10-2013

Understanding Vehicular Traffic Behavior from Video: A Survey of Unsupervised Approaches

Brendan Tran Morris
University of Nevada, Las Vegas, brendan.morris@unlv.edu

Mohan Manubhai Trivedi
University of California - San Diego, mtriveid@ucsd.edu

Journal of Electronic Imaging

SPIEDigitalLibrary.org/jei

Understanding vehicular traffic behavior from video: a survey of unsupervised approaches

Brendan Tran Morris
Mohan Manubhai Trivedi

Journal of Electronic Imaging 22(4), 041113 (Oct–Dec 2013)

Understanding vehicular traffic behavior from video: a survey of unsupervised approaches

Brendan Tran Morris
University of Nevada
Department of Electrical and Computer Engineering
Las Vegas, Nevada 89123
E-mail: brendan.morris@unlv.edu

Mohan Manubhai Trivedi
University of California
Department of Electrical and Computer Engineering
San Diego, La Jolla, California 92093
E-mail: mtrivedi@ucsd.edu

Abstract. *Recent emerging trends for automatic behavior analysis and understanding from infrastructure video are reviewed. Research has shifted from high-resolution estimation of vehicle state and instead, pushed machine learning approaches to extract meaningful patterns in aggregates in an unsupervised fashion. These patterns represent priors on observable motion, which can be utilized to describe a scene, answer behavior questions such as where is a vehicle going, how many vehicles are performing the same action, and to detect an abnormal event. The review focuses on two main methods for scene description, trajectory clustering and topic modeling. Example applications that utilize the behavioral modeling techniques are also presented. In addition, the most popular public datasets for behavioral analysis are presented. Discussion and comment on future directions in the field are also provided.* *[DOI: 10.1117/1.JEI.22.4.041113]*

1 Introduction

Modern governments invest heavily in the installation and maintenance of road networks due to safety concerns and their vital role in economic health. Recently, cameras have become an integral part of many transportation management centers (TMCs) because they give traffic operations engineers a way to view what is happening in the field. However, most of these cameras are only in use sporadically and rarely monitored actively. Monitors are often set to cycle through the cameras until an operator notices an incident. These cameras offer a rich data stream for understanding the roadway and is a virtually untapped operational resource.

Researchers have long recognized the potential of camera monitoring systems. A much larger number of cameras can be actively monitored with computer assistance than a human possibly could. However, computer monitoring progress has been slow. Robust detection and tracking systems for vehicles have been published, but they still suffer under adverse lighting and weather conditions or in densely occupied scenes. In addition, a shortcoming of kinematic and dynamic motion models—for predicting future state evolution in time—is that their performance typically degrades quickly with increasing prediction time-horizons. This is particularly apparent when motion is complex (e.g., U-turn) or when an internal state or intention motivates motion (decision to turn at an intersection). In these cases, the dynamic models are not able to be precisely modeled.

In order to address these shortcomings and enable longer-term predictive capabilities, researchers have developed new analysis approaches based on machine learning that are motivated by the observation that surveillance motion is typically constrained by environmental structure. By observing motion over time, the typical motion patterns can be learned and used as *a priori* knowledge for prediction. The advantage of these techniques is that they are generally applicable and do not require manual retraining for new scenes or scenarios.

This paper provides a review of the changing landscape in traffic behavior understanding. It highlights the shift from explicit high-resolution tracking of vehicle state and the definition of events of interest toward machine learning approaches to leverage video data for meaningful pattern extraction. By utilizing unsupervised machine learning techniques, complex scenes can be described and analyzed automatically through general behavior analysis frameworks. There is a need for this type of survey because it focuses on higher-level understanding of traffic scenes. Recent surveys[1,2] have examined computer vision techniques for traffic analysis, but focus more on the low-level vision processes of detection, classification, and tracking of vehicles with very little time devoted to analysis of the behaviors with respect to transportation.

This review focuses on two particular methods for automatic description: (1) trajectory clustering and (2) topic modeling. The scene descriptions resulting from these methods provide contextual priors on observable motion, which can be used to answer behavior questions, such as where a vehicle is going, how many vehicles are performing the same action, and to detect abnormalities. Finally, a list of the most

Paper 13229SSV received Apr. 24, 2013; revised manuscript received Jul. 12, 2013; accepted for publication Jul. 29, 2013; published online Sep. 10, 2013.

popular public datasets for behavior analysis is presented to promote needed benchmarking in the field.

2 Visual Traffic Behavior

The term behavior has many different definitions depending upon the discipline and field. The following section provides a description of traffic behaviors from the transportation engineering and computer vision perspectives. We further define the scope of traffic behaviors within the surveillance setting.

Transportation engineers view behavior in terms of network capacity management. Infrastructure systems, such as inductive loops and radars, have been installed on roads to collect the traditional traffic parameters of density, flow, and speed. These measures of speed and counts of vehicles are essential for network calibration, simulation, and support computation of annual average daily traffic used in planning. Since transportation engineers are concerned with network performance, their notion of behavior is at a large scale with a bird's eye view. They use the information from a large number of distributed sensors in aggregate for understanding network behavior.

In contrast, computer vision researchers consider behavior understanding to be the ability to analyze and recognize moving patterns and describe them using natural language.[1] In addition, the scope of the behaviors is limited to the view of the camera. Describing a behavior becomes complicated due to the inherent complexity in defining a behavior as well as camera imaging limitations. Behavior can be defined at various levels of complexity, e.g., simple behavior could be a single car speeding or more complex sequences of events such as checkpost violations or multivehicle maneuvers, such as lane-changing and passing. However, typical traffic video resolution is low and vehicles are small in size, making it extremely difficult to extract complex descriptors such as pose with real accuracy. Instead, only simple features, such as position and velocity, can be used to characterize behavior. In either case, a general vocabulary is needed to describe behavior based on context.

Since traffic video is intended to be viewed by personnel at TMCs, traffic behaviors need to be consistent with human operator expectations. Behavior must be considered from the far-field surveillance perspective since this is the typical traffic camera configuration. Motion (or lack of motion) is the major behavioral cue. In particular, behaviors of interest will be concerned with the origin and destination of vehicles and pedestrians as well as the maneuvers, like turns, that are made. It will be particularly important to make accurate predictions on the future behavior of a traffic participant. That is, it should be possible to predict when, where, and how the participant will move. Finally, abnormal or atypical events need to be recognized and detected in a timely manner since they likely indicate an event for more attention. An example would be an illegal U-turn at an intersection. This would need to be detected since it could be the precursor to an accident or indicate some other disturbance in the system.

Video provides rich information content that cannot be obtained with the traditional spot sensors used by transportation engineers. Images allow behavior analysis in the same modality that humans use to sense the world. This leads to a more natural definition of traffic behavior in terms of the way vehicles move in a scene—either how they move in a local (individual) sense, such as making a right turn, or in a more global (collective) view, such as a traffic jam or the phase cycling at an intersection.

3 Learning Behavior with Trajectories

A popular framework for scene understanding and behavior analysis is trajectory-based learning. This framework is attractive because it can be used to augment the typical surveillance pipeline since the basic task of most surveillance systems is the detection and tracking of objects of interest. In addition, the trajectory abstraction makes the framework applicable given any tracking algorithm that is able to provide coherent trajectories.

The general, simplified trajectory learning block diagram is provided in Fig. 1. During the training phase, moving objects, such as vehicles, are observed over a long time period. Trajectories are extracted from the video during observation to build a training database. The training trajectories are then clustered into groups of similar trajectories. These clusters represent the prototypical patterns of motion encountered in a scene. Often the patterns are parametrically modeled for efficient representation. The resulting models provide a description of both the scene and the typical behaviors and can be utilized to characterize new trajectories during online analysis. In particular, the activity models can be used to predict the future state and detect anomalies in real time.

There are two key questions that arise in trajectory learning:

- How can trajectories of varying length be compared (clustered) in a manner that ensures all semantically meaningful patterns are extracted in a completely unsupervised fashion?
- Given a grouping of similar trajectories, how should they be modeled and parameterized to capture the differences between clusters while providing a computationally efficient inferencing scheme?

The following sections highlight prominent techniques for trajectory clustering and modeling with representative works collected in Table 1. While the general framework has presented these two concepts as separate modules, it is often the case that the clustering and modeling are

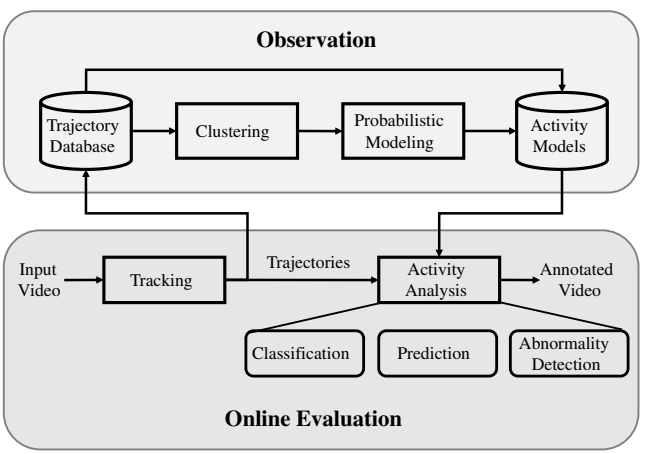


Fig. 1 The general framework for trajectory analysis.[3] During the observation (training) phase, trajectories are clustered and modeled. The learned set of typical patterns are used for live activity analysis in the online evaluation phase.

Table 1 Representative trajectory clustering studies.

Research study	Classification/ inference	Description
Makris and Ellis[5]	Route envelope	Online adaptive route envelope model between important scene zones.
Piciarelli and Foresti[6]	Probability tree	A tree of clusters is built to share common areas for prediction and anomalous track identification.
Vasquez, Fraichard, and Laugier[7]	Growing HMM	Online adaptive modeling of the different ways to move from scene goals.
Hu et al.[8]	Gaussian chain	Two-level hierarchy to first spatially cluster resampled data and then cluster with zero-padded temporal trajectory with fuzzy c-means.
Morris and Trivedi[3]	HMM	Three-level learning hierarchy accounting for goals, spatial paths, and dynamic behaviors.
Noceti and Odone[9]	Common strings	Image space quantized with Voronoi tessellation to build an alphabet for representing a trajectory as a string.

performed jointly in a single processing step since they are intimately related. The interested reader can see the 2008 review[4] for a more complete treatment of trajectory learning techniques.

3.1 *Clustering Trajectories*

When considering behavior through trajectories, it would be beneficial to utilize traditional clustering algorithms. However, the difficulty is that most traditional clustering algorithms operate on data that are assumed to lie in a fixed-dimensional space (fixed-size data). In general, trajectories will not be of fixed size due to variations in the exhibited behaviors of individuals as well as differences in the speed in which they are performed. In order to enable compatibility, various trajectory-specific similarity measures (distances) have been proposed. These trajectory-specific measures are designed to explicitly handle unequal data length in a semantically meaningful way.

Perhaps the most straightforward method for comparing trajectories is to interpolate and resample to a fixed size for L2 norm computation. However, this resample process typically considers only position data and destroys the temporal sequence information. Therefore, both a vehicle traveling slowly and another quickly will be considered as the same behavior, which may not be desirable. Other simple size normalization schemes are zero padding or replication of values to augment the trajectory to a fixed size. As an example, Hu et al.[10] designed their trajectory distance measure as the average Euclidean distance between consecutive trajectory points after length normalization. This retains some of the temporal characteristics but still may not accurately reflect behavioral semantics.

The most popular trajectory comparison methods, alignment techniques, directly account for the length variation between tracks. These alignment techniques try to capture the similarity between trajectories by finding correspondences between the individual observations in a trajectory. Dynamic time warping (DTW)[11] finds a match between all samples in a pair of trajectories using a dynamic program. The longest common subsequence (LCSS) is an extension of DTW that comes from string matching literature. LCSS relaxes the match criteria such that not every single sample must have a corresponding sample in the other trajectory. The relaxed matching provides more robustness to noise and outliers that may come from errors in the tracking process. Piciarelli and Foresti[6] defined a distance measure that accounted for temporal drift. They noted that as a trajectory got longer (observed for more time), there was more opportunity for corresponding samples between trajectories to be misaligned. Their distance measure accounted for this drift by having a growing search window for matching related to the trajectory length. The common Hausdorff distance between sets was modified to account for sequence ordering by Atev et al.[12] Their measure considers matching of trajectory samples based upon equal ratios of the total track length.

Although alignment techniques better account for unequal lengths between trajectories, they usually require appropriate setting of alignment/warping parameters, which affects the performance. The alignment techniques also require appropriate normalization of features to weight contributions between different behavior attributes, e.g., position versus velocity. Unfortunately, the appropriate similarity metric is often strongly tied to the application requirements rather than being universally obvious.

3.2 *Modeling Trajectory Clusters*

After grouping trajectories into clusters, the resulting clusters are modeled for efficient inferencing. The modeling step must compactly represent the underlying behavior while still incorporating all significant details. For example, other than just the location pedestrians travel on the sidewalk, it may be desirable to augment the location information with velocity and curvature information to enable distinction between direction of travel and how smoothly the pedestrian moves.[13]

Trajectory-based behaviors have been modeled in two main ways as shown in Fig. 2. The first considers a behavior in its entire end-to-end existence (from entry to exit from the scene). The second method decomposes the trajectory into smaller connected subparts for shared description between overlapping behaviors.

Given all trajectories in a cluster, the end-to-end behavior can be simply characterized by an average trajectory. The average could be the centroid of the cluster, when using a fixed-length representation, or a cluster prototype (e.g., a random trajectory within the set). The average trajectory can be considered as the expected observation given a particular behavior. The average can be further augmented by developing an envelope to specify the extent or variance within a cluster. The envelope developed by Makris and Ellis [Fig. 2(a)] was formed by finding the farthest point in the normal direction from each node (sample) in a cluster.[5] However, this definition was susceptible to noise and could

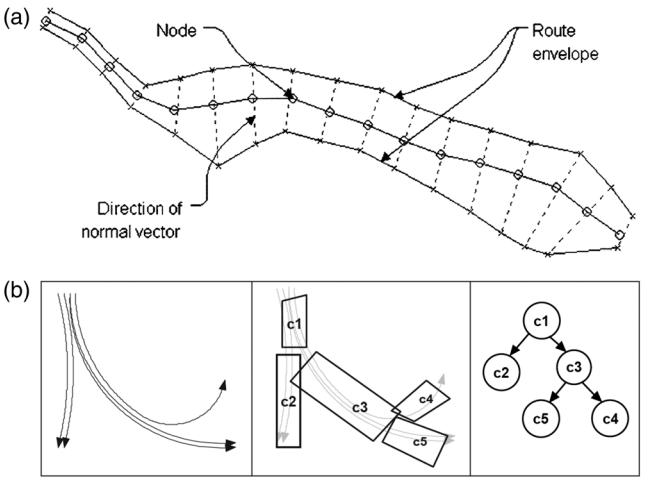


Fig. 2 (a) The trajectory route envelope defines an average trajectory in a cluster and gives bounds on the variance within the cluster.[5] (b) Representation of a trajectory as a tree of clusters[6] allows sharing of trajectory data in subclusters. In addition, a graphical tree model can be used to describe the activity at a high level based on state transitions. (This also provides a dynamic update—merge and split—ability.)

be made more robust by probabilistically defining the envelope. Gaussian distributions have been widely used for this purpose.[3,8]

In contrast, the subtrajectory methods divide a behavior into smaller shared units. The full end-to-end behavior can then be explained by the smaller subunit traversal string [Fig. 2(b)]. The smaller units are meant to represent similar regions that should have semantically similar description. The decomposition into subunits has been accomplished based on curvature or due to splits.[6] The decomposition forms a graphical tree structure in which the probability of traversal can be estimated from the data, providing a natural prediction framework.

3.3 *Discussion of Trajectory Learning*

Trajectory learning has become quite popular in the surveillance community because it nicely fits into its traditional areas of expertise—detection and tracking. Therefore, behavior analysis can be easily added on top of existing architectures. It provides a natural method to describe the behavior of individuals in the scene, which enables exact localization of anomalous events. However, it does not inherently provide a probabilistic interpretation of a behavior and does not handle behaviors defined through groups of individuals.

3.3.1 *Dependence on tracking*

While trajectory clustering is conceptually simple and fits nicely into the typical surveillance pipeline, issues in trajectory-based learning methods for activity and behavior understanding revolve around the tracking process itself. The clustering and modeling framework must be robust to imperfect tracking—incomplete and broken trajectories due to occlusion or missing detections—because this will certainly occur in a real monitoring or surveillance setting.

In an effort to limit the effect of imperfect tracking, trajectory samples can be treated independently. In the work by Noceti and Odone,[9] the image space is vector quantized based on the density of track points, resulting in a Voronoi tessellation. Each Voronoi cell represented an alphabet entry, and instead of clustering raw trajectories, they are remapped into alphabet strings. The P-spectrum kernel is used to count common substrings for graph cuts spectral clustering to provide a behavior hierarchy.

3.3.2 *Application-dependent behaviors*

Comparisons have shown that the clustering algorithm itself is not as important as selection of the type of similarity measure since the similarity metric implicitly defines a behavior.[14] However, it is still unclear how to best extract meaningful patterns from a large trajectory dataset. There is inherent ambiguity in how to describe a behavior. For example, at an intersection, should a trajectory of a vehicle that proceeds through on a green phase be thought of differently than one that is forced to stop first due to a red light?

Due to the complex nature of behaviors and difficulty developing a one-size-fits-all approach, hierarchical multilevel learning and analysis frameworks are often proposed. Morris and Trivedi[3] proposed the three-staged hierarchical learning process in Fig. 3. The different levels provide varying resolution when specifying behaviors. The first level considers only the origin-destination information (nodes), while the second level provides spatial distinctions between the nodes. Finally, the spatiotemporal dynamics are encoded with a hidden Markov model (HMM) to probabilistically characterize a behavior.

3.3.3 *Incremental learning and update*

Recent trends in trajectory clustering have focused on incremental learning techniques that enable time-dependent patterns that are able to adapt to new data and changing conditions. These are important for real-world implementation since surveillance systems are required to be operational over very long time periods and old training data may not accurately reflect the current monitoring situation (e.g., construction shutting down a lane and rerouting traffic). The challenge for operation over very long time periods is to ensure that adaption techniques only consider meaningful

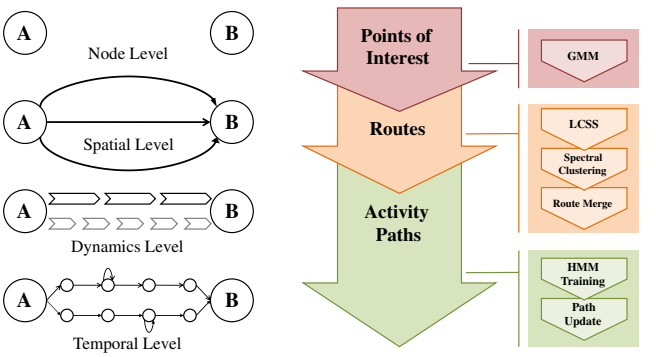


Fig. 3 A three-stage hierarchical learning procedure to model activities at various resolutions.[3] The first level learns points of interests (nodes), second locates spatial routes between nodes through clustering, and the final level probabilistically encodes spatiotemporal dynamics.

changes. Therefore, a balance needs to be maintained between historical prior models and learning from new observables.

The growing HMM is utilized in Ref. 7 in order to have an online adaptive model that changes with conditions rather than being based on a fixed observation time. Voronoi tessellation is used to build up a topological map of spatial states, which are then connected through the growing HMM. Rather than explicitly model trajectory patterns, this work looks at the probability of goals (the eventual endpoint of a trajectory) and gives a distribution of look-ahead states. Interestingly, the patterns can be learned incrementally and in parallel with prediction. Maximum likelihood linear regression (MLLR) for HMMs has been used in parallel with a batch update process to provide adaptability.[3] However, MLLR only adapted existing behavior patterns and did not provide a true incremental learning framework to handle truly time-varying patterns (such as the deletion of a behavior).

3.3.4 *Expensive computational requirements*

Another issue with trajectory clustering methods is the need to compute all pairwise similarities between trajectories. This can be quite computationally expensive depending on the selected similarity metric. When the trajectory database is large, there is significant overhead in terms of storage since a full similarity matrix requires N^2 entries for N training trajectories. This may become too large to have the entirety in memory or be very difficult to work with. Approximation methods like the Nyström method[15] may enable spectral decomposition but may not effectively utilize all available data. However, it should be noted that although clustering and learning may have high computational requirements, online analysis and usage of the learned patterns is not typically complex due to compact modeling.

4 Learning Behavior with Topic Models

The previous section highlighted a number of methods that were able to learn scene context from observation of trajectory data; however, the quality of those methods are highly dependent on robust tracking of vehicles, which is inherently difficult in general due to noise, changing lighting conditions, shadows, and occlusion. In addition, trajectory clustering requires the pairwise computation of similarity between all tracks, which may be expensive computationally and with memory. Given these shortcomings, researchers have been interested in developing methods to learn activities without explicitly tracking objects (or at least without the need for highly accurate trajectories).

Recently, there has been significant research into the use of topic models for behavior analysis. These methods have become quite popular due to their success with natural language processing—e.g., probabilistic latent semantic analysis (pLSA)[21] and latent Dirichlet allocation (LDA).[22] The topic models are able to recognize relationships through the co-occurrence of simple features at different hierarchical levels. Table 2 presents a short listing of representative works that utilize topic models for traffic behavior understanding.

Topic models, as traditionally applied in language processing, view a document as a mixture of various topics and each word in the document is generated from a single topic. The learning goal is to discover the topics of a document given the words. This discovery is achieved by mining a corpus of documents to examine the co-occurrence of words to cluster into the topics. For example, words such as "baggage," "terminal," and "flight" often co-occur in documents and could be clustered into the topic "airport." In the video analogy, a bag of visual words is constructed from motion and a topic refers to the "path."

4.1 *Basic LDA Formulation*

The most commonly used topic model is LDA.[22] The Bayesian model can be efficiently described by its plate notation as given in Fig. 4. The observed variable, shown in gray, is the word given by w_{ji}. The plate K represents the word distribution over topic k given the N_j topics and M documents. z_{ji} is the topic, governed by the topic distribution π_j. LDA has become a popular model because it enforces

Table 2 Representative topic model studies.

Research study	Input data	Classification/inference	Description
Wang et al.[16]	Trajectory	Dynamic dual-HDP	Dynamic dual-HDP nonparametric Bayesian model to automatically model activity categories and semantic regions without specifying the number of topics and with online update of model.
Song et al.[17]	Optical flow	LDA	Two-level LDA topic model learned first for single-agent motion, which is input to second level LDA for multiagent interactions.
Fowlkes et al.[18]	Foreground pixels	DPMM	Fast rank-1 robust PCA used for foreground detection with counts of pixels in blocks used as input for DPMM learning, which enables incremental learning and inference.
Fu et al.[19]	Optical flow	LDA	Sparse topical coding used for efficient learning and representation of the topic model.
Hu et al.[20]	Trajectory	tDPMM	Dirichlet process mixture model is used for unsupervised clustering and modified to handle temporal structure and ordering inherent in a trajectory sequence.

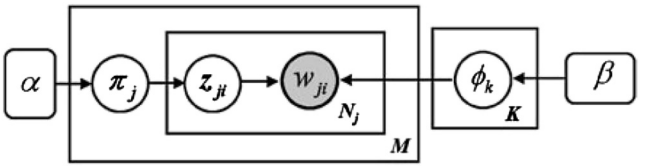


Fig. 4 The basic latent Dirichlet allocation (LDA) topic model commonly used for activity learning.[22]

a Dirichlet prior over the topic distribution and word distribution for improved performance over pLSA.

The basic methodology to learn given the LDA model is as follows: A database of video is collected and segmented into clips that represent documents. A small vocabulary (set of motion vectors) is created based on vector quantization to build a codebook of visual words. Typically, the visual words are constructed by uniformly dividing the two-dimensional image plane into small regular cells and quantizing the motion in each cell into one of four directions (north, south, east, west). A word then is a dimension 3 vector $w = [x, y, dir]$. Using an inference method, such as collapsed Gibbs sampling (Markov chain Monte Carlo algorithm) or variational Bayesian methods,[22,24] collections of words can be combined into topics (a co-occurrence of motion). The learned topics indicate the dominant behavior in a scene. The LDA process is highlighted in Fig. 5(a). Notice that since data are shared between different topics, a behavior does not necessarily have to be from an entry to exit point. A turn and straight through an intersection may have a shared topic (the approach to the intersection) before separating to complete different topics on the way out of the scene. This topic sharing is analogous to the track sharing depicted in Fig. 2(b) and provides efficient utilization of limited training data.

The standard LDA formulation can be used to determine the dominant scene flow as well as the detection of abnormal events.[23] It is popular for application papers because the implementation has been well studied and software packages are freely available online.[25–28]

4.2 *Adaptability and Automatically Generating the Number of Topics*

One of the very nice properties of topic models is their graphical representation, which makes it very easy to augment and extend the basic model for various levels of interpretation and understanding. The basic LDA structure can be augmented with new nodes that account for extra observations and hidden nodes that represent the model hierarchy. The simple topics can be extended by adding distributions over the topics to describe interactions or global behaviors (Fig. 5).

The hierarchical Dirichlet process (HDP)[31] is an extension of LDA to model the Dirichlet admixture nonparametrically, Fig 6(a). The nonparametric prior, parameterized by the concentration parameter and base measure, enables the HDP to automatically select the number of topics based on the training data. The ability to automatically determine the number of topics is incredibly important in practice because this would not be known in a new application scenario.

The HDP model was further extended by Wang et al.[32] as the dual-HDP in order to cluster motion into semantic regions (topics in the basic LDA or HDP formulation) as well as to further cluster the semantic regions into interactions without manual specification. The model, as depicted in Fig. 6(b), contains two hierarchical Dirichlet processes—one modeling the semantic regions and the other modeling the interactions.

Haines and Xiang[33] further extended the dual-HDP with the delta dual-HDP structure shown in Fig. 7. This model was designed for jointly learning both normal and abnormal behavior using weakly supervised training examples. By modeling both normal and abnormal, subtle behaviors can be learned with little training data yet still detect abnormalities as unusual behaviors. Notice the two topics, regular $H_t^{\mathcal{R}}$ and abnormal $H_t^{\mathcal{A}}$, are defined at the top of the structure and control the distribution over the topics with prior S_d.

The basic LDA formulation has been upgraded in a more straightforward staged fashion. A two-level LDA topic model[17] has been used to learn scene behaviors. The first level learned single-agent motion, while the second LDA

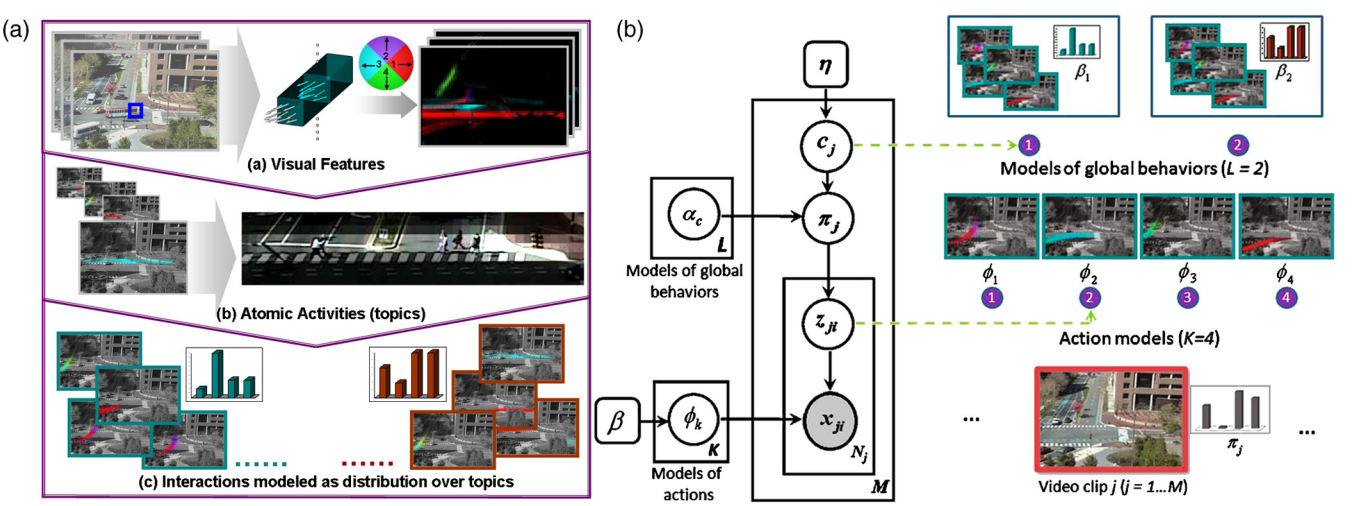


Fig. 5 (a) The basic framework for LDA topic modeling with visual features. The top shows the extraction of visual features indicating position and direction of travel, the middle gives the topics, and the bottom gives higher-level behavior interpretation as distributions over topics.[29] (b) The hierarchical LDA framework for interactions showing connections between indicator variables and behaviors.[30]

level used the first-level output to learn interactions over multiagents. This hierarchy enabled anomaly detection at both levels for every video frame rather than for clips. In a similar fashion, a two-staged cascaded LDA (Cas-LDA) model, shown in Fig. 8, was formulated in Ref. 24. The first stage learns regional behavior and context by applying LDA to all nonoverlapping windows in a video clip. The second stage learns global context over the regional models.

4.3 *Explicitly Modeling Temporal Variations*

Temporal topic models explicitly model the dynamic evolution of behaviors. That is to say, they are able to recognize the relationships and ordering between the scene behaviors. Rather than using a static model for a single clip, the evolution from clip to clip is considered by reasoning over temporal data. This enables learning the "rules of the road," e.g., the signal phase at an intersection controlling the sequence of vehicle movements.

The Markov clustering topic model (MCTM)[34] structure presented in Fig. 9(a) utilized a three-layer latent structure composed of events, actions, and behaviors. The visual events are simple pixel motion, actions are single object activities, and multiobject behaviors are composed of co-occurring activities that are correlated in time. The temporal correlation between behaviors is assumed to vary over time according to an unknown discrete distribution with parameter Ψ (defining a Markov chain). The parameter Ψ is treated as Dirichlet distributed to automatically discover the temporal evolution of behaviors between clips. Therefore, the MCTM is able to recognize the likelihood of transitioning from a behavior in one clip to another behavior in the next, which is encoded in a transition matrix.

Kuettal et al.[35] developed the dependent Dirichlet process-hidden Markov model (DDP-HMM) to extend the MCTM. Rather than a single Markov chain governing the entire scene, a number of HMMs are allowed to exist enabling analysis of more complex scenes. Therefore, various time-ordered dependencies can be discovered from video clips, resulting in a much more complicated state transition matrix than for MCTM. A comparison between the DDP-HMM and MCTM found that the DDP-HMM was able to find extra alternative traffic light cycles.

A mixed event relationship model (MERM) has relaxed the first-order Markov property to explicitly model the time-lag between dependent activities.[36] The MERM models both global rules, motion during intersection light phases, as well as local rules, the transitions from an activity to another with an associated time lag. The MERM relied on a sparse activity representation known as probabilistic latent sequential motifs (PLSM)[37] rather than visual motion words. PLSM provides dominant activities of the scene as well as the starting times to build the binary occurrence matrix input to MERM.

4.4 *Incremental Learning and Update*

The previous sections highlighted the ability of topic models to conveniently learn vehicular behavior patterns; however, the aforementioned techniques all learn and fix their outputs.

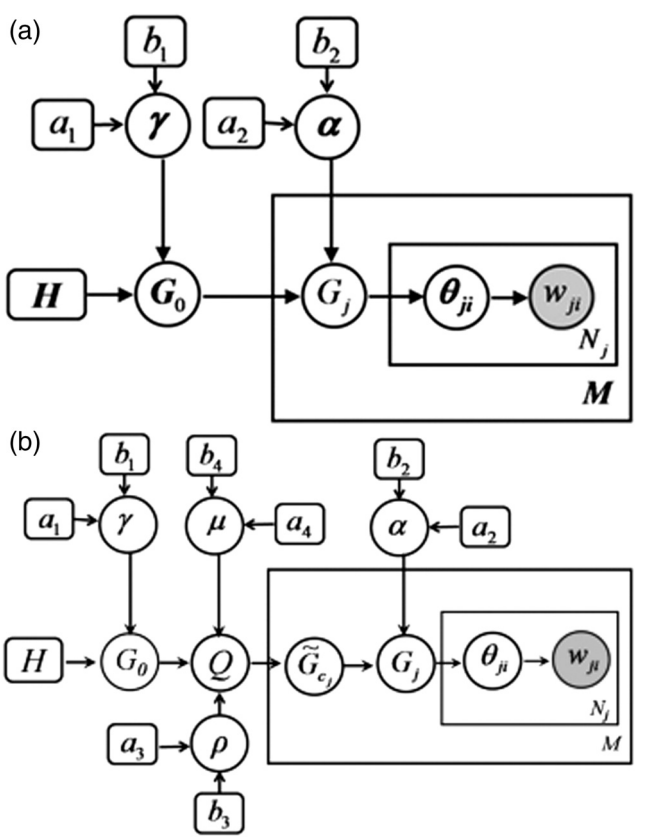


Fig. 6 LDA upgrades to automatically determine the number of topics with (a) hierarchical Dirichlet process (HDP) and the number of interactions with the (b) dual-HDP.[32]

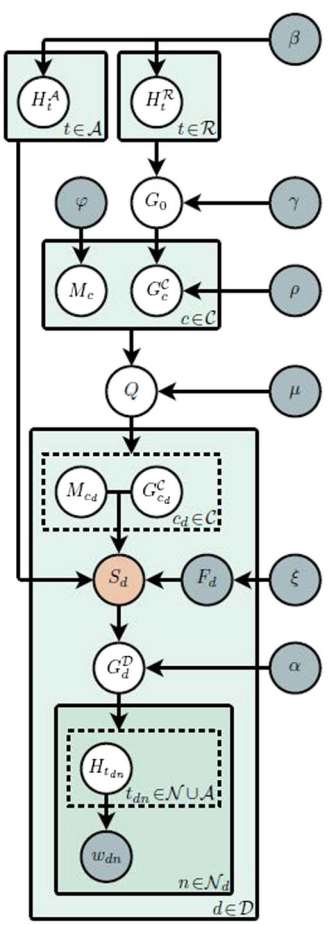


Fig. 7 Delta dual-HDP[33] used to automatically model both the normal and abnormal behaviors.

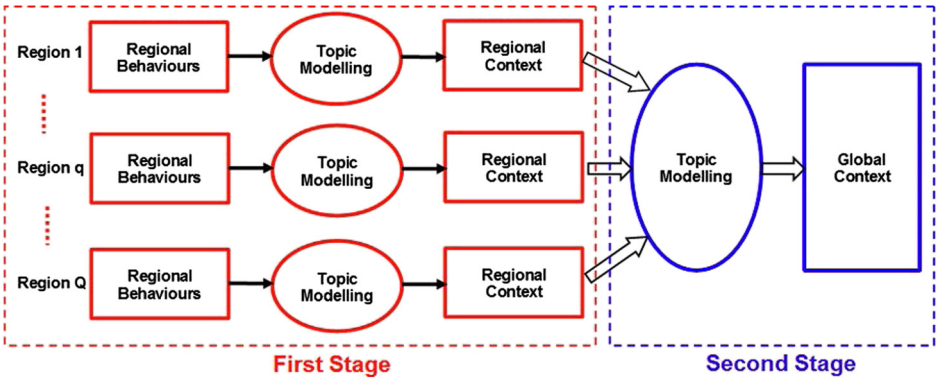


Fig. 8 Cascade LDA topic modeling of regions followed by a second LDA stage for global context from Ref. 24.

The learned patterns can then be used to classify behaviors and detect abnormalities in an online fashion but have no mechanism for updating models with passing time. Active surveillance requires observations over long periods of time where the patterns of activity may change. For example, a new activity may be observed, an old one disappear, or the sense of what constitutes an abnormality may evolve (e.g., the first appearance may be abnormal but after more examples it should become typical). These changing conditions necessitate incremental learning and update methods.

Wang et al. extended their dual-HDP model for incremental learning into the dynamic dual-HDP. As shown in Fig. 10, the dynamic dual-HDP successively learns the topic models in time slices (a fixed amount of time). All the information (topics, etc.) learned in a particular time slice $t-1$ can be used as a prior for prediction of the topics in the next time slice t. Therefore, the model is able to naturally evolve over time in a smooth fashion.

4.5 *Discussion of Topic Models*

Application of topic models is appealing for computer vision researchers because they provide a nice computational framework for machine learning. Specifically, they successfully utilize the simple bag-of-word representations to learn models in an unsupervised fashion without manual labeling of training data. They enable efficient utilization of data and avoid overfitting by sharing features and training data between actions. Most topic models for behavior understanding are hierarchical Bayesian models that allow joint modeling at different levels of complexity and enable easy extension and introduction of contextual knowledge as priors. Finally, they can use Dirichlet processes (DP) as priors to automatically learn classes from data and avoid manual specification.

These topic models are extremely useful in learning in scenarios that are difficult to process with conventional detection and tracking. Readers interested in topic models for action recognition are directed to the recent review by Wang[30] for a more detailed treatment of the subject.

While decoupling modeling from explicit tracking helps improve learning in difficult scenarios with crowds and large amounts of occlusion, the resulting generative models are based on co-occurring motion words. Due to the bag-of-words formulation, generally, temporal ordering is discarded,

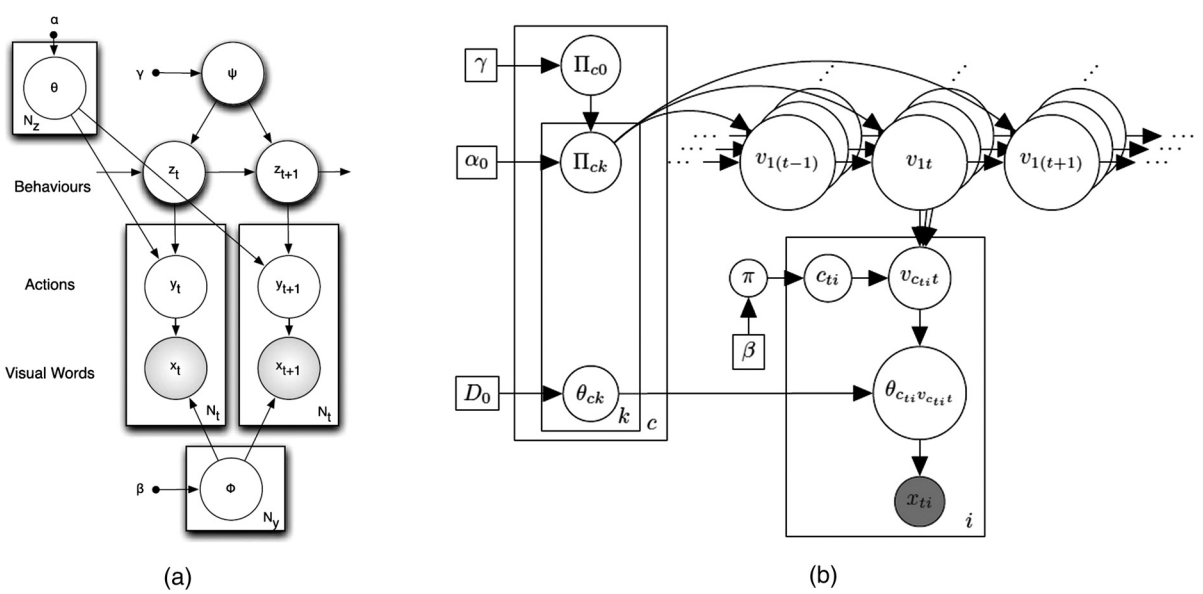


Fig. 9 Temporal topic model structures enable learning of the temporal ordering and transitions between behaviors (the "rules of the road"). (a) Markov clustering topic model (MCTM),[34] where behaviors are allowed to change between time slices based on a discrete distribution Ψ. (b) The dependent Dirichlet process-hidden Markov model (DDP-HMM)[35] extension of the MCTM to have multiple hidden Markov models for more complex scene analysis.

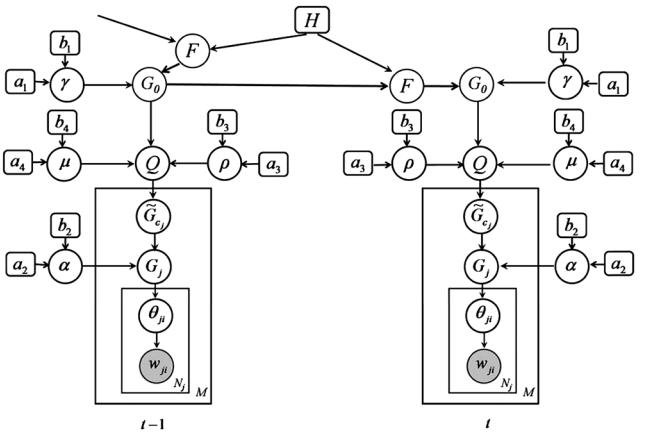


Fig. 10 The dynamic dual-HDP[16] uses semantic regions in time $t-1$ as a prior for prediction of semantic regions in time slice t enabling online incremental updating of behavior models.

which breaks the typical notion of a trajectory sequence. In addition, without explicit tracking, there may be ambiguity when making predictions about future events. For example, when an abnormality is detected, it may not be clear to which entity it is attributed. This is most easily recognized when considering frameworks that provide a clip level description (e.g., normal/abnormal clip). It is important from the monitoring and surveillance standpoint to be able to accurately gauge exactly when, where, and with whom an event occurs.

One of the appeals of topic models is their ability to automatically determine the number of behavior topics present in a scene. However, this does not mean that nonparametric topic modeling techniques are without parameters. They require the setting of hyperparameters, e.g., the Gamma priors, which may affect the modeling performance. In addition, there are hidden implicit parameters inherent in the learning process such as the burning time of the Gibbs sampling, which must also be set manually.

4.5.1 *Trajectory-topic modeling*

The topic modeling approaches can also be applied to trajectories. Rather than considering a video clip as the document, the trajectory itself is a document. The words are still the observed motion, only this time they come from the tracking process. The interpretation of topics (as well as other hierarchical distributions built upon the topics) remains the same.[30] The advantage of the fusion of trajectories and topic modeling is the ability to target individual behavior patterns rather than a full scene description of behavior.

Due to the loose constraints on pedestrian movements, Kooij et al.[39] developed a mixture of switching linear dynamic systems to discover typical actions and their temporal relations at the object level. Trajectory data were obtained from a person tracker and clustered into behaviors. These behaviors were defined by transition probabilities between actions, which represent spatial location and low-level motion dynamics. Therefore, tracks are segmented into sequences of actions and they are jointly clustered into behavior classes. The low-level actions and temporal order within high-level behaviors are inferred directly from trajectories. In addition, continuous motion distributions are utilized for better identification of dynamic variations that can be lost through quantization.

Very recent work by Hu et al.[20] used the Dirichlet process mixture model (DPMM) to cluster, model, and retrieve trajectories in an incremental fashion. Further, the time-varying information contained in a trajectory is modeled using the time-sensitive DPMM (tDPMM) over subtrajectories. The dynamic dual-HDP has also been applied to trajectories[16] to learn behaviors as distributions over shared semantic regions (subpatterns) with incremental updating. This work showed results from both vision-based trajectories as well as radar tracks, demonstrating the general applicability of trajectory learning.

4.5.2 *Computational efficiency*

The push toward incremental learning and updates is extremely important for widespread adaption of behavior understanding systems. In order for these systems to be useful in monitoring situations and TMCs, these systems must operate over extremely long periods with little supervision. It will be important in the future to develop computational efficient algorithms to leverage massive historical datasets and elegantly update new observations. A joint multinomial + Gaussian DPM framework has recently been proposed to trade-off computational complexity of the Bayesian topic model processes with scalability for large

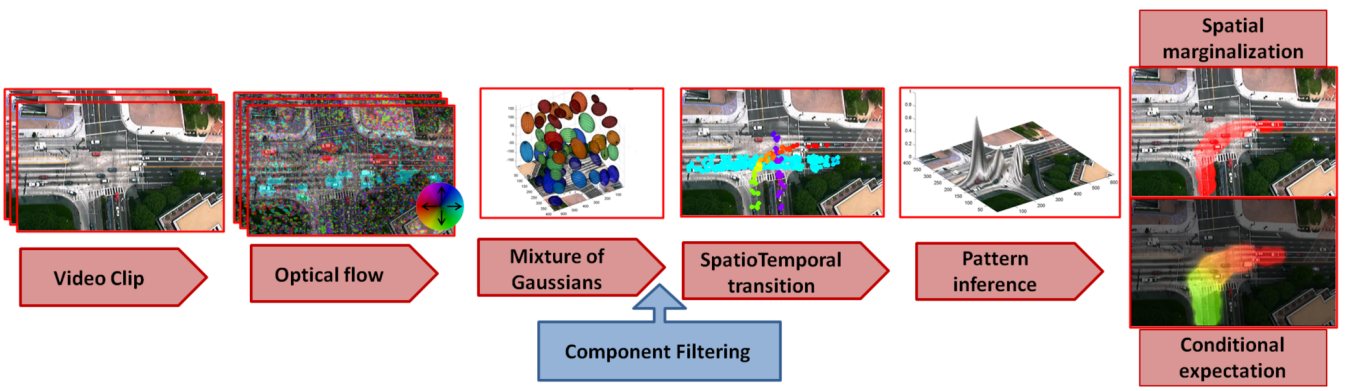


Fig. 11 Process diagram for approach of Saleemi et al.:[38] Grouping of frames into video clips, optical flow computation, Gaussian mixture model learning by k-means, filtering of noisy Gaussian components, intercomponent spatiotemporal transition computation for instance learning, pattern inference using Kullback-Leibler (KL)-divergence, pattern representation as spatial marginal density, and computation of conditional expected value of optical flow given pixel location.

datasets.[18] Fu et al.,[19] in contrast, use sparse topical coding to represent clips with a sparse set of motion patterns in an LDA framework for efficiency.

4.5.3 *Topic model alternatives*

Despite their recent popularity and success, topic models are not the only way to decouple tracking from the learning process. The same visual-bag-of-motion-words concept can be applied to video clips with other learning frameworks. In Ref. 38, dense optical flow vectors are computed as motion descriptors to avoid tracking. Video is divided into 1-s clips, and all optical flow vectors are clustered using k-means into a mixture of Gaussian distributions (Fig. 11). Mixture components were treated as nodes in a graph connected through time based on a reachability criterion that relied on a constant motion model between proximal clips. The resulting Gaussian chains defined the distribution of a motion pattern and were used for inference based on Kullback-Leibler (KL) divergence.

Tracking and learning were also decoupled without the use of topic models by Zen et al.[40,41] In this work, activity patterns are extracted through matrix factorization based on the Earth mover's distance. The authors utilize short trajectory snippets (to avoid full tracking) and background subtraction identification of static pixels to explicitly address noise and uncertainty in visual information.

5 Evaluation and Benchmarking

The increased emphasis in trajectory pattern analysis for driving behavior understanding has been fueled by the emergence of readily available datasets to provide benchmarking. This sharing of benchmark data is incredibly important for improving research results because it provides fair comparison and lowers the barrier of entry into the area. Further details on available datasets are presented in the following section along with common evaluation criteria for behavior analysis.

5.1 *Datasets*

Table 3 provides a brief description and sample image of the monitoring scene for the most popular public datasets. Detailed descriptions for each dataset are provided below.

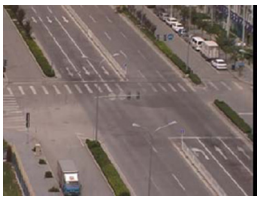
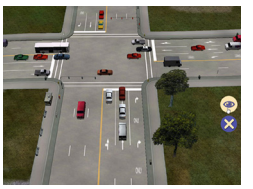
5.1.1 *MIT traffic dataset*

The Massachusetts Institute of Technology (MIT) traffic dataset[32] is for research on activity analysis and crowded scenes and has become the most widely used dataset. It contains 90 min of traffic video recorded by a stationary camera divided into 20 clips. The clips contain both vehicle traffic at a four-way signalized intersection and pedestrians on three zebra crossings. Only raw video is provided, so motion/trajectories must be extracted from the video. However, ground truth does exist for a subset of frames for pedestrian detection,[45] and there are true abnormalities present in the data.

5.1.2 *MIT trajectory dataset*

This MIT trajectory dataset[42] contains 40,453 tracks extracted from a single camera overlooking a parking lot over a week. There are a few vehicle enter/exit behaviors with the majority being pedestrian activities. The dataset

Table 3 Popular activity learning and behavior understanding traffic datasets.

	**MIT traffic dataset**[32] http://www.ee.cuhk.edu.hk/~xgwang/MITtraffic.html 90 min of video for activity analysis in crowded scenes that contain both vehicle and pedestrian traffic.
	**MIT trajectory dataset**[42] http://www.ee.cuhk.edu.hk/~xgwang/MITtrajsingle.html Over 40,000 tracks from a camera overlooking a parking lot over a week. Designed to examine computational efficiency.
	**Next-generation simulation**[43] http://ngsim-community.org 30 min of multicamera overhead arterial video of capturing intersections. Dataset contains raw video, detailed vehicle trajectories, and supporting behavioral data.
	**CSBR intersection traffic dataset**[8] May be available by special request. Complex view of an intersection with a number of possible entry and exit combinations.
	**CVRR trajectory analysis datasets**[3] http://http://cvrr.ucsd.edu/datasets Includes two highway datasets—one simulated and the other real—and a simulated intersection. The ground truth includes frame-level predictions and unusual actions.
	**i-Lids–advanced video and signal-based surveillance 2007 vehicle detection challenge**[44] http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html Three videos on a roadway in the United Kingdom with varying degrees of tracking difficulty with associated ground truth data.
	**QMUL junction dataset**[34] http://www.eecs.qmul.ac.uk/~jianli/Dataset_List.html 1 h (90,000 frames) of busy traffic video collected at a junction and with ground truth available.

was collected to examine computational efficiency in terms of space requirements, complexity of algorithms, and adaptability.

5.1.3 *NGSIM*

The next-generation simulation (NGSIM) program[43] was developed by Federal Highway Administration (FHWA) to develop a core of open behavioral models in support of microscopic modeling and traffic simulation. It included supporting validation data that have been used more recently for learning traffic patterns. The overhead intersection cameras from the Lankershim and Peachtree sets are the most commonly used. They each provide 30 min of arterial video data, detailed vehicle trajectories, as well as supporting behavioral data.

5.1.4 *CBSR intersection traffic dataset*

The Center for Biometrics and Security Research (CBSR) intersection traffic dataset[8] has been used extensively for vehicle detection and tracking. It was collected by researchers at the CBSR at the Institute of Automation, Chinese Academy of Sciences. The dataset provides a single view of a complex intersection that contains a number of possible entry and exit combinations. Although the dataset has been cited many times, it was not readily available online but seems to require special access rights; therefore, the type of data (video, trajectories, abnormalities) provided could not be confirmed.

5.1.5 *CVRR trajectory analysis dataset*

The Computer Vision and Robotics Research (CVRR) trajectory analysis datasets[3] were designed specifically for evaluation of trajectory clustering and analysis techniques. The datasets consist of a simulated intersection, an 8 lane highway, as well as video from a laboratory obtained with an omni-directional camera. The included annotations provide classification and abnormality detection as well as online prediction and localized unusual action detection for every frame.

5.1.6 *i-Lids–AVSS 2007 vehicle detection challenge*

A subset of the i-Lids dataset was used for the IEEE advanced video and signal-based surveillance (AVSS) detection and tracking algorithm challenge in 2007.[44] The Task 2—Parked vehicle challenge—video consists of three videos of a roadway in the United Kingdom of various tracking difficulty (based on density of vehicles) with ground truth data.

5.1.7 *QMUL junction dataset*

The Queen Mary University of London (QMUL) junction dataset[34] is a new dataset specifically for activity analysis and behavior understanding. This challenging dataset contains 1 h (90,000 frames) of busy traffic video collected at and 25 fps. This is quickly becoming a favorite dataset for topic modeling.

5.1.8 *Idiap traffic junction dataset:*

The Idiap traffic junction dataset[46] provides a view of a junction controlled by traffic lights and contains activities for both cars and pedestrians. The video contains multiple instances of rare or unusual events, which are provided as an abnormality annotation file. The dataset was used for topic model scene analysis and provides 44 min of video at resolution and 25 fps.

The most popular datasets for trajectory learning (specifically topic modeling) is the MIT traffic dataset and QMUL junction dataset. The NGSIM set has also been used often, thanks to the availability of trajectory, as well as ground truth behavior data. Note that only a few of the datasets are specifically for activity analysis and behavior understanding.[3,32,34,42] Most of the datasets have been used for this purpose after the fact since they are available for comparison purposes. However, even for the behavior-specific datasets, there is a shortcoming in terms of ground truth annotations of the "true" activity and abnormalities for performance evaluation. Most of these sets are just used to provide qualitative learning effectiveness rather than quantitative performance on clustering accuracy, prediction, and abnormality detection.

5.2 *Evaluation Criteria*

One of the major issues for activity analysis and behavior understanding research is consistent definitions and notions of performance. There has yet to be a consensus on the most appropriate ways to characterize performance. In fact, few works actually perform quantitative evaluation. This lack of quantitative comparison is likely due to the ambiguous nature of behavior. However, a number of methods have been used to concretely characterize performance for specific tasks in behavior understanding.

5.2.1 *Correct clustering rate*

The correct clustering rate (CCR) provides a measure of how well labels match between clustering and ground truth and has been used to verify trajectory clustering algorithms.[14,20,47] In order to evaluate a clustering result, a one-to-one mapping between ground truth labels and those returned by the clustering algorithm is found. The mapping is typically recast as the minimization of the number of mismatched labels and solved greedily using the Hungarian algorithm.[48] Given the mapping between ground truth and cluster labels, the CCR is computed as

$$\mathrm{CCR} = \frac{1}{N}\sum_{c=1}^{K} p_c, \tag{1}$$

where N is the total number of trajectories and p_c denotes the number of trajectories correctly matched to the c'th ground truth cluster label.

5.2.2 *Completeness and correctness*

In spirit similar to CCR, completeness and correctness has been used to evaluate clustering performance.[16,49] Rather than using a single metric, the pair is used to provide more detailed characterization. Correctness is a measure of how well different ground truth trajectories are separated, while completeness is a measure of how many similar trajectories are combined in a cluster. In practice, there is a tradeoff between these two aspects.

The completeness and correctness metrics are appealing because ground truth labeling can be obtained by pairwise comparisons of trajectories rather than needing to estimate (or know) the total number of clusters in the dataset. This becomes particularly important when datasets are very large and activities/behaviors are complicated since it would require expert or very-well trained annotators rather than mechanical turk.

5.2.3 *Receiver operating characteristic curve*

The receiver operating characteristic (ROC) curve has been used extensively in the detection and classification communities to provide a parameterized performance curve. The ROC curve has been adopted for abnormality detection performance reporting by casting the problem in a normal/abnormal detection framework.[3,24,33,34,50] The ROC curve is plotted by varying a sensitivity threshold and calculating the number of correctly classified versus incorrectly classified results. Example abnormality ROC curves are presented in Fig. 12. Notice that these correspond to different inputs. In (a), the abnormal classification is for every frame for each trajectory (online evaluation), while in (b), a short video clip is considered.

6 Applications

The following section provides a brief look at useful transportation applications highlighting the generality of the models learning from trajectory clustering and topic models. These are only the first applications that may come to mind and with maturation of the learning technology, exciting new application areas will emerge.

6.1 *Traffic Scene Characterization*

The learned traffic patterns, both from trajectory learning and topic modeling, provide an expression of the scene based on observables rather than manual specification. This characterizes how the particular roadway is typically used (which is often how it was designed to be used). In an unsupervised fashion, these techniques provide a method to learn the rules of the road and the geometric configuration. This gives the allowable maneuvers at an intersection, the number of lanes, etc.

This characterization is useful in the traditional traffic engineering sense when utilizing trajectory techniques. Rather than using intrusive technology, such as pneumatic tubes or inductive loops, or hiring a person to sit at a street corner, the counts can be obtained with available traffic cameras.[51] These counts could indicate the flow of vehicles on the highway, the turn counts at an intersection, etc. Also, with camera calibration it is possible to extract measures of occupancy, density, and speed profiles. Also, since cameras are wide area sensors, they can also provide statistics that transportation engineers in operations and planning would like but do not have readily available such as queue length and wait time.

More recently, more complex utilization of trajectory and imagery data has provided real-time estimations of vehicular emissions in the CalSentry system.[52] Dynamic information obtained through calibrated tracking was tied with vehicle-specific emission models to characterize the energy/emissions of all roadway vehicles. The cameras were used not only for tracking but also to classify vehicles based on their appearance and were tied together with behavioral monitoring to automatically annotate the lanes and direction of travel for each vehicle.

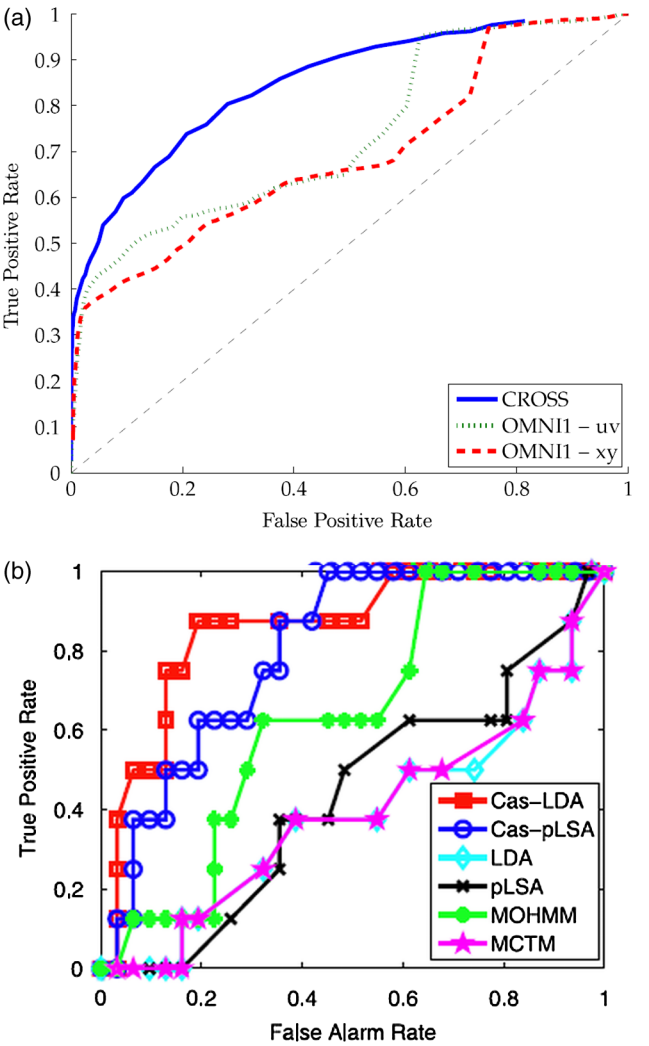


Fig. 12 Abnormality/unusual action receiver operating characteristic (ROC) curves. (a) ROC generated in online fashion for every sample of every trajectory.[3] (b) ROC generated based on abnormality classification for short video clips.[24]

6.2 *Conflict Analysis*

Perhaps the most natural extension to traditional traffic analysis one can envision with the use of video technology is advanced conflict analysis. Rather than waiting for rare collision events (and annual reports that may lack critical information), with traffic cameras, critical locations can be monitored in real time. Safety can then be analyzed in a proactive manner based on the interactions between vehicles. A hierarchy can be constructed, as shown in Fig. 13, and utilized to classify interactions.[53] At the top of the pyramid are collisions, which are rare, and the bottom has safe interactions, which make up the majority of the time. In the middle are conflicts, the situation when two vehicles could collide without intervention, which can be used as a surrogate safety measure.[54]

Surrogate safety analysis is only possible with accurate prediction of future events. However, this prediction is

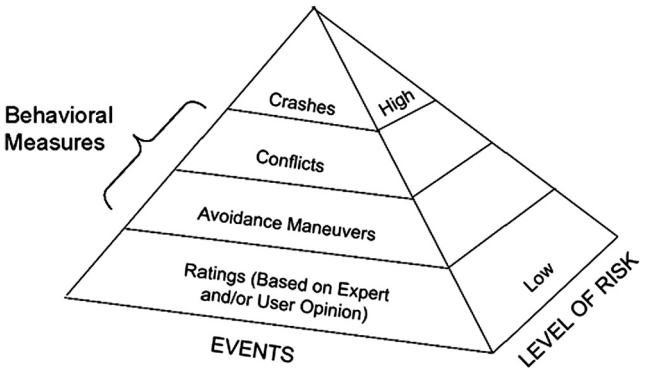


Fig. 13 Hierarchical order of safety measures.[53] While crashes are the most objective indicator of dangerous behaviors, they are rare, so surrogate safety measures can be collected based on conflicts and avoidance maneuvers, which relate to interactions between vehicles.

difficult because of the complex dynamics and control with vehicular traffic; a driver will change the kinematic state based on an internal plan or intentions. Numerical techniques, such as Markov chain Monte Carlo or multiple hypothesis tracking, can be used to enumerate many possible future outcomes as well as to attempt to guess the driver intention; however, these are computationally expensive.[55] Instead of complex motion models, patterns learned from motion or trajectory clustering can be used as priors to generate the likely future trajectories. The priors would then represent the possible intentions of a driver during tracking. This removes the need for high-resolution (sensitive) measurement of vehicle dynamic parameters such as acceleration, yaw angle, yaw angle rate, etc., and instead use historical observations to cue future trajectory of a vehicle. Conflict analysis requires methods to accurately assess future states for each individual vehicle, making it better suited for trajectory-based algorithms.

Given scene context as exemplified by the learned behavioral patterns, the next challenge for accurate behavior prediction would be to understand driver intentions as influenced by other road users.[56] This study of interactions between vehicles will be essential for real safety analysis since travel is not performed in a vacuum; the actions of other road users affect a driver's response (Fig. 14). The challenge is to understand what are the most important mental models in play during driving and how those change due to external influences from the environment (e.g., highway or intersection), driver perception and comprehension of the driving context, and the myriad of distractors faced while driving.

6.3 *Motion Database Query*

Another interesting application of trajectory characterization comes from database retrieval. Given the large amounts of visual data that are collected from traffic cameras, it will become important to succinctly index activities and events for further analysis. In particular, the models learned from trajectory learning or topic modeling can be used for pattern matching and database query.[10,20] This functionality enables a traffic engineer to find occurrences of a particular behavior, e.g., illegal U-turn or jaywalking, from a large trajectory database for closer inspection.

6.4 *Abnormality Detection*

Incident detection is one of the most challenging and useful applications in transportation. It provides a higher level of scene understanding than more basic vehicle counting. Incident detection is often treated as a recognition problem to find crashes, illegally stopped vehicles, illegal lane changing, etc.[1] These incidents can be detected as vehicles behaving abnormally since by its very definition, an incident is an unusual event. By using either trajectory learning or topic models, the scene description that is built provides not only a way to characterize traffic but also a model of typical behaviors. Motion that does not fit into these learned normative models are then classified as abnormalities.

While both trajectory learning and topic models provide means to detect abnormalities, they differ in the type of abnormalities and the localization ability. Since trajectory learning inherently analyzes individuals, it is able to localize unusual events precisely. In contrast, basic topic models provide clip classification. A clip is denoted as unusual or not without explicit localization. However, they naturally handle multivehicle situations, such as the flow in different directions of a controlled intersection.

A number of methods have been proposed to improve abnormality detection because it is one of the powerful analysis tools provided by unsupervised learning techniques. Trajectories can be analyzed in concert as highlighted above for conflict analysis.[54,55] And many augmented LDA topic structures allow for localization of anomalous motion words.

Recent work has looked to improve anomaly detection by explicitly modeling these rare events.[33] Loy et al.[57] use a cascade of dynamic Bayesian networks (CasDBNs) to provide different DBNs with sensitivity toward different types of anomalies. Active learning techniques have also been utilized to analyze live video streams and provide human direction in order to accurately identify the rare abnormality events.[58]

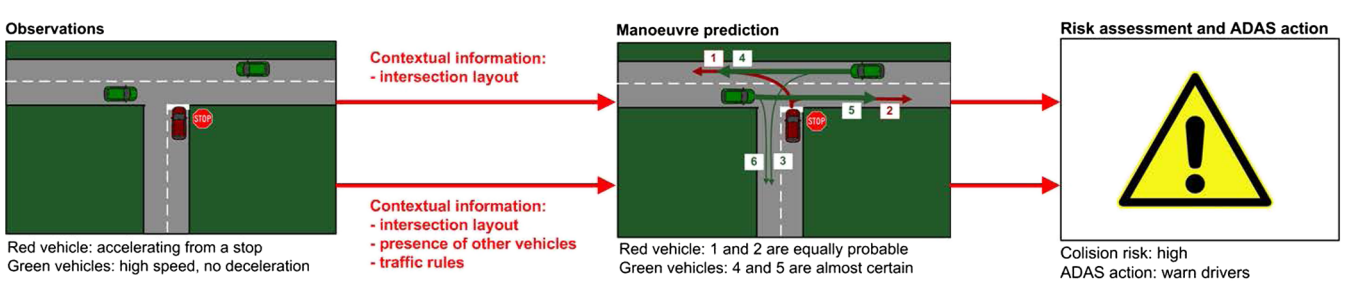


Fig. 14 Scene context is enabled with trajectory/motion clustering and provides the necessary priors for assessing critical situations without relying solely on the dynamics and kinematics of the vehicle.[56]

7 Discussion and Future Directions

Behavior analysis has become a popular field within computer vision. Much of the recent work has focused on the generative topic models due to their probabilistic description of data and interesting properties from the machine learning perspectives. But, it is important to recognize the differences between trajectory learning and topic models to understand what applications are each best suited.

Trajectory clustering fits seamlessly into a traditional surveillance setting where each object is detected and tracked. Since each vehicle is treated separately, it is possible to make direct inference of future behavior of every scene participant and to locate abnormalities precisely. Most importantly for surveillance, after learning, all analysis can occur in real time. However, these algorithms are sensitive to tracking errors, such as occlusion, which cause broken tracks, as would be found in complex and crowded scenes. In addition, these methods do not intrinsically represent multivehicle behaviors.

Topic models are more robust to occlusion since they examine the co-occurrence of motion in a short video clip. Therefore, they naturally represent multiobject behaviors and can detect anomalies defined by unusual co-occurrence of actions. However, the basic variants do not localize anomalies spatially and most do not detect anomalies if most of the simultaneous activities are typical. It is also important to select the appropriate clip size, otherwise abnormalities will not be localized temporally.

The challenges for traffic behavior understanding are how to provide the nuanced behavior specification with multiple participants at different spatial and temporal resolutions while still enabling individual counts since traffic engineers require them.

Also, new insight is required in how to handle massive data—both in terms of the number of cameras as well as the continuous stream. For reliable usage, it will be essential for these systems to accurately describe the traffic situation as the network itself changes. It is important to be able to update models effectively over time (incremental learning); however there is the problem of imbalanced data. There is significantly more typical data than unusual, which results in the difficult task of modeling the tails of a distribution. Recent stream-based learning approaches have attempted to model the tails of distributions (abnormalities) using an active learning framework.[58,59] In complex scenes with many typical activities occurring, an abnormality may be difficult to discover (unusual events overlapping with normal). The active learning framework queries a human to provide labels for these difficult situations but it is unclear if this is scalable.

In addition, the scope and view of a road behavior will change with access to larger scale data from global positioning system (GPS). Exciting new work is already underway using floating car data to learn about behaviors. The prevalence of GPS-enabled devices, both in phones and navigation systems, and public willingness to share this information offers massive data to be mined. As is common now, this data can be used to gain insight into the traffic network in terms of speed and congestion. Furthermore, it also provides rich behavioral information. This data could be used for personalized reports of traffic on your favorite routes or directed advertisements (information). But, it can also be used in aggregate to learn trends—the most popular routes between destinations, for crime prevention,[60,61] etc.

8 Concluding Remarks

This manuscript has reviewed recent progress in understanding traffic behavior with infrastructure-based video sensing. The focus of the report was on unsupervised techniques for behavior analysis, which are generally applicable to many traffic surveillance situations. Trajectory clustering and topic models have emerged as the two most popular methods for unsupervised learning and provide the context required for behavior understanding. In addition, a detailed summary of publicly available video and trajectory datasets for behavior analysis are presented. These benchmark datasets will be extremely important for progressing the field. Despite the great success in recent years, there are many open research questions that require attention before these techniques can be widely adopted and provide useful analysis.

Acknowledgments

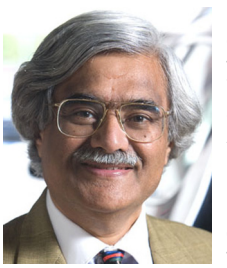
Support for this study was provided by the UC Discovery Grant program. The authors would like to thank the reviewers for their useful comments and members of the CVRR laboratory for their support.

References

1. B. Tian et al., "Video processing techniques for traffic flow monitoring: a survey," in *2011 14th Int. IEEE Conf. on Intelligent Transportation Systems*, pp. 1103–1108, Washington DC (2011).
2. N. Buch, S. A. Velastin, and J. Orwell, "A review of computer vision techniques for the analysis of urban traffic," *IEEE Trans. Intell. Transp. Syst.* **12**(3), 920–939 (2011).
3. B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: unsupervised, multilevel, and long-term adaptive approach," *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(11), 2287–2301 (2011).
4. B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Trans. Circuits Syst. Video Technol.* **18**(8), 1114–1127 (2008).
5. D. Makris and T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *IEEE Trans. Syst., Man, Cybern. B.* **35**(3), 397–408 (2005).
6. C. Piciarelli and G. L. Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recogn. Lett.* **27**(15), 1835–1842 (2006).
7. D. Vasquez, T. Fraichard, and C. Laugier, "Incremental learning of statistical motion patterns with growing hidden Markov models," *IEEE Trans. Intell. Transp. Syst.* **10**(3), 403–416 (2009).
8. W. Hu et al., "A system for learning statistical motion patterns," *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(9), 1450–1464 (2006).
9. N. Noceti and F. Odone, "Learning common behaviors from large sets of unlabeled temporal series," *Image Vis. Comput.* **30**(11), 875–895 (2012).
10. W. Hu et al., "Semantic-based surveillance video retrieval," *IEEE Trans. Image Process.* **16**(4), 1168–1181 (2007).
11. L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, New Jersey (1993).
12. S. Atev, O. Masoud, and N. Papanikolopoulos, "Learning traffic patterns at intersections by spectral clustering of motion trajectories," in *IEEE Conf. Intelligent Robots and Systems*, Beijing, China, pp. 4851–4856 (2006).
13. I. N. Junejo, O. Javed, and M. Shah, "Multi feature path modeling for video surveillance," in *Int. Conf. on Pattern Recognition*, Cambridge, England, UK, pp. 716–719 (2004).
14. B. Morris and M. Trivedi, "Learning trajectory patterns by clustering: experimental studies and comparative evaluation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Miami, Florida, pp. 312–319 (2009).
15. C. Fowlkes et al., "Spectral grouping using the nystrom method," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(2), 214–225 (2004).
16. X. Wang et al., "Trajectory analysis and semantic region modeling using nonparametric hierarchical Bayesian models," *Int. J. Comput. Vis.* **95**(3), 287–312 (2011).
17. L. Song et al., "Understanding dynamic scenes by hierarchical motion pattern mining," in *IEEE Int. Conf. on Multimedia and Expo*, Barcelona, Spain, pp. 1–6 (2011).

18. S. Rana et al., "Large-scale statistical modeling of motion patterns: a Bayesian nonparametric approach," in *Proc. of the Eighth Indian Conf. on Computer Vision, Graphics and Image Processing*, Mumbai, India, pp. 7:1–7:8 (2012).
19. W. Fu et al., "Learning semantic motion patterns for dynamic scenes by improved sparse topical coding," in *IEEE Int. Conf. on Multimedia and Expo*, Melbourne, Australia, pp. 296–301 (2012).
20. W. Hu et al., "An incremental DPMM-based method for trajectory clustering, modeling, and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(5), 1051–1065 (2013).
21. T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of the Fifteenth Annual Conf. on Uncertainty in Artificial Intelligence*, pp. 289–296, Morgan Kaufmann, San Francisco, California (1999).
22. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
23. S. Kwak and H. Byun, "Detection of dominant flow and abnormal events in surveillance video," *Opt. Eng.* **50**(2), 027202 (2011).
24. J. Li, S. Gong, and T. Xiang, "Learning behavioural context," *Int. J. Comput. Vis.* **97**(3), 276–304 (2012).
25. David M. Blei, http://www.cs.princeton.edu/~blei/ (2013).
26. D. Mochihashi, "LDA, a latent dirichlet allocation package," http://www.cs.princeton.edu/~blei/lda-c/index.html (19 August 2013).
27. A. K. McCallum, "Mallet: a machine learning for language toolkit," http://mallet.cs.umass.edu/ (19 August 2013).
28. M. Steyvers and T. Griffiths, "Matlab topic modeling toolbox," http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm (19 August 2013).
29. X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception by hierarchical Bayesian models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, pp. 1–8 (2007).
30. X. Wang, "Action recognition using topic models," in *Visual Analysis of Humans*, T. B. Moeslund et al., Eds., pp. 311–332, Springer, London (2011).
31. Y. W. Teh et al., "Hierarchical Dirichlet processes," *J. Am. Stat. Assoc.* **101**, 1566–1581 (2006).
32. X. Wang, X. Ma, and W. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(3), 539–555 (2009).
33. T. Haines and T. Xiang, "Delta-dual hierarchical Dirichlet processes: a pragmatic abnormal behaviour detector," in *Proc. IEEE Int. Conf. Computer Vision*, Barcelona, Spain, pp. 2198–2205 (2011).
34. T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *Int. J. Comput. Vis.* **98**(3), 303–323 (2012).
35. D. Kuettel et al., "What's going on? Discovering spatio-temporal dependencies in dynamic scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, California, pp. 1951–1958 (2010).
36. J. Varadarajan, R. Emonet, and J. Odobez, "Bridging the past, present and future: modeling scene activities from event relationships and global rules," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, Rhode Island, pp. 2096–2103 (2012).
37. J. Varadarajan, R. Emonet, and J.-M. Odobez, "A sequential topic model for mining recurrent activities from long term video logs," *Int. J. Comput. Vis.* **103**(1), 100–126 (2012).
38. I. Saleemi, L. Hartung, and M. Shah, "Scene understanding by statistical modeling of motion patterns," in *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, California, pp. 2069–2076 (2010).
39. J. F. P. Kooij, G. Englebienne, and D. M. Gavrila, "A non-parametric hierarchical model to discover behavior dynamics from tracks," in *Proc. of the 12th European Conf. on Computer Vision*, Volume Part VI, pp. 270–283, Springer-Verlag, Berlin, Heidelberg (2012).
40. G. Zen and E. Ricci, "Earth mover's prototypes: a convex learning approach for discovering activity patterns in dynamic scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, Colorado, pp. 3225–3232 (2011).
41. G. Zen, E. Ricci, and N. Sebe, "Exploiting sparse representations for robust analysis of noisy complex video scenes," in *Proc. of the 12th European Conf. on Computer Vision*, Volume Part VI, pp. 199–213, Springer-Verlag, Berlin, Heidelberg (2012).
42. X. Wang et al., "Trajectory analysis and semantic region modeling using a nonparametric Bayesian model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Anchorage, Alaska, pp. 1–8 (2008).
43. NGSIM community, "US DOT's Federal Highway Administration," http://ngsim-community.org/ (19 August 2013).
44. P. Hosmer, "i-Lids dataset for AVSS 2007," (2007), http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html (19 August 2013).
45. M. Wang and X. Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Colorado Springs, Colorado, pp. 3401–3408 (2011).
46. J. Varadarajan and J. Odobez, "Topic models for scene analysis and abnormality detection," in *IEEE 12th Int. Conf. on Computer Vision Workshops*, Kyoto, Japan, pp. 1338–1345 (2009).
47. Z. Zhang, K. Huang, and T. Tan, "Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes," in *Proc. IEEE Int. Conf. on Pattern Recognition*, Hong Kong, China, pp. 1135–1138 (2006).
48. H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistic Quarterly* **52**(1), 7–21 (2005).
49. B. Moberts, A. Vilanova, and J. van Wijk, "Evaluation of fiber clustering methods for diffusion tensor imaging," in *IEEE Visualization, 2005*, Minneapolis, Minnesota, pp. 65–72 (2005).
50. F. Jiang et al., "Anomalous video event detection using spatiotemporal context," *Comput. Vis. Image Underst.* **115**(3), 323–333 (2011).
51. B. T. Morris and M. M. Trivedi, "Learning, modeling, and classification of vehicle track patterns from live video," *IEEE Trans. Intell. Transp. Syst.* **9**(3), 425–437 (2008).
52. B. T. Morris et al., "Real-time video-based traffic measurement and visualization system for energy/emissions," *IEEE Trans. Intell. Transp. Syst.* **13**(4), 1667–1678 (2012).
53. Turner-Fairbank Highway Research Center, "Pedestrian and bicyclist intersection safety indices final report," http://www.fhwa.dot.gov/publications/research/safety/pedbike/06125/index.cfm (19 August 2013).
54. N. Saunier, T. Sayed, and K. Ishmail, "Large-scale automated analysis of vehicle interactions and collisions," *J. Transp. Res. Board* **2147**, 42–50 (2010).
55. M. G. Mohamed and N. Saunier, "Motion prediction methods for surrogate safety analysis,"Transportation Research Board 92nd Annual Meeting. No. 13-4647 (2012).
56. S. Lefèvre et al., "Modelling dynamic scenes at unsignalised road intersections," Rapport de recherche RR-7604, INRIA (2011).
57. C. C. Loy, T. Xiang, and S. Gong, "Detecting and discriminating behavioural anomalies," *Pattern Recogn.* **44**(1), 117–132 (2011).
58. C. C. Loy, T. Xiang, and S. Gong, "Stream-based active unusual event detection," in *Proc. of the 10th Asian Conf. on Computer Vision*, Volume Part I, pp. 161–175, Springer-Verlag, Berlin, Heidelberg (2011).
59. C. Loy et al., "Stream-based joint exploration-exploitation active learning," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1560–1567 (2012).
60. L. Liao, D. Fox, and H. Kautz, "Learning and inferring transportation routines," *Artif. Intell.***171**(5–6), 311–331 (2007).
61. T. Long et al., "E3tp: a novel trajectory prediction algorithm in moving objects databases," *Lec. Notes Comput. Sci.* **5477**, 76–88 (2009).

Brendan Tran Morris is an assistant professor in electrical and computer engineering at the University of Nevada, Las Vegas. He received his BS degree from the University of California, Berkeley (2002) and his PhD degree from the University of California, San Diego (2010). His dissertation research on "Understanding activity from trajectory patterns" was awarded the IEEE ITSS Best Dissertation Award in 2010. His research focus has been in real-time sensing and processing for understanding environments and situations with emphasis on transportation. His interests include unsupervised machine learning for recognizing and understanding activities, real-time measurement, monitoring, and analysis, and driver assistance and safety systems.



Mohan Manubhai Trivedi is a professor of electrical and computer engineering and the founding sirector of the Computer Vision and Robotics Research Laboratory and Laboratory for Intelligent and Safe Automobiles at the University of California, San Diego. He and his team are currently pursuing research in machine and human perception, machine learning, human-centered multimodal interfaces, intelligent transportation, driver assistance and active safety systems. He serves as a consultant to industry and government agencies in the U.S. and abroad, including the National Academies, major auto manufacturers and research initiatives in Asia and Europe. He is a fellow of IEEE (for contributions to intelligent transportation systems field), a fellow of the International Association of Pattern Recognition (IAPR) (for contributions to vision systems for situational awareness and human-centered vehicle safety), and a fellow of SPIE (for distinguished contributions to the field of optical engineering).