-JTM addresses the typical PTM weakness of inference speed and exploits weak supervision.

Unsupervised detection of unusual or abnormal behaviors has recently been a topical problem in behavior analysis to which statistical models including DBNs [1], [20], [19], PTMs [4], [3], and hybrids [5], [12] have been applied. In each case, a generative model of typical scene statistics is learned and abnormal behaviors are then detected if they have low likelihood under the learned model. This approach has the advantage of fully automatic operation and no supervision requirements. However, it also has crucial limitations. In addition to limited accuracy, constraints on data separability, and inability to categorize identified earlier, there is also a visual subtlety constraint. Unusual behaviors of genuine interest are often visually subtle (possibly intentionally) compared to more obvious and numerous ongoing typical behaviors. A video containing a subtle unusual behavior may still be typical, on average, and many sophisticated outlier detectors will fail [3], [5]. Alternatively, supervised classification methods have also been applied to behavior analysis [1], [2]. These can perform well given *sufficient and unbiased* labeled training data, and unlike unsupervised methods, they can deal with nonseparable data and classify behaviors. However, they are intrinsically limited in modeling *rare* behaviors due to their absolute sparsity and relative imbalance [22] to typical classes, making it difficult to build a good decision boundary. Moreover, there are still the problems of subtlety: 1) The vast majority of features in a rare class video may be typical due to ongoing typical behaviors and 2) since subtle rare behaviors often have much in common with typical behaviors, there may be few features upon which to discriminate them. These problems mean that even if large amounts of training data are available, conventional classifiers will fail without very specific and expensive supervision localizing each behavior of interest in space and time. In contrast, WS-JTM is capable of learning from sparse and weakly labeled training examples.

Other domains also encounter practical problems in providing full supervision, for example, visual object recognition, where generating training data requires tedious object segmentation. To this end, weakly supervised (WSL) [23], [24] and, in particular, multi-instance learning (MIL) [25], [26] have been exploited. Labels are provided at

image level and weakly supervised algorithms simultaneously learn to localize and classify objects of interest. Typical MIL algorithms are, however, unsuited to modeling behavior because they treat instances independently within each bag. In contrast, our approach builds a topic model to represent the correlations that define complex behaviors. MIL approaches [25], [26], moreover, rely on exploiting large quantities of data (positive instances are not assumed to be rare in absolute number). Our task of learning both rare and subtle behaviors is therefore harder still. We address this challenge by exploiting background class modeling to good effect: Rare behaviors are implicitly identified by their deviation from normality without explicit supervised localization. Thus, we achieve rare and subtle behavior learning where existing methods require more specific supervision (typical supervised classifiers) or numerous examples (typical WSL or MIL).

Other related modeling efforts to ours should be explicitly contrasted: supervised latent Dirichlet allocation (sLDA) [27], delta latent Dirichlet allocation (Δ LDA) [28], and one-shot constellation models [29], [30]. Δ LDA [28] is a weakly supervised topic model applied to understanding bugs in computer software. Our WS-JTM is partially inspired by Δ LDA and improves on it in the following ways:

1. Δ LDA is binary while WS-JTM models multiple classes.
2. Δ LDA is mathematically ill-defined. By constraining Dirichlet parameters to be zero, the likelihood of a Δ LDA model cannot be computed (which prohibits parameter learning, etc.). WS-JTM provides a multiple-model formulation with well-defined likelihoods.
3. Δ LDA requires hand-tuned parameters while WS-JTM exploits hyperparameter learning.
4. Δ LDA is defined only for learning—lacking an inference algorithm to classify new data.

We show how to perform efficient inference in WS-JTM, permitting real-time behavior analysis. sLDA [27] learns a topic model with the additional objective of finding topics that help to discriminate data classes. We will demonstrate, however, that it fails in our subtle and rare behavior context due to making no provision for the imbalanced [22] nature of the problem. Finally, constellation models for object recognition [23] have been learned in a rare class context [29], [30] by transferring prior knowledge learned from common classes to rare classes [31]. There are a few contrasts to be made here. First, this approach is synergistic to ours in that we do not currently exploit transfer learning, but could potentially do so to further improve performance. Second, object recognition is easier than our problem in that 1) it is static, while behaviors are temporally extended, and 2) it is not subtle. Target objects are present in every positive image, and tend to be the main foreground component of the image, and tend to preferentially attract the preprocessing interest-point detectors. All of these points greatly reduce the difficulty of the weak supervision aspect of the object recognition problem compared to our subtle behaviors, which are potentially visible in a minority of pixels for a minority of frames within a training clip. Our joint modeling approach, which implicitly localizes target class features, is therefore more appropriate than transfer learning [29], [30] to model rare and subtle behaviors.

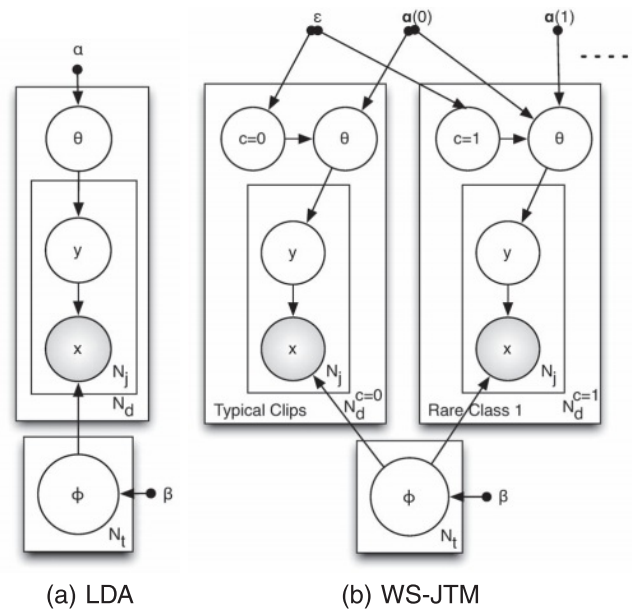


Fig. 2. (a) LDA [21] and (b) our WS-JTM graphical model structures (only one rare class shown for illustration). Shaded nodes are observed.

3 VIDEO FEATURE REPRESENTATION

Our approach uses quantized low-level motion and position features to represent video, as adopted by [3], [5], [14]. For each pixel, we compute an optical flow vector using the Lucas-Kanade algorithm. Next, we spatially divide a scene into $N_a \times N_b$ nonoverlapping square cells, each of which covers $H \times H$ pixels. For each cell, we compute a motion feature by averaging all optical flow vectors in the cell. Finally, each cell motion feature is quantized into one of N_m directions. We note that discretization necessarily imposes a loss of spatial and directional fidelity [15]. This loss can be reduced by increasing discretization resolution at a cost of increased training data requirement and computation time. We found it straightforward to set a suitable discretization such that no object was small enough that its motion was missed in the discrete encoding.

After spatial and directional quantization, we obtain a codebook \mathbf{V} of $N_v = N_a \times N_b \times N_m$ visual words: $\mathbf{V} = \{v_f\}_{f=1}^{N_v}$. This codebook is used to index all the cell motion features and establish a bag-of-words representation of video. To create visual documents, we temporally segment a video into N_d nonoverlapping clips and the visual words from each clip compose the corresponding visual document. Throughout this paper, we denote a corpus of N_d documents as $\mathbf{X} = \{\mathbf{x}_j\}_{j=1}^{N_d}$ in which each document \mathbf{x}_j is a bag of N_j words $\mathbf{x}_j = \{x_{j,i}\}_{i=1}^{N_j}$.

4 WEAKLY SUPERVISED JOINT TOPIC MODELING (WS-JTM)

Our model builds on Latent Dirichlet Allocation (LDA) [21]. Applied to unsupervised behavior modeling [32], [5], [3], LDA learns a generative model of video clips \mathbf{x}_j (Fig. 2a). Each visual word $x_{j,i}$ in a clip is distributed according to a discrete distribution $p(x_{j,i}|\phi_{y_{j,i}}, y_{j,i})$, with parameter Φ indexed by its (unknown) parent activity $y_{j,i}$. Activities are distributed as $p(y|\theta_j)$ according to a per-clip Dirichlet

distribution θ_j . Learning in this model effectively clusters co-occurring visual words in \mathbf{X} and thereby discovers regular activities y in the data set. This activity-based representation of video can facilitate, e.g., querying and similarity matching by searching for clips containing a specified activity profile, or detecting unusual clips by their low likelihood $p(\mathbf{x})$. It also promotes robustness by permitting similarities between clips to be discovered even with few visual words in common.

4.1 Model Structure

In contrast to standard LDA, WS-JTM (Fig. 2b) has two objectives: 1) learning robust and accurate representations for both typical behaviors which are statistically sufficient and a number of rare behavior classes of which few (nonlocalized) examples are available; and 2) classifying clips in test data using the learned model. These will be achieved by jointly modeling the shared aspects of typical and rare clips.

Given a database of N_d clips $\mathbf{X} = \{\mathbf{x}_j\}_{j=1}^{N_d}$, assume that \mathbf{X} can be divided into $N_c + 1$ classes: $\mathbf{X} = \{\mathbf{X}^c\}_{c=0}^{N_c}$ with N_d^c clips per class. The crucial assumption we will make in WS-JTM—which will permit rare behavior modeling in busy scenes—is that clips \mathbf{X}^0 of class 0 contain only typical activities, while class $c > 0$ clips \mathbf{X}^c may contain *both* typical and class c rare activities. We enforce this modeling assumption by partially switching the generative model of clip according to its class (Fig. 2b). Specifically, let T_0 be the N_t^0 element list of typical activities and T_c be the N_t^c element list of rare activities unique to each rare behavior c . Then, typical clips $\mathbf{x} \in \mathbf{X}^0$ are composed from a mixture of activities from T_0 (Fig. 2b, left), while clips $\mathbf{x} \in \mathbf{X}^c$ of each rare class c are composed from a mixture of activities $T_{0,c} \triangleq T_0 \cup T_c$ (Fig. 2b, right). So, while there are $N_t = \sum_{c=0}^{N_c} N_t^c$ activities in total, each clip may be explained by a class specific of subset of activities—in proportions and locations which are unknown and to be determined by the algorithm. By way of contrast to standard LDA, which uses a fixed (and usually uniform) activity hyperparameter α , in WS-JTM the dimension of the per-clip activity proportions θ and prior α are now class dependent. So, if $\alpha(0)$ are the typical activity priors, $\alpha(c)$ the priors unique to class c , and $\alpha \triangleq [\alpha(0), \alpha(1), \dots, \alpha(N_c)]$ is the list of all the activity priors; then typical clips $c = 0$ are generated with parameters $\alpha^{c=0} \triangleq \alpha(0)$ and rare clips $c > 0$ with $\alpha^c \triangleq [\alpha(0), \alpha(c)]$. In this way, we explicitly establish a shared space between common and rare clips which will enable us to differentiate their subtle differences, overcome the problem of sparse rare behaviors, and improve classification accuracy. We can summarize the generative process of WS-JTM as follows:

1. For each activity k , $k = 1, \dots, N_t$;
 - a. Draw a Dirichlet word-activity distribution $\phi_k \sim \text{Dir}(\beta)$;
2. For each clip j , $j = 1, \dots, N_d$;
 - a. Draw a class label $c_j \sim \text{Multi}(\epsilon)$;
 - b. Choose the shared prior $\alpha^{c=0} \triangleq \alpha(0)$ or $\alpha^{c>0} \triangleq [\alpha(0), \alpha(c)]$.

TABLE 1
Summary of Model Parameters

$x_{i,j} = 1 \dots N_v$	i th visual words in clip j
$y_{i,j} = 1 \dots N_t$	i th topic/activity in clip j
$c_j = 1 \dots N_c$	Class of clip j
ϕ_y	Word probability vector for topic y
θ_j	Activity probability vector for clip j
β	Dirichlet word prior
α	Dirichlet activity prior vector
$\alpha(0)$	Typical activity prior vector
$\alpha(c)$, $c > 0$	Rare activity c prior vector

- c. Draw a Dirichlet class-constrained activity distribution $\theta_j \sim \text{Dir}(\alpha^c)$;
- d. For observed words $i = 1, \dots, N_w^j$ in clip j :
 - i. Draw an activity $y_{j,i} \sim \text{Multi}(\theta_j)$;
 - ii. Sample a word $x_{j,i} \sim \text{Multi}(\phi_{y_{j,i}})$.

The probability of variables $\{\mathbf{x}_j, \mathbf{y}_j, c_j, \theta_j\}$ and parameters Φ given hyperparameters $\{\alpha, \beta, \epsilon\}$ in a clip j is

$$p(\mathbf{x}_j, \mathbf{y}_j, \theta_j, \Phi, c_j | \alpha, \beta, \epsilon) = \prod_{t=1}^{N_t} p(\phi_t | \beta) \cdot \prod_{i=1}^{N_w^j} p(x_{j,i} | y_{j,i}, \phi_{y_{j,i}}) p(y_{j,i} | \theta_j) p(\theta_j | \alpha^{c_j}) p(c_j | \epsilon). \quad (1)$$

The probability $p(\mathbf{X}, \mathbf{Y}, \mathbf{c} | \alpha, \beta, \epsilon)$ of a video data set $\mathbf{X} = \{\mathbf{x}_j\}_{j=1}^{N_d}$, $\mathbf{Y} = \{\mathbf{y}_j\}_{j=1}^{N_d}$, $\mathbf{c} = \{c_j\}_{j=1}^{N_d}$ can be factored as

$$p(\mathbf{X}, \mathbf{Y}, \mathbf{c} | \alpha, \beta, \epsilon) = p(\mathbf{X} | \mathbf{Y}, \beta) p(\mathbf{Y} | \mathbf{c}, \alpha) p(\mathbf{c} | \epsilon). \quad (2)$$

Here, the first two terms are products of Polya distributions over activities k and clips j , respectively,

$$p(\mathbf{X} | \mathbf{Y}, \beta) = \int p(\mathbf{X} | \mathbf{Y}, \Phi) p(\Phi | \beta) d\Phi = \prod_{k=1}^{N_t} \frac{\Gamma(N_v \beta)}{\prod_v \Gamma(\beta)} \frac{\prod_v \Gamma(n_{k,v} + \beta)}{\Gamma(\sum_v n_{k,v} + \beta)}, \quad (3)$$

$$p(\mathbf{Y} | \mathbf{c}, \alpha) = \prod_{c=0}^{N_c} \prod_{j=1}^{N_d^c} \int p(y_j | \theta_j) p(\theta_j | \alpha, c_j) d\theta_j = \prod_{c=0}^{N_c} \prod_{j=1}^{N_d^c} \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(n_{j,k} + \alpha_k)}{\Gamma(\sum_k n_{j,k} + \alpha_k)}, \quad (4)$$

where $n_{k,v}$ and $n_{j,k}$ indicate the counts of activity-word and clip-activity associations and k ranges over activities $T_{0,c}$ permitted by the current document class c_j . For convenience, Table 1 summarizes the model parameters.

We next show how to learn a WS-JTM model (training) and use the learned model to classify new data (testing). For training, we assume weak supervision in the form of labels c_j , and the goal is to learn the model parameters $\{\Phi, \alpha, \beta\}$. For testing, parameters $\{\Phi, \alpha, \beta\}$ are fixed and we infer the unknown class of test clips \mathbf{x}^* .

4.2 WS-JTM Learning

We first address learning our WS-JTM from training data $\{\mathbf{X}, \mathbf{c}\}$. This is nontrivial because of the correlated unknown latents $\{\mathbf{Y}, \theta\}$ and parameters $\{\Phi, \alpha, \beta\}$. The correlation

between these unknowns is intuitive as they broadly represent “which activities are present where” and “what each activity looks like.” A standard EM approach to learning with latent variables is to alternate between inference—computing $p(\mathbf{Y}|\mathbf{X}, \mathbf{c}, \boldsymbol{\alpha}, \beta)$ and hyperparameter estimation—optimizing

$$\{\boldsymbol{\alpha}, \beta\} \leftarrow \operatorname{argmax}_{\boldsymbol{\alpha}, \beta} \sum_{\mathbf{Y}} \log p(\mathbf{X}, \mathbf{Y}|\mathbf{c}, \boldsymbol{\alpha}, \beta) p(\mathbf{Y}|\mathbf{X}, \mathbf{c}, \boldsymbol{\alpha}, \beta).$$

Neither of these subproblems have analytical solutions in our case, but we develop approximate solutions for each in the following two sections.

4.2.1 Inference

Similarly to LDA, exact inference in our model is intractable, but it is possible to derive a collapsed Gibbs sampler [33] to approximate $p(\mathbf{Y}|\mathbf{X}, \mathbf{c}, \boldsymbol{\alpha}, \beta, \epsilon)$. The Gibbs sampling update for the activity $y_{j,i}$ is derived by integrating out the parameters Φ and θ in its conditional probability given the other variables

$$p(y_{j,i}|\mathbf{Y}_{-j,i}, \mathbf{X}, \mathbf{c}, \boldsymbol{\alpha}, \beta) \propto \frac{n_{y,x}^{-j,i} + \beta}{\sum_v (n_{y,v}^{-j,i} + \beta)} \frac{n_{j,y}^{-j,i} + \alpha_y}{\sum_k (n_{j,k}^{-j,i} + \alpha_k)}. \quad (5)$$

Iteration of (5) draws samples from the posterior $p(\mathbf{Y}|\mathbf{X}, \mathbf{c}, \boldsymbol{\alpha}, \beta)$. $\mathbf{Y}_{-j,i}$ denotes all activities excluding $y_{j,i}$; $n_{y,x}$ denotes the counts of feature $x_{j,i}$ being associated to activity $y_{j,i}$; $n_{j,y}$ denotes the counts of activity $y_{j,i}$ in clip j . Superscript $-j,i$ denotes counts excluding item (j,i) . In contrast to standard LDA (Fig. 2a), the topics which may be allocated by our joint model (Fig. 2b) are constrained by clip class c to be in $T_0 \cup T_c$. Activities $T_{c=0}$ will be well constrained by the abundant typical data. Clips of some rare class $c > 0$ may use extra activities T_c in their representation. These will therefore come to represent the unique aspects of interesting class c .

Each sample of activities \mathbf{Y} entails Dirichlet distributions over the activity-word parameter Φ and per-clip activity parameter θ_j — $p(\Phi|\mathbf{X}, \mathbf{Y}, \beta)$ and $p(\theta_j|\mathbf{y}_j, c_j, \boldsymbol{\alpha})$. These can then be point estimated by the mean of their Dirichlet posteriors

$$\hat{\phi}_{k,v} = \frac{n_{k,v} + \beta}{\sum_v (n_{k,v} + \beta)}, \quad (6)$$

$$\hat{\theta}_{j,k} = \frac{n_{j,k} + \alpha_k}{\sum_k (n_{j,k} + \alpha_k)}. \quad (7)$$

4.2.2 Hyperparameter Estimation

The Dirichlet prior hyperparameters $\boldsymbol{\alpha}$ and β play an important role in governing the activity-word $p(\mathbf{X}|\mathbf{Y}, \beta)$ and clip-activity $p(\mathbf{Y}|\mathbf{c}, \boldsymbol{\alpha})$ distributions. β describes the prior expectation about the “size” of the activities—how many visual words are expected in each. More crucially, elements of $\boldsymbol{\alpha}$ describe the relative dominance of each activity within a clip of a particular class. That is, in a class c clip, how frequently are observations related to rare activities T_c expected compared to ongoing normal activities T_0 ? Direct optimization

$\{\boldsymbol{\alpha}, \beta\} \leftarrow \operatorname{argmax}_{\boldsymbol{\alpha}, \beta} \sum_{\mathbf{Y}} \log p(\mathbf{X}, \mathbf{Y}|\mathbf{c}, \boldsymbol{\alpha}, \beta) p(\mathbf{Y}|\mathbf{X}, \mathbf{c}, \boldsymbol{\alpha}, \beta)$ in an EM framework is still intractable because of the sum over \mathbf{Y} with exponentially many terms. However, we can use N_s Gibbs samples $\mathbf{Y}_s \sim p(\mathbf{Y}|\mathbf{X}, \mathbf{c}, \boldsymbol{\alpha}, \beta)$ drawn during inference (5) to define a Gibbs-EM algorithm [34], [35] which approximates the required optimization as

$$\{\boldsymbol{\alpha}, \beta\} \leftarrow \operatorname{argmax}_{\boldsymbol{\alpha}, \beta} \frac{1}{N_s} \sum_s \log(p(\mathbf{X}|\mathbf{Y}_s, \beta) p(\mathbf{Y}_s|\mathbf{c}, \boldsymbol{\alpha})). \quad (8)$$

We learn β by substituting (3) into (8) and maximizing for β . The gradient g with respect to β is

$$\begin{aligned} g &= \frac{d}{d\beta} \frac{1}{N_s} \sum_s \log p(\mathbf{X}|\mathbf{Y}_s, \beta) \\ &= \frac{1}{N_s} \sum_s \sum_k \left(N_v \Psi(N_v \beta) - N_v \Psi(\beta) \right. \\ &\quad \left. + \sum_v \Psi(n_{k,v} + \beta) - N_v \Psi(n_{k,\cdot} + N_v \beta) \right), \end{aligned} \quad (9)$$

where $n_{k,v}$ is the matrix of topic-word counts for each E-step sample, $n_{k,\cdot} \triangleq \sum_v n_{k,v}$, and Ψ is the digamma function. This leads [36], [34] to the iterative update

$$\beta^{new} = \beta \frac{\sum_s \sum_k \sum_v \Psi(n_{k,v} + \beta) - N_t N_v \Psi(\beta)}{N_v \sum_s (\sum_k \Psi(n_{k,\cdot} + N_v \beta) - N_t \Psi(N_v \beta))}. \quad (10)$$

Compared to β , learning the hyperparameters $\boldsymbol{\alpha}$ is harder because they are class dependent. A simple approach is to define a completely separate $\boldsymbol{\alpha}^c$ for each class and maximize $p(\mathbf{Y}^c|\boldsymbol{\alpha}^c)$ independently for each c , but this leads to poor estimates for the frequency of typical activities from the point of view of each rare class. This is because the rare class $\boldsymbol{\alpha}^{c>0}$ parameter updates would take into account only a few (possibly 1) clips to constrain the typical elements of $\boldsymbol{\alpha}^c$ although much more data about typical activity are actually available.

To alleviate the problem of statistical insufficiency in learning $\boldsymbol{\alpha}$ we exploit the novel shared-structure approach of WS-JTM (Section 4.1 and Fig. 2b) to develop a new learning algorithm. Specifically, by defining $\boldsymbol{\alpha}^c \triangleq [\boldsymbol{\alpha}(0), \boldsymbol{\alpha}(c)]$ we established a shared space of typical activities $\boldsymbol{\alpha}(0)$ for all classes. This will alleviate the sparsity problem by allowing data from all clips to help constrain these parameters. In the following, we will use K^0 to represent the indices into $\boldsymbol{\alpha}$ of the N_t^0 typical activities, K^c to represent the N_t^c indices of the rare activities in class c , and $K^{c,0} = K^0 \cup K^c$ both typical and class c activities. Hyperparameters $\boldsymbol{\alpha}$ are learned by fixed point iterations (derived in Appendix A, which can be found in the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.81>) of the form

$$\alpha_k^{new} = \alpha_k \frac{a}{b}. \quad (11)$$

For typical activities $k \in K^0$, the terms are

$$a = \sum_{s=1}^{N_s} \sum_{j=1}^{N_d} (\Psi(n_{j,k} + \alpha_k) - \Psi(\alpha_k)), \quad (12)$$

$$b = \sum_{s=1}^{N_s} \sum_{c=1}^{N_c} \sum_{j=1}^{N_d^c} (\Psi(n_{j,c}^{0,c} + \alpha^{0,c}) - \Psi(\alpha^{0,c})) \\ + \sum_{s=1}^{N_s} \sum_{j=1}^{N_d^0} (\Psi(n_{j,c}^0 + \alpha^0) - \Psi(\alpha^0)), \quad (13)$$

where

$$\alpha^0 \triangleq \sum_{k \in K^0} \alpha_k, n_{j,c}^0 \triangleq \sum_{k \in K^0} n_{j,k}, \alpha^{0,c} \triangleq \sum_{k \in K^{0,c}} \alpha_k, n_{j,c}^{0,c} \triangleq \sum_{k \in K^{0,c}} n_{j,k}.$$

For class c rare activities $k \in K^c$, the terms are

$$a = \sum_{s=1}^{N_s} \sum_{j=1}^{N_d^c} (\Psi(n_{j,k} + \alpha_k) - \Psi(\alpha_k)), \quad (14)$$

$$b = \sum_{s=1}^{N_s} \sum_{j=1}^{N_d^c} (\Psi(n_{j,c}^{0,c} + \alpha^{0,c}) - \Psi(\alpha^{0,c})). \quad (15)$$

Iteration of (10) and (11)-(15) estimates the hyperparameters $\{\alpha, \beta\}$ and is used periodically during sampling (5) to complete the Gibbs-EM algorithm.

4.3 WS-JTM Online Classification

In this section, we address inference for unseen video given the learned model $\{\alpha, \Phi\}$ from Section 4.2. Specifically, we classify each test clip \mathbf{x}^* , i.e., determine if it is better explained as a clip containing only typical activities ($c = 0$), or typical activities and some rare activities c ($c > 0$). Note that in contrast to the E-step inference problem of Section 4.2.1 where we computed posterior of *activities* \mathbf{y} via Gibbs sampling, we are now computing the posterior *class* c , which will require the harder task of integrating out activities \mathbf{y} . In this section, we will show how to perform this integration efficiently. The desired class posterior is given by

$$p(c|\mathbf{x}^*, \alpha, \varepsilon, \Phi) \propto p(\mathbf{x}^*|c, \alpha, \Phi)p(c|\varepsilon), \quad (16)$$

$$p(\mathbf{x}^*|c, \alpha, \Phi) = \int \sum_{\mathbf{y}} p(\mathbf{x}^*, \mathbf{y}|\theta, \Phi)p(\theta|\alpha)d\theta. \quad (17)$$

The challenge is that of accurately and efficiently computing the class-conditional marginal likelihood in (17). Efficiently and reliably computing the marginal likelihood in topic models is an active research area [37], [38] due to the intractable sum over correlated \mathbf{y} in (17). We take the view of [37], [38] and define an importance sampling approximation to the marginal likelihood

$$p(\mathbf{x}^*|c) \approx \frac{1}{S} \sum_s \frac{p(\mathbf{x}^*, \mathbf{y}^s|c)}{q(\mathbf{y}^s|c)}, \quad \mathbf{y}^s \sim q(\mathbf{y}|c), \quad (18)$$

where we drop conditioning on the parameters for clarity. Different choices of proposal $q(\mathbf{y}|c)$ induce different estimation algorithms. The (unknown) optimal proposal $q_o(\mathbf{y}|c)$ is $p(\mathbf{x}^*, \mathbf{y}|c)$. We can develop a mean field approximation $q_{mf}(\mathbf{y}|c) = \prod_i q_i(y_i|c)$ with minimal Kullback-Leibler divergence to the optimal proposal by iterating

$$q_i(y_i|c) \propto \left(\alpha_y^c + \sum_{l \neq i} q_l(y_l|c) \right) \hat{\phi}_{y_i, x_i}. \quad (19)$$

The new importance sampling proposal in (19) results in much faster and more accurate estimation of the marginal likelihood (17) than the standard approach [39], [37] of using posterior Gibbs samples. The latter results in the harmonic mean approximation for the likelihood [39], [37] and suffers from 1) requiring a (slow) Gibbs sampler at test time (prohibiting online computation) and 2) the high variance of the harmonic mean estimator (making classification inaccurate). The new proposal is crucial for us because classification speed and accuracy is determined by the speed and accuracy of computing marginal likelihood (17).

In summary, to classify a new clip, we use the importance sampler defined in (18) and (19) to compute the marginal likelihood (17) for each class c (i.e., typical 0, rare 1, 2, ...) and hence the class posterior for that clip (16). Interestingly, classification in our framework is essentially a model selection [40] computation. We must determine (in the presence of numerous latent variables \mathbf{y}) which generative model provides the better explanation of the data: A simple one involving only typical behavior (Fig. 2b, left) or one of a set of more complex models involving both typical and rare behaviors (Fig. 2b, right). The simpler typical-only model is automatically preferred by Bayesian Occam's razor [40], but if there is any evidence of a particular rare activity, the corresponding complex model is uniquely able to allocate the associated rare topic and thereby obtain a higher likelihood.

4.4 WS-JTM Localization

Once a clip has been classified as containing a particular rare behavior, we may moreover be interested in localizing the behavior of interest in space and time. In notable contrast to unsupervised outlier detection approaches to rare behavior detection [3], [5], WS-JTM provides a principled means to achieve this. Specifically, for a test clip j of type c , we determine its activity profile $p(\mathbf{y}_j|\mathbf{x}_j, c, \alpha, \hat{\Phi})$ with Gibbs sampling by iterating

$$p(y_{j,i}|\mathbf{y}_{-j,i}, \mathbf{x}_j, c, \alpha, \hat{\Phi}) \propto \hat{\phi}_{y,x} \frac{n_{j,y}^{-j,i} + \alpha_y}{\sum_k (n_{j,k}^{-j,i} + \alpha_k)}. \quad (20)$$

We then list the visual words i for which the corresponding sampled activity $y_{j,i}$ is a class c rare activity, i.e., $\mathcal{I} = \{i\}$ s.t. $y_{j,i} \in T_c$. The indices of these visual words \mathcal{I} within the clip provide an approximate spatio-temporal segmentation of the behavior of interest. Because parameters α and Φ are already estimated, this Gibbs localization procedure needs many fewer iterations than the initial model learning and is hence much faster.

5 EXPERIMENTS

5.1 Classifying Simulated Data with Ground Truth

In this section, we apply our proposed framework to a simulated data set. This serves three purposes: to illustrate the mechanisms of our model, to validate its correct behavior on data which are nontrivial yet have known ground truth, and to provide insight into its properties compared to other standard approaches as a function of input sparseness which we can control precisely here.

The experiment is illustrated in Fig. 3. We created 11 2D patterns $\{\phi_y\}$ to represent ground-truth activities, in

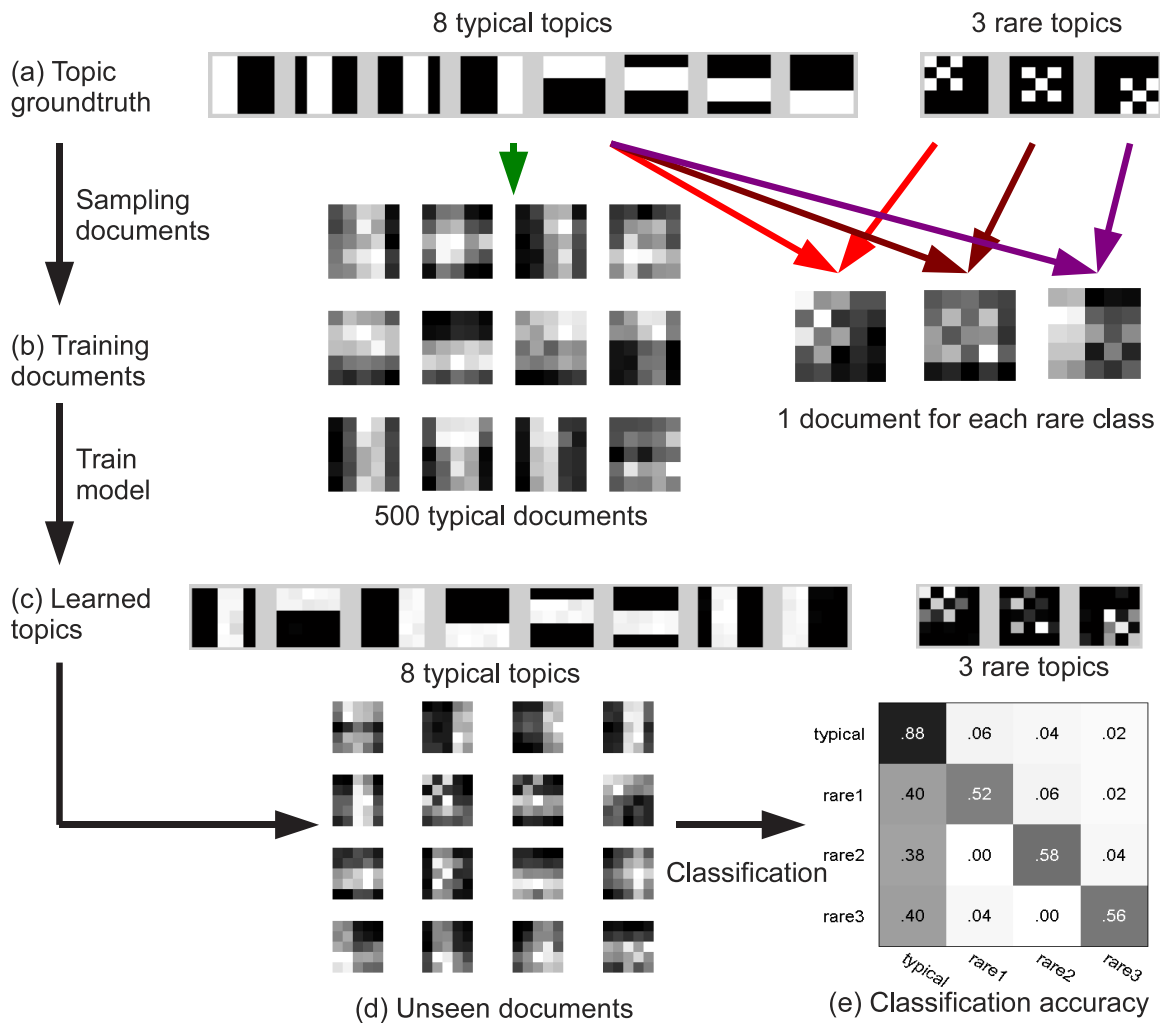


Fig. 3. Illustration and validation of WS-JTM using synthetic data.

which eight (bars) were typical and three (stars) were rare. Following the generative process in Section 4.1, we sampled training documents as follows: First, we generated 500 documents with only typical activities (Fig. 3, second row, middle) and 500 documents for each rare class by sampling both the eight typical activities and corresponding rare activity (Fig. 3, second row, right). We assumed 500 typical documents and varied the number of rare documents for training from 1 to 500, resulting in training set sizes from 503 to 2,000. We also generated a separate test set with 50 documents per class. All Dirichlet hyperparameters α were set to 0.5 and all documents contained 1,000 tokens.

One-shot learning. We trained our WS-JTM with 500 typical documents and one from each rare class, i.e., one-shot learning. The 11 learned activities are shown in Fig. 3c. The model learns a fairly good representation of each rare activity despite having only one noisy and nonlocalized example each (Fig. 3b). This is because it is able to leverage the shared structure and the typical activity representation, which is better constrained by the more numerous typical documents to implicitly localize the rare patterns in the training set. Next we applied the learned model to classify test documents. Fig. 3d shows some test document examples in which each row illustrates a class. The

classification accuracies are shown in the confusion matrix in Fig. 3e.

Quantifying the effect of data sparsity. To illustrate the challenge involved in rare-class learning and validate our contribution over existing state-of-the-art models, we performed a second experiment in which we varied the number of documents in each rare class from 1 to 500. We compared against the following methods:

1. Supervised LDA (sLDA [41], [42]). A supervised topic model classifier. It jointly learns a topic model for the data and a topic profile-based classifier. We utilized the implementation of [42] and set the number of topics to 11.
2. LDA Classifier (LDA-C). Treating LDA as a class-conditional generative model, we learned a separate model [39], [21] for each class of documents with eight topics per class. Dirichlet hyperparameters α and β were learned using [43]. For classification, we computed the test document likelihood under each LDA model class using importance-sampling method proposed in Section 4.3 and then calculated the class posterior assuming equal priors.
3. Multiclass SVM (MC-SVM). A Gaussian kernel classifier was trained directly on visual word counts

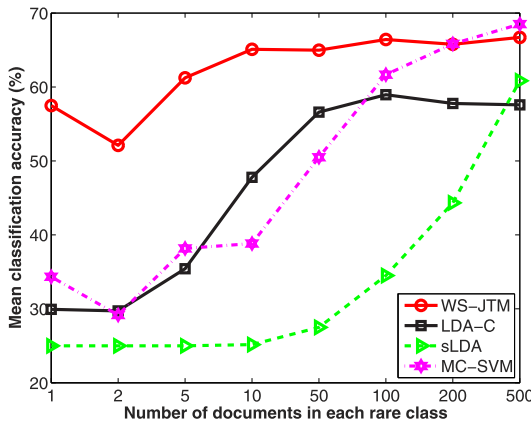


Fig. 4. Synthetic data classification performance as a function of rare-class example sparsity. One-shot learning corresponds to the y -axis. Our WS-JTM exhibits dramatically superior performance in the low data domain.

with hyperparameters $\{C, \gamma\}$ optimized by grid search. To account for data set imbalance, the misclassification cost parameter C_c was weighted on a per-class basis according to the inverse proportion of that class in each training data set [22].

The classification results are shown in Fig. 4. Given only one or a few rare training documents, all existing approaches performed very poorly. Specifically, existing methods classify most test documents as typical, resulting in average accuracy around 30 percent. In contrast, the proposed WS-JTM achieved average accuracy of 58 percent even with one-shot learning. Existing methods approach but do not outperform WS-JTM as the numbers examples per class becomes balanced. This key result shows that for the important task of weakly supervised rare-class learning and classification, WS-JTM provides a decisive advantage over existing techniques.

5.2 Classifying Real-World Rare Behaviors

We evaluate our WS-JTM on classifying rare behaviors in two real-world video data sets: the MIT data set [3] (30 Hz, 720×480 pixels, 1.5 hour) and the QMUL data set [4] (25 Hz, 360×288 pixels, 1 hour). Both scenes featured numerous objects exhibiting complex behaviors concurrently. Figs. 5a and 5d illustrate the behaviors that typically occur in each scene. Unlike many existing studies [3], [4], [5], [12] which focus on learning typical behaviors and their concurrence or temporal correlation, our objective is to learn to classify rare behaviors which are of particular interest to visual surveillance applications. In the MIT data set (Figs. 5a and 5b), we are interested in the illegal left-turn and right-turn at different locations of the scene (red arrows). In the QMUL data set (Figs. 5a and 5b), our targets are the U-turn at the center of the scene and the near-collision situation in which horizontal flow vehicles drive into the junction (red arrow) before turning traffic finishes.

To create a visual word vocabulary, we followed the procedure in Section 3. The videos were spatially quantized into 72×48 cells (MIT) and 72×57 cells (QMUL), respectively, and motion direction in each cell quantized into four orientations (up, down, left, right) resulting in codebooks of 13,824 and 16,416 words, respectively. Each data set was

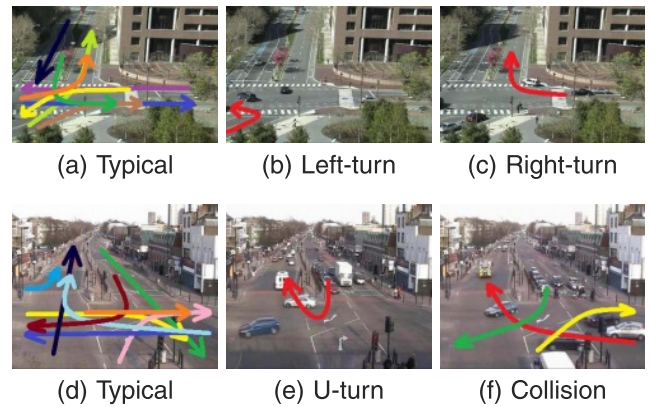


Fig. 5. Example typical and rare behaviors in the (a), (b), (c) MIT and (d), (e), (f) QMUL surveillance data sets.

temporally segmented into nonoverlapping video clips of 300 frames each. We manually labeled the clips into three classes, typical, rare 1 and rare 2, according to which, if any, rare behavior existed in each clip. The total available numbers of clips for each behavior class are detailed in Table 2. Throughout our experiments, we varied the number of training clips for each rare class while keeping constant the number of typical training clips and testing clips for all classes. In each case, we used 20 typical activities and one activity per rare class for ease of visualization.

5.2.1 Learning Activity Models of Rare Behaviors

We first evaluate the proposed WS-JTM on learning activity representations for both typical and rare behaviors. In this experiment, we assume a one-shot learning condition, i.e., the training corpora contain only 1 clip per rare behavior (see Table 2). We apply our Gibbs-EM learning algorithm proposed in Section 4.2 for 2,000 iterations of burn-in with hyperparameter updates every 100 iterations and then draw five samples from the posterior (6). All quantitative results are averages of three folds of cross validation.

MIT data set. The dominant visual words (large $p(x|\phi_y)$) are used to illustrate the some of the learned typical and rare activities ϕ_y . The illustrative typical activities in Fig. 6 show: pedestrians crossing (left column), straight traffic (center column), and turning traffic (right column). Fig. 7 shows that the learned rare activities match the examples in Figs. 5b and 5c, and that these have generally been disambiguated from ongoing typical behaviors (Fig. 5a). This is despite 1) one-shot learning of rare behaviors, 2) rare behavior subtlety in that more numerous typical behaviors were co-occurring overwhelmingly (Fig. 5a), and 3) only

TABLE 2
Number of Clips Used in the Experiments

MIT			
Total	Typical (300)	Rare 1 (26)	Rare 2 (28)
Train	200	1, 2, 5, 10	1, 2, 5, 10
Test	100	16	18
QMUL			
Total	Typical (200)	Rare 1 (12)	Rare 2 (5)
Train	100	1, 2	1, 2
Test	100	10	3

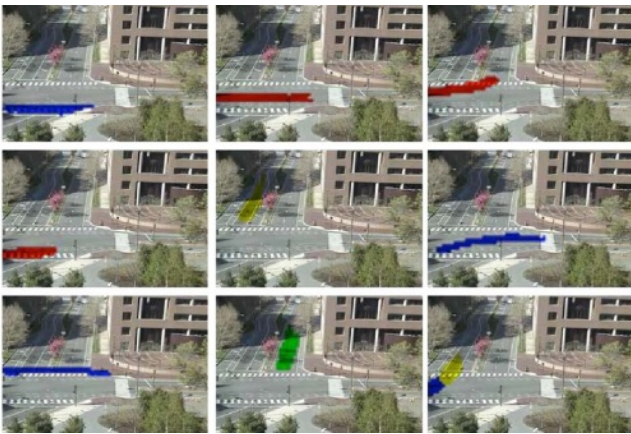


Fig. 6. Typical activities learned from the MIT data set. Red: Right, Blue: Left, Green: Up, Yellow: Down.

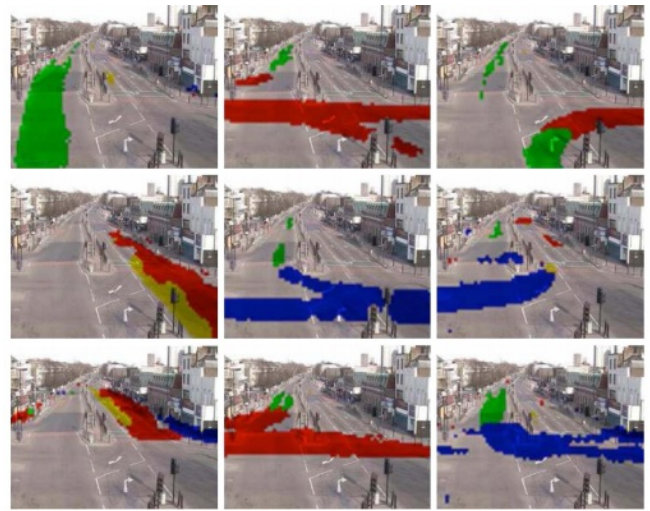


Fig. 8. Typical activities learned from the QMUL data set. Red: Right, Blue: Left, Green: Up, Yellow: Down.



(a) Rare 1: Left Turn

(b) Rare 2: Right Turn

Fig. 7. One shot learning of rare activities: MIT data set.

small differences (potentially confusing similarity) between some rare and typical activities (some arrow segments in Figs. 5b and 5c overlap those in Fig. 5a).

QMUL data set. In Fig. 8, some learned typical activities are illustrated, including vertical traffic (left column), horizontal traffic (center column), and turning traffic (right column). Learning rare activities in the QMUL data set is harder because the scene is much busier, objects are highly occluded, and motion patterns were frequently broken. For example, in the U-turn activity (Fig. 5e), vehicles often drive to the central area and wait for a break in oncoming traffic before continuing. Furthermore, the rare activities are very subtle in that they are both composed mostly of visual words common to typical activities. Our results show that the proposed model is able to cope with such variations to learn descriptions of typical (Fig. 8) and rare (Fig. 9) behaviors.

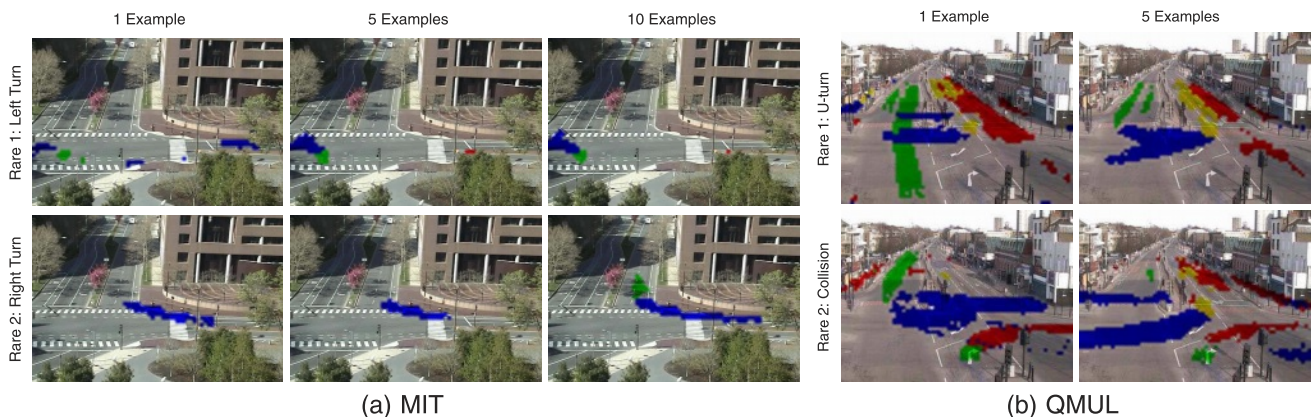


(a) Rare 1: U-turn

(b) Rare 2: Near collision

Fig. 9. One shot learning of rare activities: QMUL data set.

Dependence on data sparsity. The above results were based on one-shot learning. We also explored how additional rare-class examples can improve the learned behavior representation. Fig. 10 illustrates the learned rare behavior models for an increasing number of examples. The simpler MIT data set has more (10) rare-class examples available, and the learned behavior representations are meaningful and accurate (Fig. 10a). The QMUL data set (Fig. 10b) is more challenging and also has fewer available examples (5), so while the learned models capture the essence of the U-turn and collision behaviors, they are not yet perfectly disambiguated from ongoing typical behavior.



(a) MIT

(b) QMUL

Fig. 10. Improvement of rare behavior models with an increasing number of training examples.

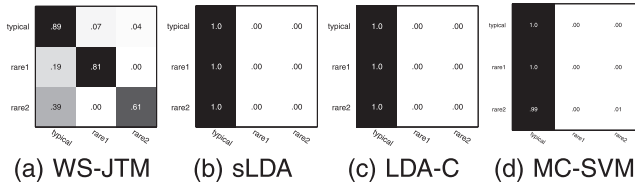


Fig. 11. Classification confusion matrices after one-shot learning: MIT data set.

5.2.2 Classifying Rare Behaviors Online

In this section, we compare the classification performance of WS-JTM against contemporary approaches including sLDA, LDA-C, and MC-SVM (as described in Section 5.1). In each case, results are quantified in terms of the average classification accuracy for each class (i.e., the mean along the diagonal of the normalized confusion matrix). This ensures that errors of each type are weighted equally although the test data are imbalanced.

One-shot learning. All models were learned with one clip from each rare class (see Table 2). For WS-JTM, we used the model learned in last section and classification was performed as described in Section 4.3. For sLDA, we learned 22 topics, to match the total number from WS-JTM. For LDA-C, we learned eight topics per class. We assumed a uniform class prior $p(c|\varepsilon) = 1/3$ for each model. The resulting confusion matrices are shown in Fig. 11 (MIT data set) and Fig. 12 (QMUL data set). It is clear that the model selection approach to classification induced by WS-JTM and implemented by our importance sampler is qualitatively superior to other approaches. WS-JTM achieved 77 and 66 percent average classification rate on the MIT and QMUL data sets. The other approaches generally failed to classify instances of rare behaviors, interpreting almost every behavior as typical resulting in their average accuracy of 33 percent.

Thus far we have considered unbiased maximum likelihood classification. It is also possible to tune a classifier by applying a nonuniform threshold to the posterior $p(c|x^*)$ for declaring the winning class or, equivalently, by setting the class prior $p(c|\varepsilon)$ nonuniformly. This is useful, for example, in many real-world applications with constant false alarm rate (CFAR) constraints. In such applications, it is more important to control the rate at which false alarms distract the operator than to detect every instance of interest. In our context, “false alarms” are instances of typical clips being classified as any of the rare behaviors. We therefore perform a CFAR evaluation by quantifying the average classification accuracy while varying the class label prior $p(c|\varepsilon)$ parameter ε from 0 to 1 such that $p(c=0) = \varepsilon$ and $p(c=1) = p(c=2) = (1-\varepsilon)/2$. This has the effect of biasing classification completely toward the typical behaviors for $\varepsilon = 1$ and toward the rare

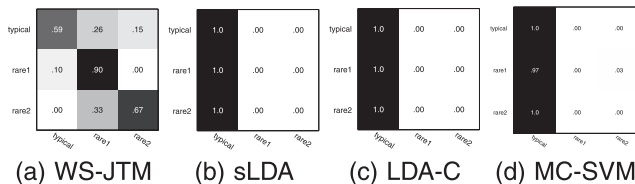


Fig. 12. Classification confusion matrices after one-shot learning: QMUL data set.

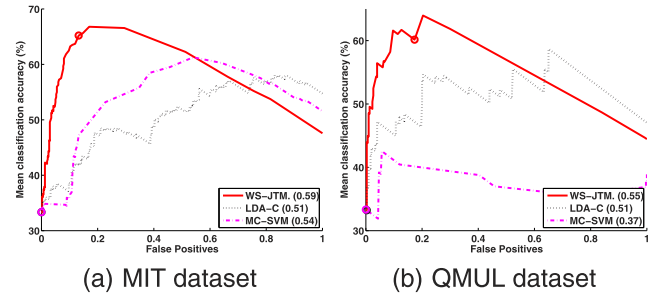


Fig. 13. Average classification accuracy achieved while controlling false alarm rate. Quantity in brackets indicates area under the curve.

behaviors for $\varepsilon = 0$. Fig. 13 details the results. Circles indicate the points corresponding to the unbiased case $p(c|\varepsilon) = 1/3$ from Figs. 11 and 12. The sLDA implementation of [42] provides no posterior estimates, so we do not compare it here. Note that the curves which approach the top left (high classification accuracy with low false alarm rate) or which enclose a greater area should be considered better. WS-JTM's accuracy curves enclose the greatest area (Fig. 13, legends) and, more importantly, approach the top left with a dramatic margin over the other models. That is, performance is most clearly superior in the practically valuable domain of low false alarm rate.

Dependence on data sparsity. We next increased the number of rare class training clips while fixing the typical clips (see Table 2). The average classification accuracy for all methods is reported in Fig. 14. In all cases, WS-JTM produced superior classification accuracy. Comparing Fig. 14 to the classification results for synthetic data in Fig. 4 suggests that this dramatic improvement is because, although we increased the number of rare class training examples, we still do not reach the domain where alternative methods begin to perform well (Fig. 4). This suggests that our WS-JTM is of great value for learning rare behaviors with 1 to 10 examples.

Alternative models. To provide insight into our contribution, we consider why the alternative methods perform poorly on this classification task. The reasons come down to the challenges of learning from weakly labeled subtle examples and from sparse imbalanced data which are better addressed by our model. MC-SVM failed to classify the clips directly from the visual word counts. The subtle and weakly supervised nature of the problem (i.e., the few relevant features corresponding to the rare behaviors are unknown) is too challenging. Without provision for weakly supervised learning, the ongoing and more visually obvious typical

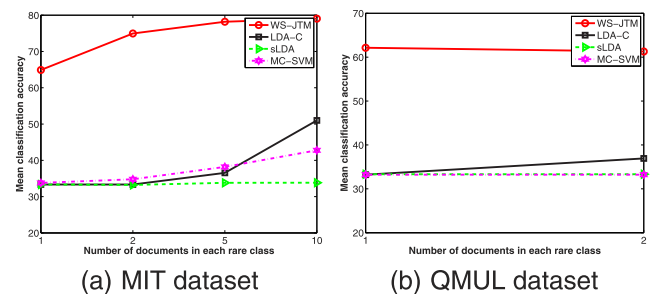


Fig. 14. Average classification accuracy given varying number of rare class training clips.

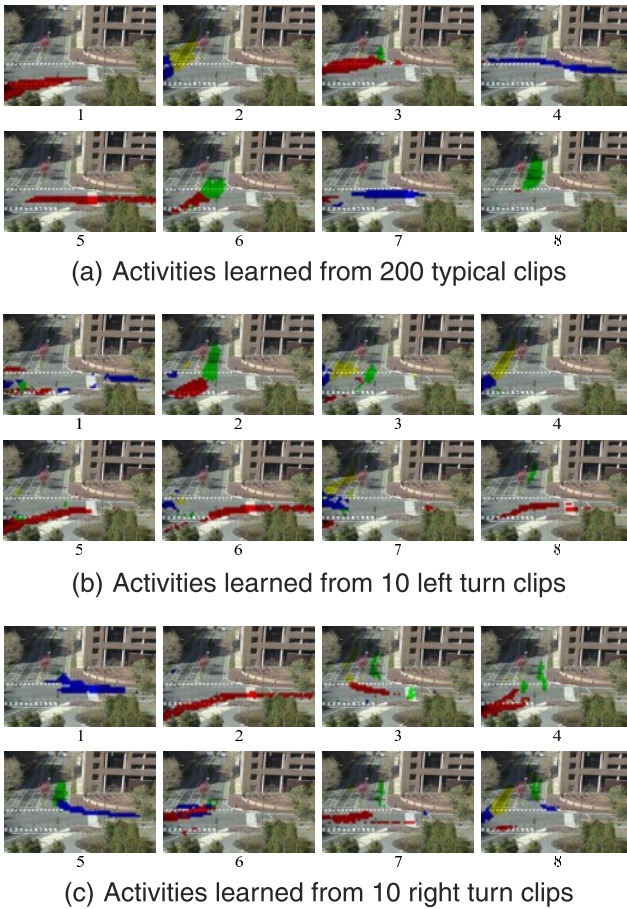


Fig. 15. Activities learned by LDA-C.

behaviors dominate the classification. Multi-instance SVM learning could be considered [26] as an alternative; however, as discussed earlier, the simplifying assumptions made by this technique do not provide a good model for behaviors and, moreover, would require a change of representation to tracks or video cuboids. This would increase computational complexity and susceptibility to noise.

sLDA [41], [42] learned a good typical behavior model, but did not learn any topics corresponding to rare activities, preventing correct classification (figure omitted due to space constraint). This seems to be due to sLDA balancing learning a good generative model of all the data with learning discriminative topics [42]. The benefit of spending topics to learn a better typical behavior model overwhelms the benefit of spending them on learning a discriminative (rare activity) topic since there are so many more typical clips. In other words, sLDA fails due to the imbalanced data. In contrast, by reserving topics/activities for each rare behavior (Fig. 2b), WS-JTM is much more successful at learning from imbalanced data.

LDA-C learns an independent LDA model for each class. In the one-shot learning experiments (Figs. 11, 12, and 13) this was insufficient to learn any clear activities from the rare data. In the experiment with the most data (MIT data set, 10 examples per rare class), LDA-C was able to learn a fair model of each class (Fig. 15). The learned rare activities include the correct ones (Fig. 15b, activity 7; Fig. 15c, activity 5). However, there is a key factor which limits classification accuracy: Each



Fig. 16. Localization of rare activities by WS-JTM.

LDA model has to independently learn the ongoing typical activities. For example, Fig. 15a, activity 1, Fig. 15b, activity 5, and Fig. 15c, activity 2 all represent “right turn from below.” By learning separate models for the same typical activities LDA-C introduces a significant source of noise. In contrast, by learning a single model for ongoing typical activities which is shared between all classes (Figs. 6 and 8), WS-JTM avoids this source of noise. Moreover, by permitting the rare class models to leverage a well-learned typical activity model, they can more accurately learn the rare activities (e.g., Fig. 15b, activity 7 is noisier than the left turn in Fig. 10a).

5.2.3 Localizing Rare Behaviors

In this section, we illustrate the ability of WS-JTM to approximately segment rare behaviors in space and time within a clip classified as rare. Fig. 16 illustrates examples of localizing specific behaviors within correctly classified clips from the previous section. The brighter areas in each image are specific visual words which were labeled as corresponding the rare behavior. The segmentation is rough, due to the noisy optical flow features and the MCMC-based labeling. The bounding boxes of rare behavior words (red rectangles) nevertheless provide a useful localization in space and time. Note that in Fig. 16d, the near-collision behavior is intrinsically multi-object—being defined by the concurrent state of two traffic flows. The bounding box therefore contains elements of both contributing flows—the fire engine and the turning traffic.

5.3 Model Learning and Complexity Analysis

In this section, we provide some additional intuition into our model’s function and validate the significance of two of our specific contributions, namely, the hyperparameter estimation method proposed in Section 4.2.2 and the online inference method proposed in Section 4.3.

Activity profile illustration. To provide additional insight into the mechanism of our model, Fig. 17 illustrates the topic profiles inferred for each of the clips shown earlier

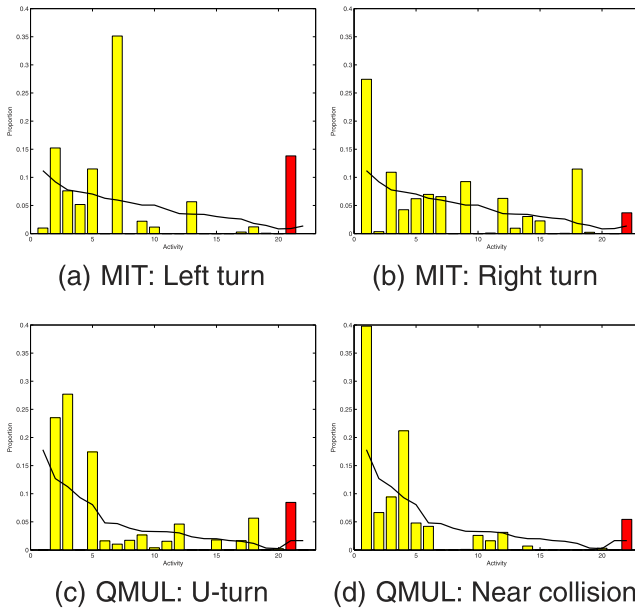


Fig. 17. Estimated topic profile for rare clips from Fig. 16. Bar color indicates typical versus rare activities. Black line indicates the average profile over the entire data set.

in Fig. 16. Clearly the model describes each clip in terms of a unique mixture (Fig. 17, yellow bars) of the learned typical activities (Figs. 6 and 8). The black lines in Fig. 17 illustrate the overall representation of each activity in the test data sets, within which the rare activities (21 and 22) are a small proportion. In contrast, these particular rare class clips require a greater than average proportion (Fig. 17, red bars) of rare activities (Figs. 7 and 9) to be explained.

Hyperparameter estimation. Thus far, hyperparameters $\{\alpha^c\}$ were learned using the method proposed in Section 4.2.2—by allowing typical behavior parameters $\alpha(0)$ to be shared across all classes. In this section, we compare this against two other simpler methods. In the first, all values in $\{\alpha^c\}$ were set to 0.1 (constant). In the second, we naively learned $\{\alpha^c\}$ for each class independently using only clips of the corresponding class and without a shared component. The results in Fig. 18 verify that our proposed algorithm performs best. As we discussed earlier, learning the hyperparameters independently is an especially poor choice for rare-class problems because (e.g., Figs. 18a and 18b) because their typical-class parameters are poorly constrained by the limited data.

Likelihood computation. In the previous experiments, classification was performed via the likelihood computed by the algorithm proposed in Section 4.3. We compared this against the classification performance of WS-JTM using the

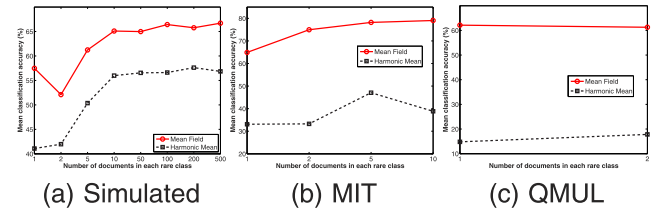


Fig. 19. Classification accuracy using different methods for computing likelihoods of unseen documents.

commonly used harmonic mean likelihood approximation [37], [39]. As seen in Fig. 19, and especially in real-world data, our proposed method significantly outperformed the harmonic mean, despite being more than an order of magnitude faster.

Computational complexity. Quantifying the computational cost of MCMC learning in any model is challenging as assessing convergence is itself an open question [33]. In training, our algorithm is dominated by the $\mathcal{O}(N_w N_T)$ cost of resampling N_w visual words for N_T activities per Gibbs-sweep. In inference, our algorithm is dominated by the $\mathcal{O}(N_w N_T)$ cost of iterating (19). In practice, training in our model (C implementation) proceeded at 11 and 2 FPS for the MIT and QMUL data sets, respectively, while testing (Matlab implementation) proceeded at 84 and 46 FPS, respectively, on a 3 GHz PC.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced the task of weakly supervised learning and classification of rare and subtle behaviors. Rare and subtle behavior learning is of great practical use for behavior-based video analysis, as behaviors of interest for identification are often both rare and subtle. Moreover, the ability to work with weak supervision is increasingly important given the need to increase automation and reduce manpower requirements. These tasks have proven challenging due to existing approaches being challenged by some combination of relevant conditions including:

1. busy and cluttered scenes with occlusion,
2. relative imbalance between few rare behavior examples and numerous typical behaviors,
3. absolute sparsity of rare behavior examples,
4. subtlety of rare behaviors compared to ongoing typical behaviors, or
5. onerous supervision requirements/inability to deal with weak supervision.

We introduced WS-JTM to address these issues. WS-JTM leverages its model of typical activities to help locate and accurately model specific rare and subtle activities. This permits learning of rare and subtle activities despite the challenging combination of weak supervision and sparse data learning for which contemporary methods fail dramatically. Classification in WS-JTM is performed by Bayesian model selection: computing whether the typical model alone is sufficient to explain a new clip or if a more complex model which can allocate additional rare activities is required. Classification accuracy is enhanced by our shared typical activity model and by our shared hyperparameter learning approach, which alleviate the effect of data

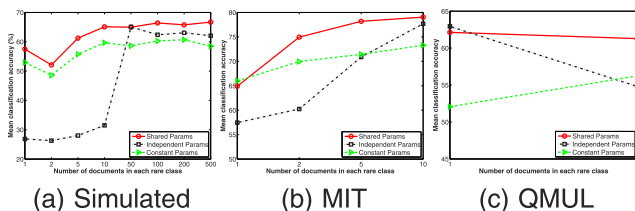


Fig. 18. Classification accuracy using different methods for estimating Dirichlet hyperparameters $\{\alpha^c\}$.

sparsity in learning. Finally, our inference algorithm permits dramatically faster and more accurate online inference than the typical harmonic mean approach. The result is a framework which uniquely enables robust one-shot learning of nonlocalized rare and subtle activities in clutter, and real-time activity classification and localization in test video.

In this study, we have applied our model in a weakly supervised context and solely to video surveillance data. In principle, the same model can be used for fully unsupervised learning and can be usefully applied to any classification or detection problem where interesting instances are rare and embedded within typical instances [28]. We will explore these avenues in future work.

We see two main limitations of our current model: There is no leveraging of typical activities as components to explain rare activities, and our learning approach thus far is nonadaptive. These suggest two ways to generalize our approach in future research, specifically transfer learning [31], [30] and online active learning [44]. Transfer learning aims to leverage underlying commonalities between classes so as to better learn rare classes using generic knowledge obtained from similar classes with more examples. This idea is synergistic with our approach in this paper and is amenable to implementation with topic-model-based hierarchical behavior models [5], [3]. Online learning is also relevant to our rare class categorization problem and the general aim of increasing automation as, with few rare class instances, it is especially helpful to incorporate any new examples. More interestingly, rare and subtle instances may also be nonobvious to humans, so a useful capability to further reduce supervision requirements and increase accuracy is online active learning [45]. The model itself can search for potential rare class examples or those which help to distinguish them and actively query these for labels.

REFERENCES

- [1] T. Xiang and S. Gong, "Beyond Tracking: Modelling Activity and Understanding Behaviour," *Int'l J. Computer Vision*, vol. 61, no. 1, pp. 21-51, 2006.
- [2] N. Robertson and I. Reid, "A General Method for Human Activity Recognition in Video," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 232-248, 2006.
- [3] X. Wang, X. Ma, and E. Grimson, "Unsupervised Activity Perception by Hierarchical Bayesian Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539-555, Mar. 2009.
- [4] J. Li, S. Gong, and T. Xiang, "Global Behaviour Inference Using Probabilistic Latent Semantic Analysis," *Proc. British Machine Vision Conf.*, 2008.
- [5] T. Hospedales, S. Gong, and T. Xiang, "A Markov Clustering Topic Model for Behaviour Mining in Video," *Proc. IEEE 12th Int'l Conf. Computer Vision*, 2009.
- [6] I. Saleemi, K. Shafique, and M. Shah, "Probabilistic Modeling of Scene Dynamics for Applications in Visual Surveillance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1472-1485, Aug. 2009.
- [7] N. Johnson and D. Hogg, "Learning the Distribution of Object Trajectories for Event Recognition," *Image and Vision Computing*, vol. 8, pp. 609-615, 1996.
- [8] W. Hu, T. Tan, L. Wang, and S. Maybank, "A Survey on Visual Surveillance of Object Motion and Behaviors," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Rev.*, vol. 34, no. 3, pp. 334-352, Aug. 2004.
- [9] X. Wang, K. Tieu, and E. Grimson, "Learning Semantic Scene Models by Trajectory Analysis," *Proc. European Conf. Computer Vision*, 2006.
- [10] Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal Events Detection Based on Spatio-Temporal Co-Occurrences," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.
- [11] M.D. Breitenstein, H. Grabner, and L.V. Gool, "Hunting Nessie—Realtime Abnormality Detection from Webcams," *Proc. IEEE Int'l Workshop Visual Surveillance*, 2009.
- [12] D. Kuettel, M. Breitenstein, L. Van Gool, and V. Ferrari, "What's Going On? Discovering Spatio-Temporal Dependencies in Dynamic Scenes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.
- [13] R. Mehran, A. Oyama, and M. Shah, "Abnormal Crowd Behavior Detection Using Social Force Model," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.
- [14] Y. Yang, J. Liu, and M. Shah, "Video Scene Understanding Using Multi-Scale Analysis," *Proc. IEEE 12th Int'l Conf. Computer Vision*, 2009.
- [15] I. Saleemi, L. Hartung, and M. Shah, "Scene Understanding by Statistical Modeling of Motion Patterns," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.
- [16] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A System for Learning Statistical Motion Patterns," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1450-1464, Sept. 2006.
- [17] J. Berclaz, F. Fleuret, and P. Fua, "Multi-Camera Tracking and Atypical Motion Detection with Behavioral Maps," *Proc. European Conf. Computer Vision*, 2008.
- [18] T. Duong, H. Bui, D. Phung, and S. Venkatesh, "Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.
- [19] T. Xiang and S. Gong, "Video Behavior Profiling for Anomaly Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 893-908, May 2008.
- [20] T. Xiang and S. Gong, "Activity Based Surveillance Video Content Modelling," *Pattern Recognition*, vol. 41, pp. 2309-2326, 2008.
- [21] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
- [22] H. He and E. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowledge and Data Eng.*, vol. 21, no. 9, pp. 1263-1284, Sept. 2009.
- [23] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, 2003.
- [24] L. Cao and L. Fei-Fei, "Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes," *Proc. IEEE Int'l Conf. Computer Vision*, 2007.
- [25] P. Viola, J. Platt, and C. Zhang, "Multiple Instance Boosting for Object Detection," *Proc. Neural Information Processing Systems*, 2005.
- [26] M.H. Nguyen, L. Torresani, F. de la Torre, and C. Rother, "Weakly Supervised Discriminative Localization and Classification: A Joint Learning Process," *Proc. IEEE 12th Int'l Conf. Computer Vision*, 2009.
- [27] X. Wang and E. Grimson, "Spatial Latent Dirichlet Allocation," *Proc. Neural Information Processing Systems*, 2007.
- [28] D. Andrzejewski, A. Mulhern, B. Liblit, and X. Zhu, "Statistical Debugging Using Latent Topic Models," *Proc. European Conf. Machine Learning*, 2007.
- [29] L. Fei-Fei, R. Fergus, and P. Perona, "A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories," *Proc. IEEE Ninth Int'l Conf. Computer Vision*, 2003.
- [30] L. Fei-Fei, R. Fergus, and P. Perona, "One-Shot Learning of Object Categories," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594-611, Apr. 2006.
- [31] S.J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Trans. Knowledge and Data Eng.*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.
- [32] J. Li, S. Gong, and T. Xiang, "Discovering Multi-Camera Behaviour Correlations for On-the-Fly Global Activity Prediction and Anomaly Detection," *Proc. IEEE Int'l Workshop Visual Surveillance*, 2009.
- [33] *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson, and D. Spiegelhalter, eds. Chapman & Hall, 1995.
- [34] H. Wallach, "Topic Modeling: Beyond Bag-of-Words," *Proc. 23rd Int'l Conf. Machine Learning*, 2006.
- [35] C. Andrieu, N.D. Freitas, A. Doucet, and M.I. Jordan, "An Introduction to MCMC for Machine Learning," *Machine Learning*, vol. 50, pp. 5-43, 2003.

- [36] T.P. Minka, "Estimating a Dirichlet Distribution," technical report, Microsoft, 2000.
- [37] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation Methods for Topic Models," *Proc. Int'l Conf. Machine Learning*, 2009.
- [38] W. Buntine, "Estimating Likelihoods for Topic Models," *Proc. Asian Conf. Machine Learning*, 2009.
- [39] T.L. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proc. Nat'l Academy of Sciences USA*, vol. 101, pp. 5228-5235, 2004.
- [40] D. Mackay, "Bayesian Interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415-447, 1991.
- [41] D. Blei and J. McAuliffe, "Supervised Topic Models," *Proc. Neural Information Processing Systems*, 2007.
- [42] C. Wang, D. Blei, and F.-F. Li, "Simultaneous Image Classification and Annotation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.
- [43] G. Heinrich, "Parameter Estimation for Text Analysis," technical report, Univ. of Leipzig, 2005.
- [44] B. Settles, "Active Learning Literature Survey," Technical Report 1648, Univ. of Wisconsin-Madison, 2009.
- [45] C.C. Loy, T. Xiang, and S. Gong, "Stream Based Active Anomaly Detection," *Proc. Asian Conf. Computer Vision*, 2010.



Timothy M. Hospedales received the PhD degree in neuroinformatics from the University of Edinburgh in 2008. Currently, he is a postdoctoral researcher with the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include probabilistic modeling and machine learning applied variously to computer vision, behavior understanding, sensor fusion, human-computer interfaces, and neuroscience.



Jian Li received the PhD degree in computer vision and experimental psychology from the University of Bristol in 2008. From 2007 to 2010, he worked as a postdoctoral research assistant in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include machine learning and computer vision-based video behavior and activity analysis, scene context understanding, and classification.



and recognition; face and expression recognition; gesture and action recognition; visual behavior profiling and recognition. He is a fellow of the Institution of Electrical Engineers and a fellow of the British Computer Society.



Tao Xiang received the PhD degree in electrical and computer engineering from the National University of Singapore in 2002. Currently, he is working as a lecturer in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, statistical learning, video processing, and machine learning, with focus on interpreting and understanding human behavior.

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.