Topic-based Video Analysis: A Survey

RATNABALI PAL and SEKH ARIF AHMED*, UiT The Arctic University of Norway, Norway
DEBI PROSAD DOGRA, Indian Institute of Technology Bhubaneswar, India
SAMARJIT KAR, National Institute of Technology Durgapur, India
PARTHA PRATIM ROY, Indian Institute of Technology Roorkee, India
DILIP K. PRASAD, UiT The Arctic University of Norway, Norway

Manual processing of a large volume of video data captured through CCTV is challenging due to various reasons. Firstly, manual analysis is highly time-consuming. Moreover, as surveillance videos are recorded in dynamic conditions such as in the presence of camera motion, varying illumination, or occlusion, conventional supervised learning may not work always. Thus, computer vision-based automatic surveillance scene analysis is carried out in unsupervised ways. Topic modelling is one of the emerging fields used in unsupervised information processing. Topic modelling is used in text analysis, computer vision applications, and other areas involving spatio-temporal data. In this paper, we discuss the scope, variations, and applications of topic modelling, particularly focusing on surveillance video analysis. We have provided a methodological survey on existing topic models, their features, underlying representations, characterization, and applications in visual surveillance's perspective. Important research papers related to topic modelling in visual surveillance have been summarized and critically analyzed in this paper.

CCS Concepts: • **Computing methodologies** → *Probabilistic reasoning*; *Scene understanding*; *Activity recognition and understanding*; *Visual content-based indexing and retrieval*; *Scene anomaly detection*; *Motion capture*.

Additional Key Words and Phrases: Video Analysis, Topic Model, Unsupervised Learning

ACM Reference Format:
Ratnabali Pal, Sekh Arif Ahmed, Debi Prosad Dogra, Samarjit Kar, Partha Pratim Roy, and Dilip K. Prasad. 2020. Topic-based Video Analysis: A Survey. 1, 1 (March 2020), 34 pages. https://doi.org/xxx

1 INTRODUCTION

CCTV camera setups record and store a huge volume of video data that are unexplored due to the absence of interesting events and shortage of manpower. Interpreting and visualizing large volumes of videos can be challenging due to various reasons such as unavailability of computational hardware, limitation of supervised learning methods, complex nature of the scene, etc. Surveillance videos are summarized and processed with the help of object motion patterns [16, 119, 147, 177]. Motion-guided video analysis systems first extract the motion information by tracking

Authors' addresses: Ratnabali Pal, ratnabali3@gmail.com; Sekh Arif Ahmed, skarifahmed@gmail.com, UiT The Arctic University of Norway, P.O. Box 9019, Tromsø, Tromsø, Norway; Debi Prosad Dogra, Indian Institute of Technology Bhubaneswar, Bhubaneswar, Odisha, India, dpdogra@iitbbs.ac.in; Samarjit Kar, National Institute of Technology Durgapur, Durgapur, West Bengal, India, samarjit.kar@maths.nitdgp.ac.in; Partha Pratim Roy, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India, proy.fcs@iitr.ac.in; Dilip K. Prasad, UiT The Arctic University of Norway, P.O. Box 9019, Tromsø, Tromsø, Norway, dilipprasad@gmail.com.

Manuscript submitted to ACM

the moving objects [126]. Next, the motion tracks are analysed to identify the events of interest. In applications such as traffic monitoring [27], video forensic [118], or crowd monitoring [133], recordings may contain varying motion patterns. Unsupervised methods can be highly productive to deal with such a large volume of data.

"Topic" is defined by the semantic feature that can denotes the category. Topic model [9, 10, 151] is a popular approach used to identify "Topic" automatically from a collection of features by analyzing the occurrences and correlations [85, 135]. Topic models have also been successfully applied in mining textual information and natural language processing [50, 142, 155, 163]. In recent years, similar concepts have been used in various computer vision (CV) tasks [37, 65, 148].

Cameras attached to different sources such as CCTV, smartphones, or drones generate a large volume of unexplored video data. Video is considered one of the most complicated and challenging sources of information for researchers due to: **(i)** complex spatio-temporal relations and **(ii)** variations in visual representations. This makes the processing of videos in a supervised learning framework hard. Hence unsupervised methods are preferred for indexing, searching, and understanding of video contents. Unsupervised clustering approaches such as K-means [68] are popular in many data understanding and grouping. The main drawback of such clustering algorithms is the demand of suitable feature selection and similarity measures. The choice of K (number of clusters) is also important. Hierarchical cluster analysis [130] bridges the gap of cluster selection and interpretation. However, it needs expert inputs for better interpretation. Unsupervised deep neural networks primarily deal with the learning of visual features [72]. In this category, generative approaches [99] can process data in an unsupervised way. Context-Based methods [105] utilize context similarity such as patches or temporal structure to extract similarity among unlabelled data. These methods are highly domain specific and unable to handle the complex nature of data. Other unsupervised methods such as semantic label-based methods [23] use algorithms, simulations, game engines, etc. to generate synthetic labelled data for training. Cross modal-based methods [62] use labelled data to generate labels for similar unlabelled data points. These methods have limitations and cannot discover hidden patterns automatically. Topic-based analysis of large volume complex data such as text and video has shown some potential in various data-driven applications. The topic models are suitable for large volume complex data, where supervised learning is difficult. It is used in searching, recommendation, indexing, event detection, and many more. Unlike unsupervised methods such as cluster analysis, topic-based analysis can discover underlying patterns automatically and it is suitable for video analysis applications too.

1.1 Motivations and Contributions

The main motivation of this work is to summarize the applications of topic models for semi-supervised and unsupervised clustering of actions and events in surveillance videos, classification of events, and learning of distinct events (topics). None of the existing reviews summarized in Table 1 discusses topic models for video analysis. Therefore, a review of topic models used for video surveillance can be a timely contribution to this field of research. We have made the following research contributions in this paper:

- We have summarized topic models, methodologies, and how they have been used in video analysis applications.
- We have provided an overview of the publicly available video datasets applicable to video analysis.

The rest of the paper is organized as follows. In Section 2, we discuss the details of the topic models. The section starts with the state-of-the-art topic models used in text-based analysis. Next, we discuss the possibility of extension from text-based analysis to video-based analysis. This section includes details of the topic models. In Section 3, we discuss the algorithmic comparison of different topic models including time complexity, advantages, and disadvantages.

Table 1. Recent surveys in topic modelling

Year	Ref	Broad Topics
2010	[87]	An empirical comparison of four text mining methods
2012	[69]	Topic models and advanced algorithms for profiling of knowledge in scientific papers
2013	[31]	A Survey on Topic modelling
2015	[5]	A survey of topic modelling in text mining
2016	[82]	A Survey on Interactivity in Topic Models
2016	[110]	LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis.
2017	[70]	Latent Dirichlet Allocation (LDA) and Topic modelling: models, applications, a survey
2018	[41]	A Study of Topic Modelling Methods
2019	[70]	Latent Dirichlet Allocation (LDA) and Topic modelling: models, applications, a survey

In Section 4, we discuss the information representation in video analysis, different applications, and details of the datasets, and evaluation methods. Finally, Section 5 concludes the article.

2 TOPIC-BASED ANALYSIS

Several pattern analysis tasks are solved using machine learning and statistics [5]. Finding patterns of different features in collections of data using a hierarchical probabilistic model, is popular in literature. These models are called topic models. Topic modelling is a kind of unsupervised classification, where a natural group of items, their occurrences, and the distribution of the groups are used for learning and classification. The groups are called "topics". The methods have been primarily designed for text analysis. Fig. 1 shows a typical topic modelling setup used in various text and video analysis setups. First, the unique words and the frequency of occurrences are extracted from the set of documents. Next, the words are grouped semantically, known as "Topics". Finally, a document is classified based on the topics and the distribution of the topics in the document. The topic models are easily generalized to other kinds of data. The topic models analyse different forms of data such as images, biological data, or videos. Here, we first discuss the possibility of extension of the existing models from text to video data analysis. Next, we discuss different topic models that have already been applied for video analysis.

Extension of Text Analysis Models to Video Analysis: Although the majority of the topic models developed so far focus on text analysis, however, these models can be extended to video analysis. The main components in a typical topic-based video analysis framework are (a) feature representation and extraction, (b) defining semantics, and (c) designing suitable topic models. Text data usually contains hierarchical information, namely document, sentences, and words. In a similar manner, a video is represented using a sequence of activities and interactions of objects. A topic of a text data is represented by a "bag of words". A "bag of features" can represent video. This similarity leads to an easy extensibility of the existing topic models to video data. From the dimension of semantics, "topic" is represented by objects, behaviors, activity, events, abnormal events, stories, etc. In the temporal dimension, the topic is denoted by duration, correlated position, sequence of events, etc. However, there are a few challenges still remains. For example, text data comes with additional features such as corpus and word-to-vector representation that help topic models to measure similarity. This is not available with the video data. Table 2 compares the terminologies of topic models used in text and video analysis.

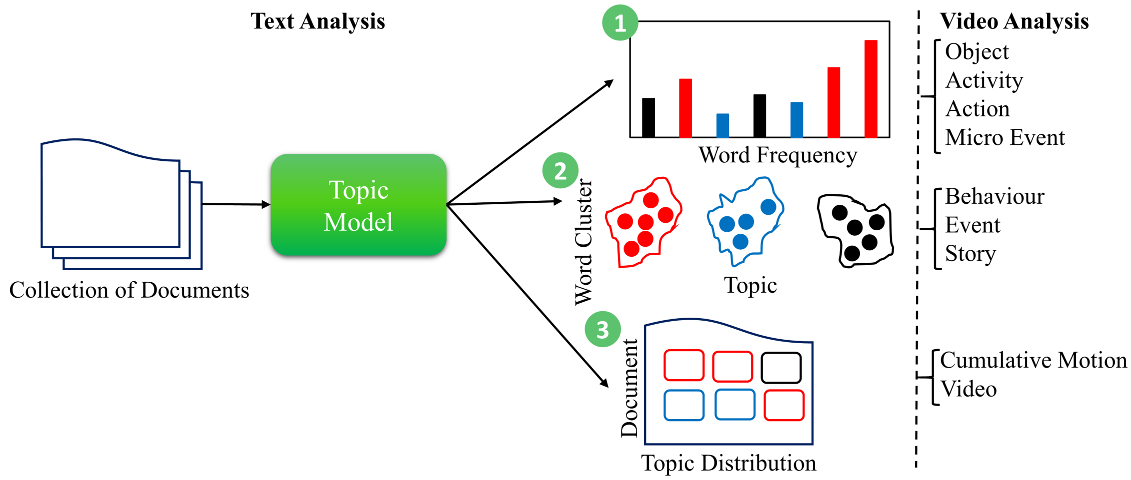


Fig. 1. Generic presentation of topic modelling. The method combines word frequency, word cluster, and topic distribution over documents for extracting the topics.

Table 2. Comparisons of the terminologies used in topic models on text and video data

TEXT Analysis	VIDEO Analysis
A document	A set of trajectories / video clip
A word	An activity / action/event
A topic	An unique activity / pattern

State-of-the-art Topic Models: Several variations of the topic models used for text analysis have been reused for video analysis. The models are primarily categorized into two groups: **(a)** time-independent models and **(b)** time-dependent models. Models such as probabilistic latent semantic analysis (PLSA), Latent Dirichlet allocation (LDA), Co-related Topic Model (CTM), and other extensions of LDA [19, 103, 114] are popular amongst the first category. Analyzing and modelling topics through observations on different trends over time is called "Topic Evolution Models". These types of models are grouped into continuous-time models and discrete-time models. Topic evolution modelling such as Non-Markov Continuous Model and Dynamic Topic Model (DTM) [65], and Multi-scale Topic Model (MST) usually consider a discrete distribution of the topics over time, whereas Topics over Time (TOT), dynamic mixture model (DMM), and Hierarchical Dirichlet Process (HDP) perform parameterization with continuous distributions over time associated with each topic. Fig. 2 depicts the categorization of topic models used in video analysis.

These generative models have been used in semi-supervised and unsupervised ways to perform automatic video analysis and video information retrieval. Next, we discuss the state-of-the-art topic models used in video-based applications. The important notations commonly used in research articles are mentioned in Table 3.

The topic is a probability distribution over features and the data can be modelled by the probabilistic behaviour of the features. Generally, topic models are formulated using (i) observed variables, (ii) latent variables (hidden), (iii) sampling methods, and (iv) conditional dependency among variables. The process of finding hidden topics (latents) from the observed features, is referred to as topic modelling. This can be achieved by finding the probability distribution of features over the data. It is a method for constructing a topic (shared feature) z for a given data d considering the

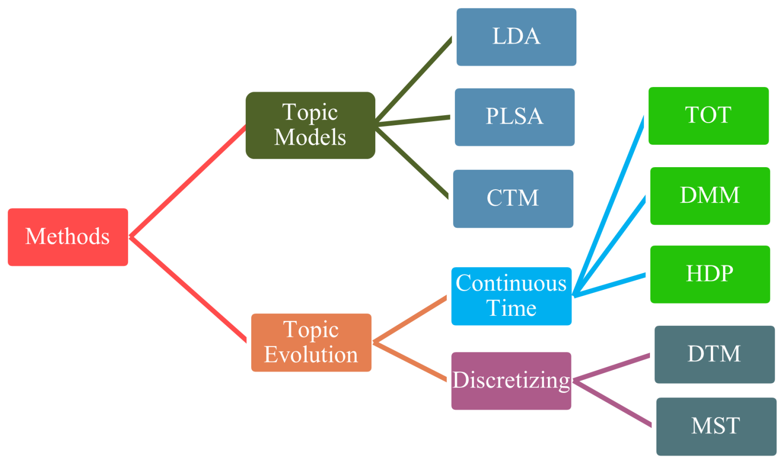


Fig. 2. Categories that can be considered in the field of topic modelling applicable to video analysis.

Table 3. Descriptions of commonly used variables

Variable	Description
T	Number of targets in a video clip
N	Total number of activities in a video clip
X	Observed activity
Z	An atomic activity/topic assigned to X
θ	Probability of topic in a given activity
ϕ	Probability of activity in a given topic

probability distribution of features in a given set of features F. A topic model is interpreted using plate notation. We have also used such notations to demonstrate the topic models used in video analysis. The symbolic representations of different components are shown in Fig. 3.

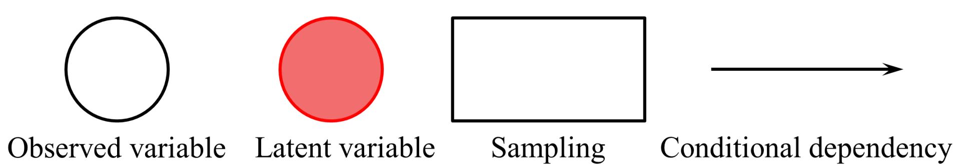


Fig. 3. Symbols used in graph plate notations.

There are mainly two types of statistical topic models available in the literature, namely probabilistic latent semantic analysis (PLSA) [55] and Latent Dirichlet Allocation (LDA) [14]. PLSA provides a co-occurrence perspective to extract topics or themes and LDA is based on Bayesian approach. The methods use different frameworks for modelling topics such as maximum likelihood estimation (MLE) through the Expectation Maximization (EM) algorithm [54, 55], Bayes inference [14, 53, 107, 138], Gibbs sampling [45, 117, 138], correlation based [67], and maximum a posteriori probability (MAP) estimation [13].

- ***PLSA:*** Probabilistic Latent Semantic Analysis [54] is a statistical learning method used to find a mapping between high-dimensional count vectors to a low-dimension space that describes semantic relationships within co-occurrence data. First, a video is represented by the collection of trajectories $T = \{t_1, ..., t_n\}$. Trajectories can be extracted by tracking the objects present in the video clip. A set of micro-activities $X = \{x_1, ..., x_m\}$ of targets is defined by ignoring the order. The video is summarized as a co-occurrence matrix with the terms $c(x_m, t_n)$ that denote how much time the activity x_m has appeared in the clip (t_n). The latent variable $z \in Z = \{z_1, ..., z_k\}$ in PLSA is called as an aspect model and in a video, it represents topics. The joint probability model over video clips and activities can be defined using (1), where $P(t)$ is the probability of an event.

$$P(t, x) = P(t) \sum_{z \in Z} P(x \mid z) P(x \mid t) \tag{1}$$

The conditional probability $P(x \mid t)$ is the probability of observing an activity x given the topic z. $P(z \mid t)$ is video-specific conditional multinomial probability. The parameters of PLSA are estimated using the maximum likelihood principle. For example, given a training video containing a trajectory set (T), θ defines the log-likelihood of the model parameters. PLSA is defined in (2), where the probability model is derived from (1) and $n(t, x)$ represents the number of occurrences of activity x in the video.

$$L(\theta \mid T) = \sum_{t \in T} \sum_{x} n(t, x) \log P(x \mid t) \tag{2}$$

EM algorithm is used for optimizing the classification accuracy described in (2). The graphical presentation of PLSA is presented in Fig. 4. The video clips are then represented in the latent topic space using models like PLSA and used in the traffic flow direction, movement pattern, human action, etc.

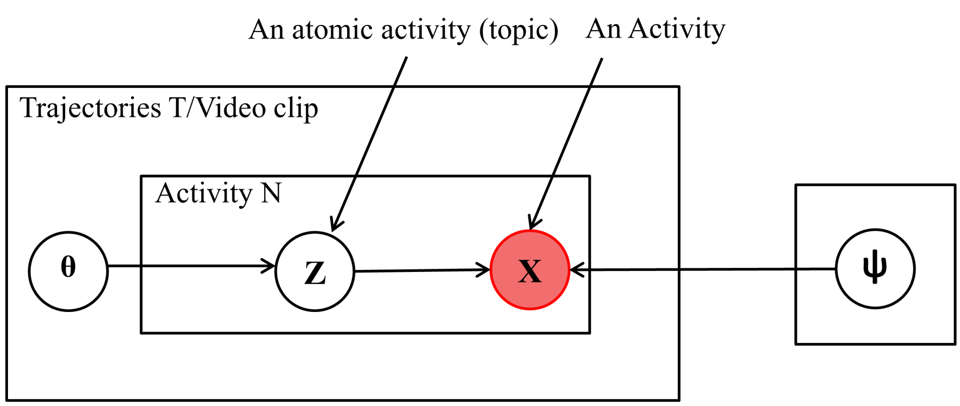


Fig. 4. Graphical representation of PLSA applicable to video analysis. Here, a trajectory/video clip is used for feature extraction and represented by a sequence of atomic activities (Z). The main target is to estimate the high level characteristics of the video/object such as actions, behavior, etc. The document specific topic distribution (θ) and the topic ψ are known or estimated in some cases.

- ***LDA:*** The main drawback of PLSA [14] is, it is not a well-defined model suitable for fully generative probabilistic models as it cannot assign a probability to unknown observations. Latent Dirichlet Allocation (LDA) [14] improves upon PLSA by introducing a Dirichlet prior on θ and ψ. α is the Dirichlet prior with multinomial distribution and β represents the Dirichlet prior parameters that tell how latent topics are mixed in a given video. The joint distribution of a topic mixture θ, a set of activities x observed in the video of length N, and their corresponding topic z are expressed using (3).

$$P(\theta, z, x \mid \alpha, \beta) = P(\theta \mid \alpha)\Pi_{n=1}^{N} P(z_n \mid \theta)P(x_n \mid z_n, \beta) \tag{3}$$

The method can be used further to compute the marginal distribution of patterns by integrating over θ using equation (4). Fig. 5 depicts the graphical representation of the LDA model.

$$P(x \mid \alpha, \beta) = \int P(\theta \mid \alpha)\Pi_{n=1}^{N} \Sigma_{z_n} P(z_n \mid \theta)P(x_n \mid z_n, \beta)d\theta \tag{4}$$

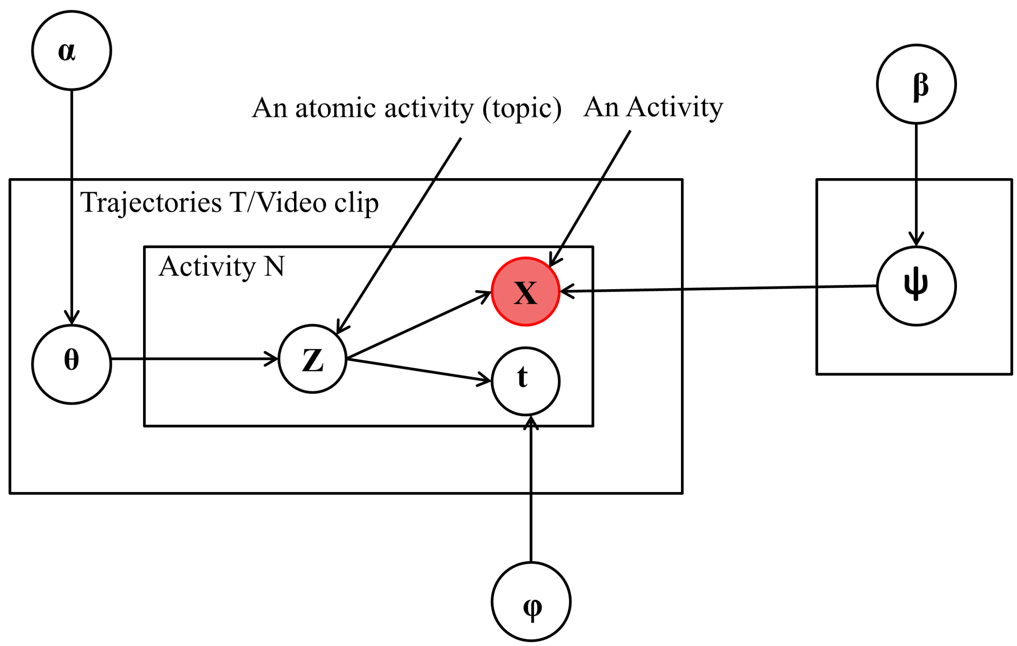


Fig. 5. Graphical representation of LDA in video analysis. This is fully generative as compared to PLSA. Here, a trajectory / video clip is used for feature extraction and represented by a sequence of atomic activities (Z). The newly added parameters α and β are Dirichlet prior with multinomial distribution and the Dirichlet prior with parameters that tell the distribution of topics in the dataset. φ is the activity distribution in the video.

- ***CTM:*** Correlated Topic model (CTM) [12] is an extension of LDA that uses a logistic normal prior to explicitly model correlation patterns with a Gaussian covariance matrix. CTM is capable to model dependencies between different behaviours in an unsupervised framework [120]. Belief-based CTM has been used to learn discriminative middle level features (topics) for trajectory analysis and clustering [186]. Fig. 6 depicts the graphical representation of CTM, where η is assumed to follow a joint Gaussian distribution $\aleph(\mu, \Sigma)$ and z is a latent variable being assumed to follow a parameterized multinomial distribution $f(\eta)$.

Topic Evolution Methods: Topic evolution methods [18, 29, 175] are generative methods that have been used to analyse the evolution of unobserved topics from the video over time for surveillance applications. Evaluation methods

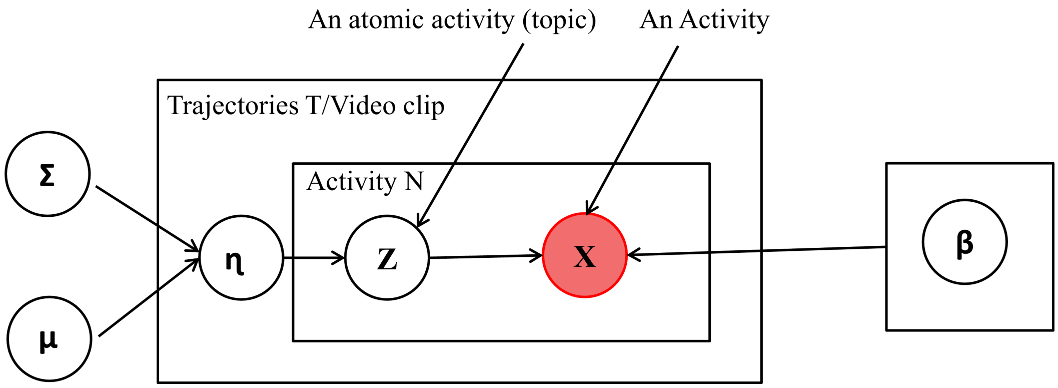


Fig. 6. Graphical representation of correlation topic model applicable to video analysis. Here, the activity (Z) depends on joint Gaussian distribution of μ and σ, where μ, σ represent activity-level topics' mean and covariance.

may be used in various time-dependent models and their research applications to model the associativity between topics and extracted activities provide an efficient tool for monitoring and visualizing the strength of the topic depending on time. In general, topic evolution can be categorized as modelling topic evolution by continuous-time models and discrete-time models.

Continuous models are obtained when observations are collected continuously over a defined period. Neo et al. [106] have introduced a topic evolution method for browsing events based on users' choice and proposed question answering on top of the topic hierarchy to manipulate different functional video search queries. In the topic over time method [154], a topic is considered as being related to a continuous multinomial distribution over time and sampled through a Dirichlet. Fig. 7 depicts the graphical representation, where the β distribution of each topic generates a time stamp and used in topic discovery. Another approach uses Gibbs sampling [52] to discover the topics shown in Fig. 8. In various applications [144], non-parametric hierarchical Bayesian time modelling is used to provide correctness in anomaly detection with a sampling strategy for posterior estimation in activity analysis.

A dynamic topic model is a generative model that implements topic changes over time in sequentially arranged text documents and shows a word-topic distribution that helps to view the topic trends. It is an extension of LDA proposed by Blei et al. [11]. In this model, the data is divided into time slices and it models the documents of each slice with a k-component topic model where topics related to slice t evolve from topics related to slice $t-1$. i^{th} component of the natural parameter $\beta_i = log(\pi_i|\pi_V)$, where EM-MCTM [66] is used for abnormality detection. It shows more effectiveness rather than using Gibbs sampling-based inference when experimenting on both real and synthetic datasets.

- ***HDP:*** Hierarchical Dirichlet Processes (HDP) [153] is a Bayesian non-parametric topic model. Unlike the LDA, HDP does not require the number of topics as a parameter. The number of topics is automatically estimated from the data, hence the method is a popular choice in video analysis [7, 139, 166]. The method initially clusters similar patterns (features) and co-occurring patterns together as topics. In video processing, a global list of activities is represented by (G_0), and its distribution is a Dirichlet Processes. The activities are represented by the concentration parameter α and Dirichlet prior H. For each video segment (t), G_t is randomly chosen from G_0 and concentration parameter β. For any i^{th} activity in t, a topic is chosen as θ_{ti} and the activity (X_{ti}) is a multinomial distribution of G_t. Although the number of topics is determined automatically in HDP, rare and low frequent activities are treated as noise or abnormal events in

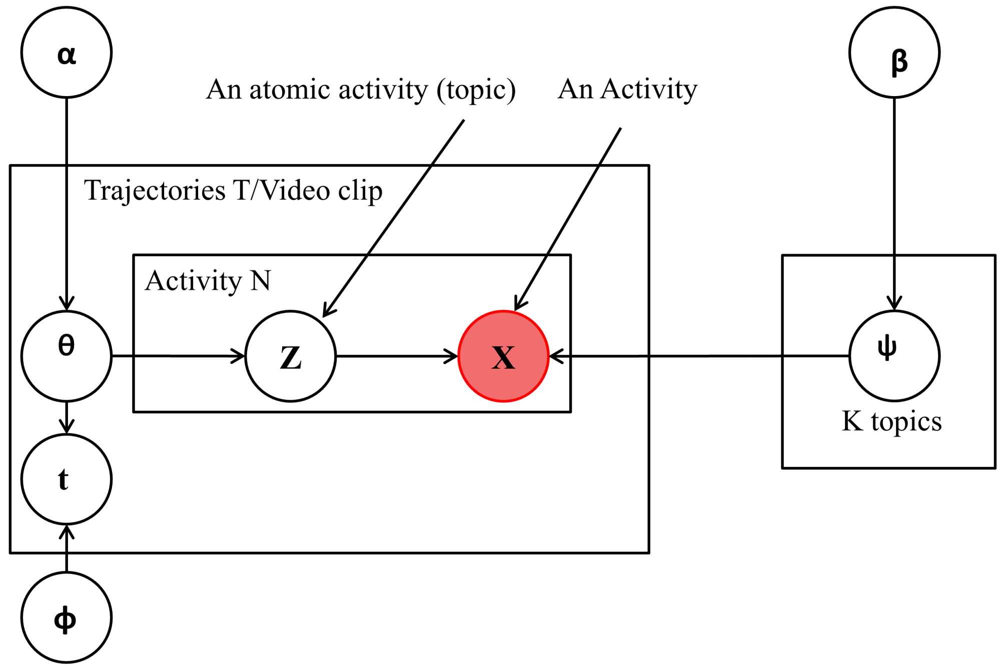


Fig. 7. Topic Over Time (TOT) in video analysis: graphical representation. Here, the parameters α and β are multinomial distribution of activity and topics in the video. φ denotes temporal distribution of activity.

the video. This property is useful in video analysis and abnormality detection. Fig. 9 depicts a graphical representation of HDP.

- ***Random-Field Topic Model:*** Random-field topic model [180] approaches use Markov Random Field that has been used to identify tracklets (fragments of trajectories). Such methods are useful to discover coherent events like follow, together, cross, interaction, etc. Fig. 10 depicts a graphical representation of such a system, where a point on the tracklet is represented by four variables (x, h, z, m) such that x is the fully observed visual word. h and m are the labels of sources and sinks related to past observations. The parameter A denotes the MRF connection between two neighbouring tracklets. θ_i is the distribution of document i over topics. Φ_k is the spatial distribution over topics, where the sources and sinks are denoted by Ψ_k and ω_k.

A space-time MRF model [81] reveals that robustly localized automatic abnormalities in a crowded video clip can simultaneously capture global-level activities via irregular interactions between local activities. Moreover, in the case of moving object tracking, a compress-domain method can use a spatio-temporal Markov Model [80] in H.264/AVC for fast and accurate performance. The model is defined in (5), where π is the mean parameter of V-dimensional multinomial. To model the sequence of compositions of random variables of each topic $\beta_{t,k}$ by chaining Gaussian distributions, an extension of the logistic normal distribution [3] to time-series simplex [158], has been introduced.

$$\beta_{t,k} \mid \beta_{t-1.k} \sim \chi(\beta_{t-1,k}, \sigma^2 I) \tag{5}$$

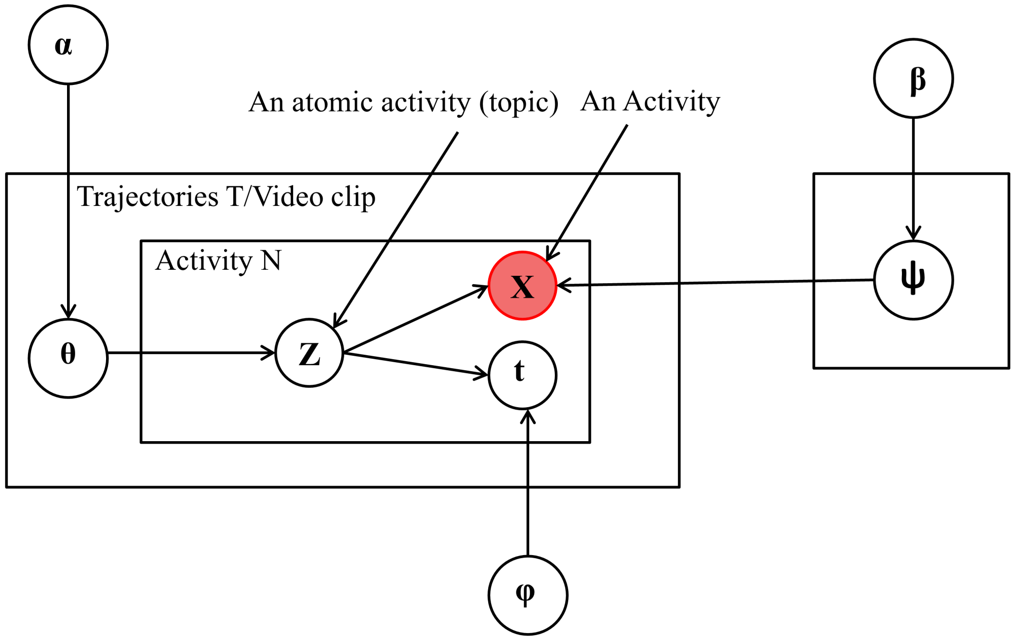


Fig. 8. Topic Over Time (TOT) in video analysis: graphical representation of Gibbs sampling. The main difference with TOT is that the activity distribution (φ) is sampled in a bounded time (t).

The sequential structure between the models is again captured by a logical normal with mean α and uncertainty over proportions. The modified model is defined in (6).

$$\alpha_t \mid \alpha_{t-1} \sim \chi(\alpha_{t-1}, \sigma^2 I) \tag{6}$$

The palate diagram for this generative process is shown in Fig. 11, where π maps the multinomial natural parameters to mean parameters.

$$\pi(\beta_{k,t})_w = exp(\beta_{k,t,w}) \mid \Sigma_w exp(\beta_{k,t,w}) \tag{7}$$

- ***MSTM:*** Another variation of the dynamic topic model, named as Markov Clustering Topic Model (MCTM) [57], is more sensitive, robust, and efficient in handling computational challenges. Markov Chain Monte Carlo (MCMC)-based Gibbs sampling or variational Bayesian inference is another method of such category that can be used for activity discovery in surveillance applications [7, 101, 173]. Another variation of MCTM, namely Latent Dirichlet Markov Clustering (LDMC) [184], has been proposed for modelling human action categorization and correlates them over time. The method has been successfully applied to sensor data [22] and videos for automatic action categorization. Fig. 12 depicts the graphical representation of such systems.
- ***MST:*** One main feature of the topic model is its ability to discover meaningful key motion patterns in a happening scenario by observing a video clip for an extended period. For the problem of pattern recognition, there is a nice probabilistic explanation [169] that uses diffusion maps following low-level feature quantization to identify dominant motion patterns that occur simultaneously at different scales. Processing information in different scales is known as Multi Scale Topic model (MST).

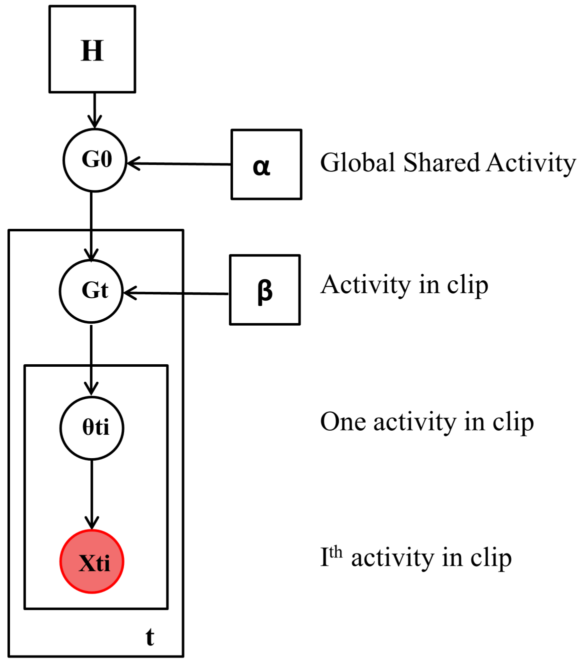


Fig. 9. Hierarchical Dirichlet Processes (HDP) applicable to video analysis. It contains two Dirichlet Processes (DP), namely α and β. The first DP is used to extract a global level activities (G_0) and the second one is a subset of activities from the global set for a clip (G_t). Finally, visual bag-of-words are drawn from activities.

3 ALGORITHMIC COMPARISONS

A model which generates an output considering the prior distribution of some objects, is known as a generative model. Here, we discuss the comparative analysis of the generative models of different algorithms, advantages and drawbacks, and their complexities. PLSA models each feature in a video as a sample from a mixture model, where the components of the mixture model are multinomial random variables. Each video is considered as a variety of mixture models (topic). On the other hand, LDA uses a generative process to infer the topics. The generative process is assumed that the videos (a collection of activities) are represented as random mixtures over latent topics. The generative processes of PLSA and LDA are demonstrated in Algorithms 1 and 2. The CTM uses a logistic normal distribution replacing the state-of-the-art Dirichlet process. This produces more flexibility to the model. CTM incorporates a covariance structure among the different components. This gives a more realistic model of the latent topic structure, where the presence of one latent topic may be correlated with the presence of another. The algorithm is presented in Algorithm 3. The TOT is an updated method of LDA that includes the time information of the LDA. It uses Gibbs sampling procedure as shown in Algorithm 4. HDP is a non-parametric Bayesian approach. It is a hierarchical version of Dirichlet process (DP). The generative process is presented in Algorithm 5, where H is the base distribution and α and β are hyper-parameters. Unlike the standard LDA, MCTM uses a three-layered latent structure. The behaviour is assumed to vary systematically over time. The generative model is shown in Algorithm 6.

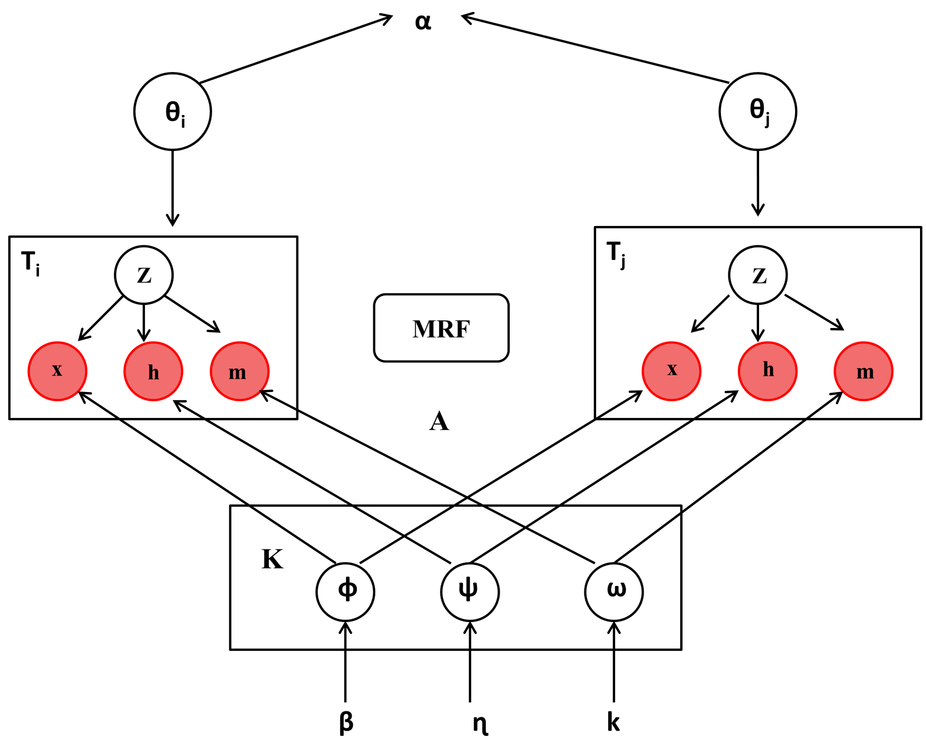


Fig. 10. Random-field topic model used in video analysis. The distribution of video i is defined by θ_i. (Φ, ψ, ω) are the parameters of the specific topic. β, η, κ are the hyper-parameters of a Dirichlet distribution. x, h, m are discrete variables sampled from a discrete distribution from the MRF.

Algorithm 1 PLSA

PLSA (video):
Select an activity with probability $P(\theta)$
for Every feature in the activity θ, Z **do**
    Select topic Z_i from conditional distribution with probability $P(Z|\theta)$
    Select a feature with probability $P(X|Z)$ ▷ Joint probability discussed in equations 1, 2
end for

Algorithm 2 LDA

generativeProcessLDA (video)
$\theta_i \sim Dir(\alpha)$ (Where $i = 1, ..., N; \theta_i \in \Delta_K$) ▷ $\theta_{i,k}$ is the probability that a video $i \in \{1, ..., M\}$ belongs to topic $k \in \{1, ..., K\}$
$\psi_k \sim Dir(\beta)$ (Where $k = 1, ..., K; \phi_k \in \Delta_V$) ▷ $\psi_{k,v}$ is the probability that a activity $v \in \{1, ..., V\}$ in topic $k \in \{1, ..., K\}$
Choose $Z_{i,j} \sim Polynomial(\theta_i)$ (Where $Z_{i,j} \in \{1, ..., K\}$)
Choose $X_{i,j} \sim Polynomial(\psi_i)$ (Where $X_{i,j} \in \{1, ..., V\}$)

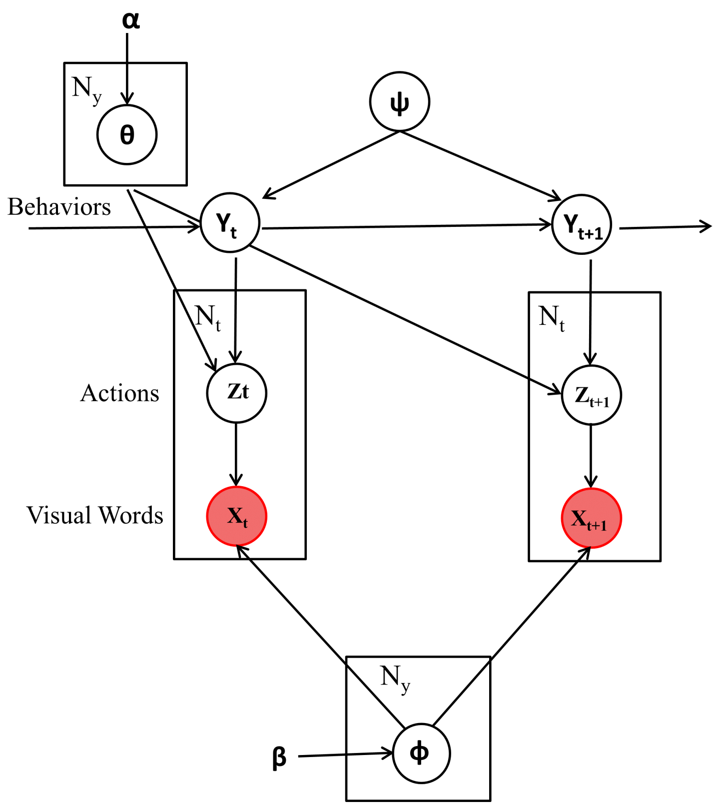


Fig. 11. MCTM model in video analysis. A particular activity is represented by z_t and it is varying systematically over time and assumed to some unknown multinomial distribution $p(z_t|z_{t+1}, \psi)$. Each observed event is chosen based on the multinomial parameters (ϕ, ψ, θ) that are unknown Dirichlet priors.

Algorithm 3 CTM

generativeProcessCTM (video):
for Every feature in the activity $\forall t \in T$ **do**
  Draw $\eta_d|\{\mu, \sum\} \sim M(\mu, \sum)$
  for Every activity in the video $\forall n \in N$ **do**
    Select topic assignment $Z_{n,d}|\eta_d \sim Categorical(f(\eta_d))$
    Select visual words $X_{d,n}|\{Z_{d,n}, \beta_{1:K}\} \sim Categorical(\beta_{Zn})$
  end for
end for

Algorithm 4 TOT

inferenceTOT (video)
Assigns a random topic for all activity
$\theta_i \sim Dir(\alpha)$ (Where $i = 1, ..., N; \theta_i \in \Delta_K$) ▷ $\theta_{i,k}$ is the probability that a video $i \in \{1, ..., M\}$ belongs to topic $k \in \{1, ..., K\}$
$\psi_k \sim Dir(\beta)$ (Where $k = 1, ..., K; \phi_k \in \Delta_V$) ▷ $\psi_{k,v}$ is the probability that a activity $v \in \{1, ..., V\}$ in topic $k \in \{1, ..., K\}$
Choose $Z_{i,j} \sim Polynomial(\theta_i)$ (Where $Z_{i,j} \in \{1, ..., K\}$)
Choose $X_{i,j} \sim Polynomial(\psi_i)$ (Where $X_{i,j} \in \{1, ..., V\}$)

Algorithm 5 HDP

generativeProcessHDP (video):
Select $G_0|\alpha, H \sim DP(\alpha, H)$
Select $G_t|\beta, G_0 \sim DP(G_t,)$
$\theta_t|G_t \sim G_t$
$X_t|\theta_t \sim F(\theta_t)$ ▷ $F = Mult(\theta)$

Algorithm 6 MCTM

generativeProcessMCTM (video):
$p(\psi_z|\gamma) = Dir(\psi_z, \gamma)$
$p(\theta_z|\alpha) = Dir(\theta_z, \alpha)$
$p(\phi_y|\beta) = Dir(\phi_y, \beta)$
$p(z_{t+1}|z_t, \psi) = Multi(z_t, \psi_{zt})$
$p(y_{i,t}|z_t, \theta) = Multi(y_{i,t}, \theta_{zt})$
$p(x_{i,t}|y_{i,t}, \phi) = Multi(x_{i,t}, \phi_{yi,t})$

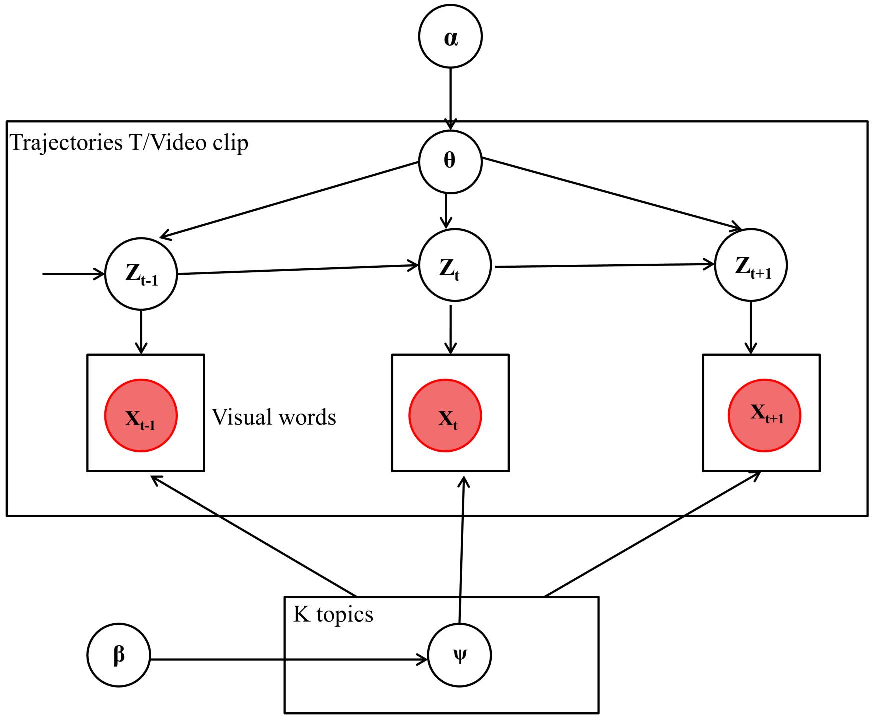


Fig. 12. Graphical overview of LDMC model applicable to video analysis. A particular activity is represented by z_t and it is varying over time. Each observed event is chosen based on the multinomial parameter ψ that is a Dirichlet priors to β.

Advantages and Limitations: Each variation of the topic model is designed for some specific task. For example, statistical models such as LDA or PLSA are suitable for spatial features, whereas the topic evolution is suitable for spatio-temporal features. The addition of temporal features also increases the computational cost in many cases. In Table 4, we have summarized the characteristics and limitations of the different topic models.

Time Complexity: The state-of-the-art PLSA algorithm uses Expectation Maximization (EM) algorithm. The method is a two-stage method involving expectation and maximization. In the expectation step, the posterior probability of a topic is calculated and in the maximization step, the log-likelihood is computed. The computational cost is defined as:

$$C_{time}(PLSA) = O(C_{Iteration}(C_{Estep} + C_{Mstep})) \tag{8}$$

TLDA is an iterative process. In each iteration, it counts short-duration activities and assigns a topic distribution. The complexity per iteration is linear in the size of the data and linear in the number of topics. The number of iterations necessary to get convergence will depend on the video. For a fixed number of iterations, LDA is highly efficient. A major part of the complexity of the task goes into estimating the appropriate number of topics and figuring out the stopping times. HDP is a non-parametric implementation of LDA. Hence it shares similar complexity like LDA. The advantage of HDP is that the number of topics is determined by the data. TCTM is highly effective in several applications, but limited due to the high computational cost. The method uses a pairwise correlation and the non-conjugacy of logistic normal inference. Hence the complexity is $O(K^3)$, where K is the number of latent topics. TMCTM adopts both the

Table 4. Characteristics and Limitations of Topic Model Methods.

Method	Necessity in video analysis	Advantages	Weakness
PLSA	PLSA can filter unimportant information from the data. Hence it is useful in many video applications such as activity recognition and abnormality detection. It is also used in semantic modelling.	(i) PLSA considers local and global activity co-occurrences together. It uses a mixture of conditionally independent multinomial distributions. (ii) Unlike clustering, it uses a mixture model. (iii) It is much interpretable in terms of probability. (iv) It allows multiple combinations of different models.	(i) Computation cost is higher. (ii) Sometimes leads to a local maximua due to the expectation maximization (EM). (iii) May overfit. (iv) Not fully generative.
LDA	Due to the nature of generalizability, it is useful to model different action.	i) Fully generative. ii) Easy to implement.	i) Does not consider correlation among topic. ii) Evolution of topics over time is not considered.
CTM	Due to the use of temporal correlation, CTM is used to model trajectory in video.	(i) Consider the correlation among topics. (ii) Ability to model heterogeneity in number of topics by normal logistic prior.	(i) Inability to construct medium-level features among different clusters.
HDP	Due to the nature of non-parametric, it is useful for unsupervised event modelling.	(i) Non-parametric and number of topics can be estimated.	Sometimes infinite number of topics is not suitable and applications demand finite topics.
MCTM	Several surveillance applications demands real time processing. MCTM is useful in such cases.	(i) Generative model by adding Gibbs sampling theory. (ii) Can be used in online manner.	(i) While an online inference is fast, the procedure is slow enough to provide a barrier to learning on truly large and complex datasets.
RFTM	As it uses Markov random field, it can model the spatial and temporal coherence, hence useful in various scene and motion analysis.	(i) It is an extension of LDA by integrating space-time Markov random Field.	(i) Lower completeness accuracy as it does not able to modify the neighboring topics information during learning.
Multi-scale	Based on low-level features hence, able to model pixels and optical flow.	(i) Based on low-level features. (ii) Can model different scale.	(i) High computation cost.

concept of LDA and HMM. The time complexity of MCTM is hard to quantify because of the data-specific convergence time consumption. In general, the method demands $O(F_iT)$ training time, where F_i is the number of input features and T is the number of topics. During testing, the time cost is $O(S^2) + O(F_iTS)$, where S number of states are present in the HMM. Due to the convergence criteria, RFTM also shares similar computational complexity with MCTM. Multi-scale topic models are the most expensive models due to the use of low-level features in multiple scales.

Deep Learning and Topic Models: The topic models primarily deal under a probabilistic framework and these models are used in various data modelling and understanding tasks. On the other hand, in the last few years, we have been witnessing a rapid progress in deep neural networks and machine learning applications. The majority of such deep learning methods use supervised learning strategies that demand labelled data, whereas unsupervised learning methods primarily use clustering techniques. The main advantage of topic model is, it can automatically discover interpretable patterns from the data. It does not require labelled data. Only a few research works have been reported that combine deep learning and topic models together. In text processing, Cao et al. [17] have proposed a neural topic model and an extension using a supervised approach. The method has been used to classify texts into different classes. Dieng et al. [30] have presented a topic-based recurrent neural network (RNN) for sentiment analysis. Lv et al. [98] have used LDA and deep learning to describe videos using language. Recently, Dong et al. [32] have used LDA-based topic discovery and learning to produce interpretable deep learning for video description. Yu et al. [171] have used topic discovery combined with CNN for image caption generation. Chen et al. [21] have used a Latent topic for discovery and video narration generation.

4 REPRESENTATION, APPLICATIONS, AND DATASETS

Here, we discuss **(i)** different features and information embedding methods used in topic discovery and analysis, **(ii)** different video-based applications, and **(iii)** benchmark datasets and evaluation methods.

4.1 Topic Representation and Feature Embedding:

State-of-the-art topic models are designed for language models. Hence the conventional features used in computer vision may not be suitable for topic models. Various topic modelling methods have been applied to discover activity patterns in video clips or motion trajectories. The models use a similar concept that is adopted in text mining, named "bag-of-words (BoW)" in the form of a bag of features (BoF) in video analysis. Statistical topic models use spatial features such as patches, pixels, shapes, etc. as the baseline. Time-dependent models use trajectories, optical flow, motion, etc. for extracting the topics. Table 5 summarizes the features used in various video-based analysis. In typical modelling frameworks, the whole video sequence is divided into non-overlapping short clips as documents, where the clips are random mixtures over latent topics (activity categories) extracted from the features. Next, we discuss the embedding methods used in different topic models.

Table 5. Comparison of terminologies of topic models in text and video analysis

Feature Types & Representation	**References**
Object trajectory (Position, geometrical shapes, sizes, centroid, velocities, etc.)	[49, 57, 94, 128, 186] [78, 125, 136, 139, 145, 147, 180, 182, 187]
Spatio-temporal features (including patches and saliency)	[36, 80, 108, 113, 141, 160, 162, 183, 185]
Visual words such as direction and motion	[65]
Histograms of Motion, direction, color, texture, pixel change, etc.	[95, 153]
Inter-frame color distribution	[149]
Optical flow (pixel motion)	[42]

Embedded Topic Model (ETM) is focused on the word embedding mechanism. Topic models relay on the smallest unit of information such as words in NLP or visual words in video analysis. Hence the representation and information embedding play a vital role in the success of the models. The word embeddings begin with the natural language model [14]. In that method, the words are represented using a "one hot" encoding method. The main drawback of such a system is, similar words are represented using different encoded values. Later, the problem is solved using a lower-dimensional vector representation, where similar words are in close in space [123]. Text-based topic models have used various word embedding methods in different ways [86, 102]. Information embedding in video analysis is different because of the unavailability of common representations such as language. Bag-of-words based representation is popular in many video analysis applications [61, 121]. The bag-of-words is constructed using low-level features such as pixels [43] or motion tracks [152]. The problem of representing similar concepts using similar bag-of-words is solved using the contextual relevance representation [59, 98]. The method uses language embedded with visual words for finding the similar concepts and applied in video analysis. Joint embedding of visual information and textual information is popular in many topic-based video analysis [100, 111]. A graph-based embedding method is used in [172]. The method uses a graph to model the appearance of a human used to classify actions in video. Habibian et al. [48] have proposed to use the description of videos for information embedding and it has been used in video story generation. A velocity pattern-based embedding [71] is used to identify abnormal traffic activity. Jing et al. [73] have proposed to use a multimodal information embedding to classify micro videos. The method uses visual, acoustic, social, and textual modalities. Visual state binary embedding method [170] is used for event classification, where a small activity is known as a state.

4.2 Applications

The primary application of topic modelling is to use the observed patterns of the targets in a video sequence to infer hidden topic structures or patterns. Fig. 13 depicts the general structure of topic modelling applicable to motion-based video analysis. Here, the input is a set of moving object trajectories extracted from the video recordings of QMUL [57] dataset. Based on various topic models and parameters, the algorithm can infer hidden topic structures, i.e., the patterns of movements. Furthermore, we can predict the events/actions from a set of known topics. In this way, topic modelling can provide an automatic solution for surveillance scene analysis, event detection, or action recognition. The process begins with feature extraction, embedding, and ends with classifying or clustering different patterns (topics). Fig. 14 depicts such a representation of topics applied on publicly available surveillance videos. Next, we discuss specific applications in detail.

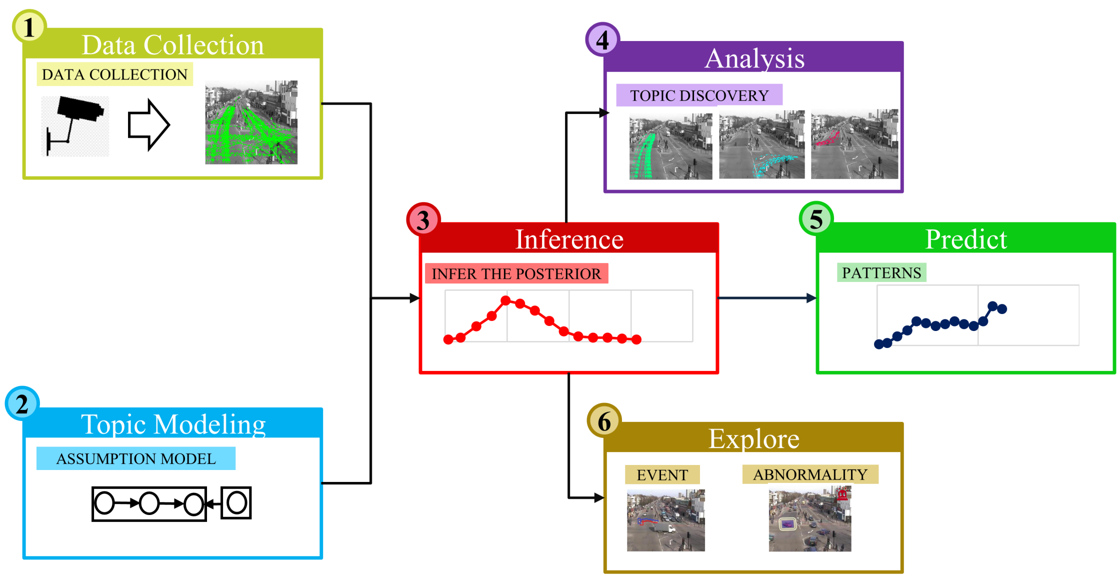


Fig. 13. A typical framework of topic-based video analysis application. (1) Collect the motion information by feature tracking [8] or multi-object tracking [51]. (2) Setup topic model parameters such as topic domination, classification, etc. (3) Infer the posterior based on the model. (4) Discover distinct topics and classify them. (5) Predict the future movements. (6) Explore events, patterns, and actions (topics).

Behavior and Event Analysis: Probabilistic topic model-based action / pattern identification has been used to discover and learn real-world events and event relations [160]. MCTM [65] has been used in transportation systems, security, and surveillance for activity behavior analysis. A dynamic casual topic model [36] has been proposed to mine activities in crowded and complex scenes, where all temporal relationships are updated at every time step using noisy OR distribution. Dynamic Bayesian model [57] is proposed for mining and screening irregular spatio-temporal patterns by clustering visual events into actions and discovering behaviors. Xue et al. [164] have used HDP-based methods for sequential event detection. Saleemi et al. [125] have proposed such a probabilistic model based on MCTM framework that is used to improve foreground detection and object tracking. Fig. 16(c) depicts one event ⟨cleaning table⟩ from a set of events (topics). Recently, Al et al. [4] have proposed graph-based extensions of LDA and CTM, referred to as GLDA

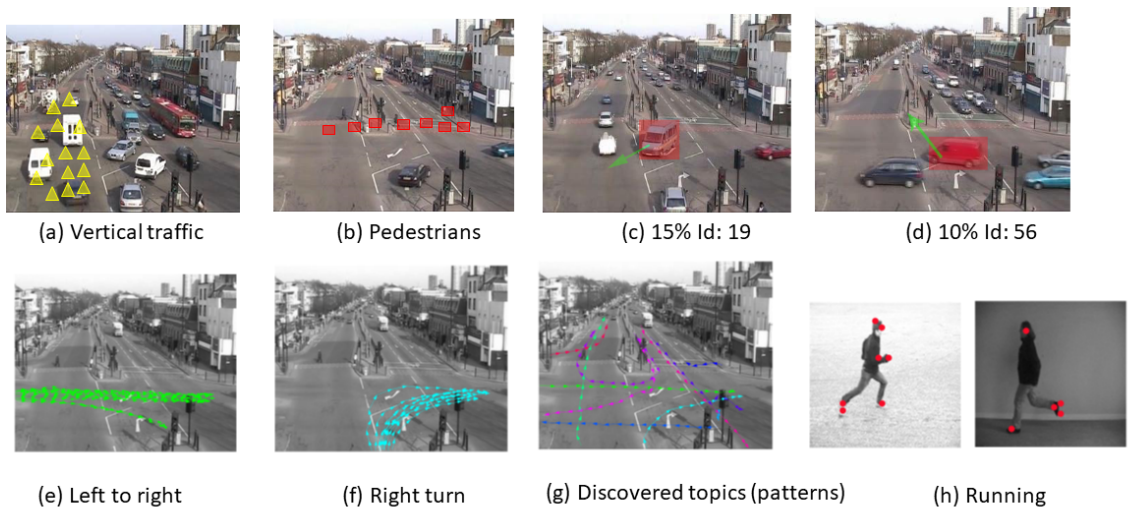


Fig. 14. Various topics and representation of topics in public video dataset. (a)-(g) taken from QMUL [57] junction video, (h) Taken from KTH dataset. (a) Vertical traffic [57], (b) Pedestrian activity [147], (c) A sample target (car), the movement pattern has 15% weight (i.e 15% targets follow the pattern), (d) Another target (red car) with 10% weight, (e) A discovered topic [147] (left to right movement), (f) Right turn [147], (g) Automatically discovered patterns using topic modelling [147], each colour represents different pattern/topic and (h) Bag of Features (BoF) representation of a running event [174].

and GCTM, to learn and analyze motion patterns by trajectory clustering. Xue et al. [165] have proposed a supervised sequential symmetric based HDP model for multi-class video classification. Speech word topic embedding based lecture video classification combined with deep neural network has been reported in [76]. Long short-term memory (LSTM) combined with a topic model [156] has been used to segment interesting segments in videos.

Abnormal-behavior Detection: MCTM [64] uses the temporal dynamics of behaviour for determining activity distributions in each video. The authors have used the Expectation-Maximization (EM) algorithm for optimization and threshold-based abnormality detection. Isupova et al. [65] have developed a maximum a posteriori (MAP) [24] estimation using EM algorithm and variational Bayes inference [13] for anomaly detection. An unsupervised approach such as Bi-Layer sparse topic model has been proposed to discover semantic motion patterns and to detect abnormalities in a dynamic scene [147]. In [185], a new framework is proposed for spatio-temporal point clustering-based normal behavior patterns identification and online abnormality detection. The work combines HMM [46] and LDA. Semi-supervised sparse topic model guided abnormal event detection has been proposed in [148]. Probabilistic Latent Space Model based video abnormality detection has been presented in [131]. Fig. 16(b) depicts an abnormal situation ⟨ illegal crossing ⟩ from a set of normal events (topics) in VIRAT dataset. The HDP-based method can identify noisy and low frequent patterns and it can be used in various video abnormality detection such as abnormal traffic activity [94, 182], crowd patterns [128], unusual human actions [139], etc.

Scene Analysis: Surveillance scene analysis such as traffic and crowd analysis [93] is challenging due to the nature of complex movement patterns. In [95], topic modelling has been used for tracking targets. The main advantage of such a system is its real-time processing capability. In [153], the authors have used LDA to discover and provide a summary of typical atomic activities and interactions occurring in a scene. In [40], LDA has been used as a multimodal framework to build connectivity between attributes and features of each modality that helps to make a difference for semantic and cross-modal gaps. The work proposed in [162] describes a new framework for surveillance scene understanding. Iyer et al. [67] have proposed a correlation LDA-based method for indexing and retrieval of videos. Histogram of optical flow (HOF) [34] has been used in various LDA models to classify actions and events considering the code length and patterns. Non-parametric HDP is also used in various traffic scene analysis [1, 2, 128]. Fig. 15(b) depicts the different topics (paths) that are present within a video sequence of QMUL [57] dataset.

Activity Recognition: One typical application of visual pattern analysis is human action recognition. The process for identifying an action using PLSA and other probabilistic topic models have been discussed in [116]. Unsupervised action categorization using local shape context features has been proposed in [174]. The method is based on structured PLSA with a codebook action representation. In visual pattern discovery and video analysis [33, 146], topic models have been used to build top-down approaches for diverse applications of temporal event detection. LDA-based applications [63] have been used to discover daily routines from a combination of activity patterns in an unsupervised manner. Another variation of LDA is used to classify micro events in large volume video datasets [79]. LDA and PLSA-based algorithms [108] have been used to automatically recognize and localize multiple actions in long and complex video sequences. LDA is also used for dominant codewords selection [78], where BoW based on the dominant dense trajectory [145] is used as input. Recently, an improved unsupervised object discovery and localization method, named Dirichlet allocation with a mixture of Dirichlet trees (LDA-MDT) [109], has been proposed. In [35], LDA has been used to guide an autonomous robot to collaborate on joint activities from long-term observations in crowded scenes. It has also been applied to person identification and action recognition [28]. Unsupervised HDP model is also used to identify distinct human actions [139]. Santhosh et al. [127] have proposed a non-parametric Gibbs sampling-based method for clustering traffic patterns. In [83], authors have extended it and proposed a variation of HDP model called modified Dirichlet Process Mixture Model (mDPMM), which is an unsupervised topic model. The method has been used to cluster different patterns of movement in QMUL junction. LDA combined with other modalities such as convolutional neural network (CNN) [38] is used to classify different indoor and outdoor scenes. LDA is also extended for different group activity recognition [176]. A combination of high-level and low-level features is also used in activity recognition in video [168]. For example, Fig. 15(a) depicts an action ⟨jack⟩ from a known set of actions (topics) such as a walk or running in the KTH action dataset. Fig. 15(c) depicts an event ⟨cleaning table⟩ from a set of events (topics) in the KIT Robo Kitchen [124] dataset.

Anomaly Detection: Anomaly or abnormality detection [104, 113, 115] is referred to as a process to identify the anomalous events in surveillance videos. Sometimes, it has been modeled as a typical semantic scene segmentation problem [89] to divide a scene into different regions as inputs for global behavior inference. Both PLSA and hierarchical PLSA have been used in correlation behavior modelling and anomaly detection, where the hierarchical PLSA is superior for anomaly detection due to its robustness to noise. Varadarajan et al. [141] have investigated situations where several actions can occur in the same scene concurrently. Video-based human abnormal behavior detection using methods including PLSA [42] has been discussed in [115]. The PLSA is also extended in unsupervised learning environments to find unusual activities [25]. Anomaly detection in an automated surveillance system [132] uses a multi-class approach known as multi-class delta LDA that generates new unseen topics regarded as abnormal behavior. Multi-class LDA is also useful to find rare activities [91]. In one of the recent works [42], researchers have described an LDA model for streaming video dataset and then used it to detect anomalous events by an underwater robot. LDA model has been used to recognize events that are not ordinary or surprising [49]. Recently, Li et al. [88] have proposed an LDA-based method that acts as an encoder for low-level features to locate high-level abstractions for video concept detection. Hospedales et al. [58] have proposed a weakly supervised joint topic model for rare event detection in traffic videos. LDA has also been used in biological and geological research by implementing the concept of substrate mapping [75]. LDA is a mixture model over documents and it has a latent variable for topic assignment for each word. The number of topics in a corpus (k) needs to be parameterized by the user to get promising results. To solve this sparsity problem, non-parametric hierarchical Bayesian approaches [74, 137] have been introduced. Hierarchical Dirichlet Process (HDP) [15, 128, 143, 159, 182] is a non-parametric Bayes framework that automatically finds the number of latent topics that can be used for trajectory clustering. For example, Fig. 15(c) depicts an abnormal event by analysing multiple targets and interactions in QMUL junction video. Fig. 16(a) depicts the semantic regions in GCS [181] crowd video.

Other Applications: Topic model is also used in other applications such as video description generation, video indexing, object tracking, etc. Iyer et al. [67] have proposed a Correspondence LDA for a multimedia retrieval system. The method has been applied in indexing multimedia video clips. Chen et al. [20] have proposed a topic-guided method for video description generation. The model combines the language cue with the visual cue. A similar method is also proposed in [100, 111]. Huang et al. [60] have used the topic method for object tracking. Object tracking using different topic-based methods has been proposed in [97]. Chen et al. [21] have proposed a topic model guided deep neural network for video description generation. Some key applications and publicly available datasets are summarized in Table 6.

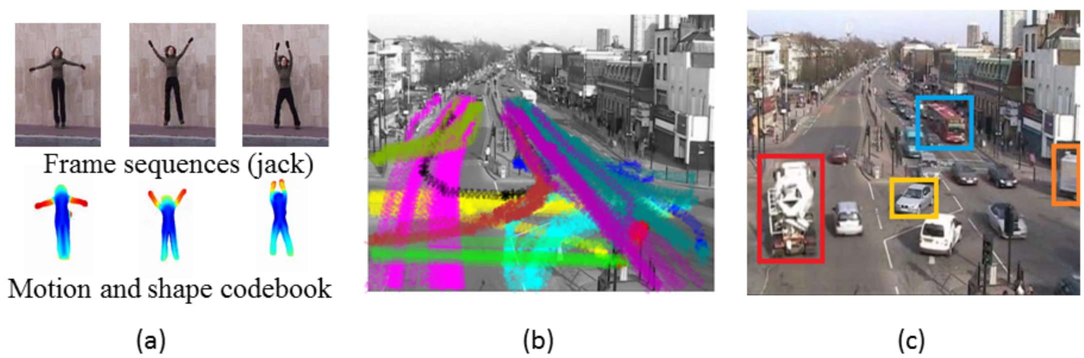


Fig. 15. (a) Action (jack) is identified in KTH action dataset using structural pLSA (SpLSA) [174], where codebook is generated using shape and motion. (b) Identified similar activities based on the movement pattern in QMUL junction video [162]. The method uses LDA to model different paths (shown by a different colour) by unsupervised topic modelling and uses to find similarities in heterogeneous surveillance videos. (c) A two-stage hierarchical pLSA model [89] is used to model abnormality. The picture depicts an abnormal situation in QMUL junction. Different classes of local behaviors in the clip that caused the anomaly are shown using bounding boxes of different colours.

Table 6. Datasets references, and applications covered in various topic models guided research work

Base Method	**Dataset and Application References**	**Applications**
PLSA	Crowded outdoor scenes[89, 141]	Scene and abnormality analysis
PLSA	Traffic, junction, highway [113]	Anomaly detection
PLSA	WEIZZMAN and MIT-CSAIL Datasets [174]	Action categorization
PLSA	Different traffic Datasets [140]	Activity pattern recognition
LDA	WEIZZMAN, KTH and figure skating Dataset [108]	Action recognition

LDA	Street and Pedestrian Path surveillance [49]	Anomaly detection
LDA	Crowed scene (busy train station, shopping mall) [153]	Atomic activity detection
LDA	Daily activities (dinner, lunch, office work) [63]	Activity pattern recognition
LDA	human computer interactions and kits navigation [95]	Object tracking, HCI, navigation
LDA	INRIA, IXMAS, NTSEL, MPII activities [78]	Action recognition
LDA	Caltech4, LabelME, PASCAL07 [109]	Action recognition
LDA	Traffic surveillance [162]	Scene analysis
LDA	Underwater vehicle monitoring [42]	Anomaly detection
LDA	Trcevid2013 for semantic indexing development dataset [88]	Anomaly detection
LDA	Human activity monitoring by mobile robot [35]	Activity recognition
LDA	Microactivity classification [79]	Action recognition
LDA	Scene classification in UIUC dataset [38]	Scene classification
LDA	Video description in MSR-VTT dataset [20]	Scene classification
LDA	Group activity recognition in USAA dataset [176]	Group activity classification
LDA	Human activity recognition in KTH dataset [168]	Human activity classification
LDA	Human action recognition in KTH and UCF dataset [167]	Human activity classification
LDA	Traffic abnormality detection in QMUL dataset [?]	Traffic activity classification
LDA	Weakly supervised traffic behaviour analysis [58]	Traffic activity classification
HDP	Traffic analysis [128]	Activity recognition
HDP	Event detection [94]	Event recognition
HDP	Behaviour analysis [182]	Anomaly detection
HDP	Action recognition [139]	Anomaly detection
HDP	Traffic activity analysis in QMUL [83]	Activity analysis
CTM	Crowd, PETS09, UCF Crowd [120]	Behaviour analysis
CTM	RGB-D activity video dataset using Kinect V2[186]	Action and object detection
DTM	QMUL Junction, Pedestrian crossing [160]	Abnormal behavior detection
DTM	QMUL Street interaction and Idiap Traffic Junction [147]	Abnormal behavior detection
DTM	QMUL Junction [65]	Behavior Analysis
DTM	QMUL Street Interaction, Pedestrian Crossing, Subway Platform, MIT Traffic [36]	Behavior analysis
DTM	Video of human and traffic with occlusions [57]	Object location tracking
DTM	MED and TREC AP88 [125]	Abnormal behavior detection
DTM	Special areas (station, airport, junctions, etc.), MIT traffic datasets, Marathon Race video [24]	Abnormal behavior detection
DTM	Road Junctions/UMN datasets [22]	Abnormality detection
DTM	Domestic activities monitoring hh120, hh122 datasets [66]	Human action recognition
RFTM	Subway station path surveillance [81]	Abnormal activity detection
RFTM	Surveillance video sequences [80]	Object tracking

MCTM	Behaviour analysis in traffic [56]	Behaviour analysis
PLSA	Unusual activity analysis [25]	Abnormality detection

4.3 Datasets, Evaluation Metrics, and Benchmark

Here, we discuss some application-specific datasets and evaluation methods used in various topic-based modelling. WEIZMANN action recognition dataset [44] consists of 90 low-resolution (180 × 144, 50 fps) videos of different activities such as "running", "walking", "jumping-jack", "jumping-forward-on-two-legs", "jumping-in-place-on-two-legs", "galloping-sideways", "waving-two-hands", "waving-one-hand", "bending", "skipping", etc. MIT-CSAIL datasetis a hand gesture-based activity recognition dataset. The dataset consists of "Expand Horizontally", "Expand Vertically", "Point and Back", "Double Back", "Flip Back", and "Shrink Vertically". KTH dataset [129] consists of six different actions such as "boxing", "hand-clapping", "hand-waving", "jogging", "running", and "walking". The dataset is a collection of indoor and outdoor videos (160 × 120, 25 fps). INRIA surgery dataset [60] is an activity dataset centering surgical tables. The dataset includes "cutting", "hammering", "repositioning", and "sitting". MPII Cooking activities datase [122] contains 65 activities in kitchen. The common activities like "cutting", "mixing", "blending", etc. are included. IXMAS action dataset [157] is focused on actions recorded in different viewpoints and in the presence of partially occluding actors. NTSEL traffic dataset [77] is a collection of different traffic activities such as "walking", "crossing", "turning", and "riding a bicycle". The dataset reported in [35] is recorded by cameras mounted on an autonomous robot. The dataset consists of different human activities such as "microwave food", "open fridge", "throw trash", etc. QMUL dataset [57] is a traffic junction video dataset and used in several applications such as object tracking, event detection, motion pattern clustering, etc. The dataset consists of the activities of vehicles and pedestrians. Grand Central station dataset [181] is also one of the popular crowd datasets used in motion clustering, scene understanding, and activity analysis tasks. The dataset consists of videos of more than a thousand people moving and interacting. Junction and Roundabout [90] dataset is a traffic dataset that contains high-quality surveillance videos of different traffic activities. The dataset is used in various behaviour analysis and abnormal behaviour classification. Trecvid 2013 semantic indexing [88] is a collection of web videos of diverse concepts such as object (aeroplanes, bus, computer, etc.); scenes (hills, oceans, fores, etc.); activity (running, walking, skating, etc.); interaction; etc. Traffic dataset reported in [141] is a 45 minute video of size 288 × 360 recorded in a busy traffic junction. The dataset is divided into small activities (125 frames each). More than 2500 such activities are labeled in the dataset. Cooking activity dataset [122] involves 12 participants to record 60 different cooking activities. The dataset is popular in indoor activity classification. UCF crowd dataset [6] consists of a variety of different crowd activities collected from different sources. The dataset is used in various crowd activity monitoring and abnormality detection. A large volume long duration video surveillance dataset is reported in [125]. The dataset contains a recording of video for 3 days containing different activities and used in various activity analysis and abnormality detection tasks. UMN dataset [22] is a large collection of different surveillance videos used in different event detection, such as detection of abandoned objects, detection of unusual crowd activity, detection of loitering individuals, etc.; activity analysis; and abnormality detection. UCF action dataset [134] is a large collection of youtube action videos. The dataset consists of 101 different events in various conditions. UIUS dataset [92] is a sports event dataset that contains different sports action videos such as "rowing", "badminton", "polo", "bocce", etc. MSR-VTT dataset [161] is a large-scale video description generation dataset. The datset contains diverse video topic such as "music", "cooking", "daily activity", etc. USAA dataset [39] is collection of social interaction and activity dataset

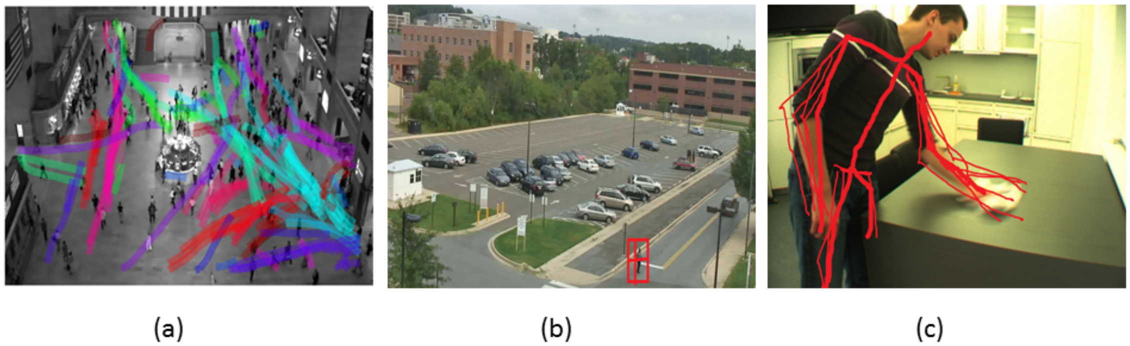
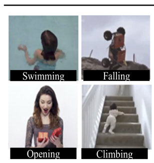


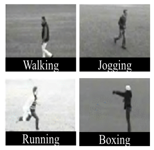
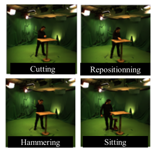
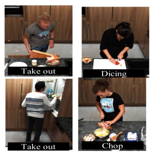
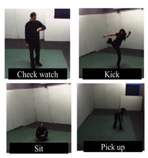
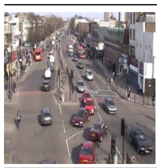
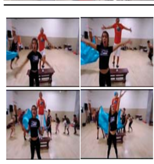
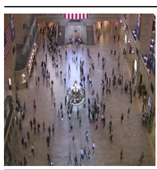
Fig. 16. (a) Depicts different crowd activities (different colour) in Grand central station (GCS) [181] videos. The method uses Hierarchical Dirichlet Processes (HDP) to model semantic regions and patterns. (b) A situation in VIRAT dataset. The pedestrian is identified and localized as abnormal when crossing the road. The method uses a Markov Random Field topic model [81] to find and localize anomaly. (c) An activity representation using dynamic topic model (DTM) [160] to relate with object motion (skeleton) and interaction. The action is ⟨ cleaning table ⟩ in Kitchen dataset [124]

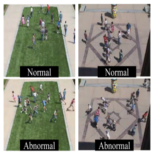
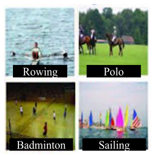
.

such as “wedding Dance”, “birthday party”, “graduation ceremony”, etc. Table 7 summarizes popular datasets appeared in various topic-based analysis.

Table 7. Popular video datasets used in various topic-based analysis

Samples	Dataset	Description	Annotation	Topic-based Analysis
 HMDB51	HMDB51	A Large Video Database for Human Motion Recognition	51 action class	Latent topic models [96], pLSA [150]
	UCF101	A Dataset of 101 Human Actions Classes From Videos in The Wild	101 action class	Multi-view topic model [59], HDP based model [139]
	WEIZMANN	Human Action Dataset	10 natural actions class	Latent topic models [96]
	MIT-CSAIL	A large-scale dataset for recognizing and understanding action videos	399 activity class	Probabilistic latent model[174]

	KTH	A dataset for recognition of human actions	6 action class	Topic guided unsupervised learning [108], Probabilistic latent model [174]
	INRIA	A dataset recorded in surgical table	4 activity class	LDA based model [78]
	MPII	Cooking Activity Dataset	78 classes	LDA based model [78]
	IXMAS	A large Dataset of Human Actions Recorded From Different Angle	11 action class	LDA based model [78]
	NTSEL	Pedestrian Activity Dataset on Road	9 activity class	LDA based model [78]
	QMUL	Video Recorded in a Traffic Junction	12 different patterns	Modified DPMM [83] Causal topic mode [160]
	MSR-VTT	A Large Video Description Dataset for Bridging Video and Language	257 popular query text	Topic-guided model (TGM) [20]
	GCS	A Large Field of View Video Recorded in a Railway Station	Not labeled	Mixture model [181] CTM [181]
	Trecvid 2013	A large Dataset of Semantic Indexing	60 concepts	Latent topic model [26]

	UMN	A large Dataset of Human Activity	16 activity class	EM-MCTM [64]
	UIUC	Sports Event Dataset	8 sports event	Supervised LDA [84] LDA based model [179]
	USAA	Unstructured Social Activity Attribute Dataset	69 attributes	Relevance topic model (RTM) [176] Latent semantic analysis [178]

Different applications use different evaluation methods. Classification-related applications such as action classification, event classification, abnormality detection, etc. use accuracy (AC), precision (PR), recall (RE), and sometimes F1 score. These are defined as:

$$\begin{aligned} PR &= \frac{TP}{TP+FP} \\ RE &= \frac{TP}{TP+FN} \\ AC &= \frac{TP+TN}{TP+TN+FP+FN} \\ F1 &= \frac{2TP}{2TP+FP+FN} \end{aligned} \tag{9}$$

In the above formulation, TP is “True Positive”, FP is “False Positive”, TN is “True negative”, and FN is “False Negative”. Different clustering approaches are used in activity mining, motion analysis, cumulative behaviour analysis, etc. to cluster similarity evaluation methods such as Rand index (RI), Jaccard index (JI), Dice Index (DI). These are defined in (10).

$$\begin{aligned} RI &= \frac{TP+TN}{TP+FP+TN+FN} \\ JI &= \frac{TP}{TP+FP+FN} \\ DI &= \frac{2TP}{2TP+FP+FN} \end{aligned} \tag{10}$$

Text embedding with visual features is a popular choice in various topic models. This is used across different applications such as image and video captioning; semantic activity detection; similarity detection; etc. Here, Bilingual Evaluation Understudy (BELU) [112] is a popular choice for evaluation. BELU is used to measure the similarity between the candidate text (t_c) and reference text (t_r). It is a modified form of precision (P) and it is defined in (11).

$$P = \frac{m}{w_t} \tag{11}$$

where m is the number of candidate translation words occurring in reference t_r and w_t is the number of words in t_c. The method is extended in BELU-1,2,3, and 4. based on the n-gram representation of the words.

Open Source Implementations: There exist a few open source libraries for using different topic-based modelling methods. Majority of these methods are built for text-based analysis of different data. Though none of them is built for video-based analysis, however, it can be easily adopted in video analysis. TOM (Topic modelling) [47] is a Python 3 library mainly focused on different variations of LDA. It includes methods of automatic number of topic identification. A library for different topic models is also included in R software[1]. This library focuses on different variations of LDA. Another popular open source library gensim[2] can be easily integrated with Python. This is a robust library that provides a suite of tools for implementing LSA, LDA, and other topic modelling algorithms.

Benchmark Results: Here, we summarize the performance of different topic-based methods used in different applications and datsets. We have arranged them according to the year of publication (2015-2020). We note that different model uses different dataset and metric depending on the underlying applications. First, we categorized the models into six major applications, namely anomaly detection, trajectory clustering and activity modeling, action recognition, video description, semantic recognition, and video classification. Next, we extend the analysis based on the methods, datasets, metric, and performance. The report is summarized in Table 8.

Table 8. Benchmark results of different topic-based video analysis methods with varying datasets

Application	Reference	Method	Metric	Dataset	Results	Remarks
Anamoly Detection	[2]	Topic Related Sparse Topical Coding (TRSTC)	Accuracy	QMUL Junction QMUL Roundabout	0.86 0.98	Number of topic is 20
	[113]	pLSA based anomaly detection	AUC	AVSS Dataset	0.75	
	[148]	Semi-supervised sparse topic model	AUC	QMUL Junction AVSS	0.93 0.95	
Trajectory Clustering and Activity Mining	[4]	Graph-based Topic Model based on LDA	Correctness completeness	CUHK	0.80 0.87	Topics varying from 2 to 20
	[7]	Unsupervised Bayesian Clustering based on Dirichlet Process Mixture (DPM)	Accuracy and AMI	Highway dataset	Accuracy 0.97 AMI 0.76	Topic range 2 to 14
	[36]	Dynamical causal topic model (DCTM) based on LDA	log likelihood convergence	QMUL Junction	$-3.6x10^7$	Topic is set to 22
	[57]	Dynamic Topic Model based on Markov Clustering Topic Model (MCTM)	TPR FPR	QMUL Junction i-LIDS MIT Traffic Dataset	TPR: 52% FPR: 1% TPR: 53% FPR: 11% TPR: 27% FPR: 0.6%	Semi-supervised method
	[65]	Dynamic Topic Modeling based on LDA	AUC Accuracy	QMUL Junction	AUC: 0.32 Accuracy: 0.95	Number of topics and behaviors are set to 8 and 4
	[83]	Modified Dirichlet Process Mixture Model	Accuracy	QMUL Junction VIRAT MIT	0.78 0.99 0.99	Clustering-based approach
	[187]	Locally Consistent Latent Dirichlet Allocation based on LDA	Accuracy	QMUL Junction GCS	0.97 0.94	3D SHIFT feature is used for clustering
Action Recognition	[16]	LDA based Type-2 Fuzzy Model	Accuracy	KTH UCF	0.90 0.86	Codewords varying from 500 to 2500
	[33]	Spatio-temporal Interest Points and PLSA	Accuracy	CASIA	0.86	9-bin 3DHOG and 5-bin HOF are used
	[78]	Codewords based LDA	Accuracy	INRIA surgery dataset IXMAS NTSEL MPII	0.80 0.95 0.91 0.62	Use feature reduction method
	[79]	Supervised LDA	Accuracy	UCF101	0.70	Feature size varying from 1 to 300
	[94]	Gaussian Process based HDP	Accuracy	QMUL Junction MIT Traffic	0.98 0.96	
	[96]	LDA based two-level beta HMM	Accuracy	Weizmann KTH UCF Sports HMDB51	0.98 0.96 0.93 0.66	SVM based classifier is used
	[139]	Multi-label hierarchical Dirichlet process	Accuracy	KTH UCF101	0.96 0.89	
	[168]	Multi-Feature Max-Margin Hierarchical Bayesian Model	Accuracy	KTH UCF Sports	0.98 0.94	3D SHIFT feature is used

[1] https://www.tidytextmining.com/topicmodelling.html
[2] https://radimrehurek.com/gensim/

Video Description	[20]	LDA based topic-guided model (TGM)	BELU-4	MSR-VTT	0.44	Topic settings 10, 20, and 30
	[21]	Latent topic-guided model (LTGM)	BELU-4	MSR-VTT	0.49	Topic settings 10, 20, and 30
Semantic Recognition	[38]	Deep Feature LDA (DF-LDA)	Accuracy	UIUC Sports	0.90	Topic varying from 10 to 150
	[88]	LDA based clustering	mAP	Trecvid2013	0.33	Topic settings 100,150,200, and 300
Classification	[59]	LDA based multi-layer multi-view topic model	Mean Accuracy	Web videos	0.82	Topic varying from 10 to 150

5 CONCLUDING REMARKS

This review mainly focuses on recent uses of topic models in video surveillance applications. The following points have been summarized from the study:

(1) Co-related topic models have been proposed to maintain interactions between topics. Several models have been proposed to improving the state-of-the-art LDA.
(2) Dynamic topic models have been introduced in video analysis to discover how topics evolve over time.
(3) Objects can be tracked and represented by motion features. In such cases, video clips are treated as documents, moving pixels are treated as words, and action classes are noted as topics.
(4) Deep learning-guided topic models are not fully explored in video analysis.

Quality and Time Management: For better visual pattern representation and to effectively model scene fragmentation, quality measurement may be needed to maintain spatio-temporal co-occurrences and their graphical associations to develop efficient implementations of visual patterns. Many of the dynamic topic model algorithms use Expectation-Maximization algorithm within the spatio-temporal learning frameworks. However, these algorithms run slower when the videos are large. This results in a slow convergence rate to the posterior distributions under consideration. Better algorithms are needed to overcome such computational problems.

Information Embedding: Unlike the language models, video analysis does not have any predefined words. Hence a global embedding method like word to vector is not suitable. This drawback leads to difficulties in semantic measurement. Even though several languages and vision combined approaches have partially solved the problem, however, it is still an open issue to design such information embedding methods applicable to topic-based video analysis.

Challenge of Large Camera Network: Topic models have been successfully experimented and applied to mine activities over a small camera network (with less than ten cameras). However, some video surveillance applications, such as monitoring activities and traffic flows in large cities or human behaviors in crowded places, require human actions under large camera networks.

Fusion of Multiple Models: Though several variations of topic-based models have been used in surveillance video analysis, however, only a few of them work by fusing multiple methods. In a complex environment, multi-model fusion can be explored for unsupervised analysis.

Deep Topic Models: Though a few variations of topic-based deep neural networks have been used in various video analysis, however, the potentials are not explored fully. It has been observed that the non-parametric topic model such as HPD can be used in unsupervised learning. It has also been understood that topic models can be used to design explainable AI models due to the inherent capability of expressing data by topics.

REFERENCES

[1] Parvin Ahmadi, Iman Gholampour, and Mahmoud Tabandeh. 2017. A new two-stage topic model based framework for modeling traffic motion patterns. In *Iranian Conference on Machine Vision and Image Processing*. IEEE, 276–280.
[2] Parvin Ahmadi, Iman Gholampour, and Mahmoud Tabandeh. 2018. Employing Topical Relations in Semantic Analysis of Traffic Videos. *IEEE Intelligent Systems* 34, 1 (2018), 3–13.
[3] John Aitchison and CH Ho. 1989. The multivariate Poisson-log normal distribution. *Biometrika* 76, 4 (1989), 643–653.

[4] Manal Al Ghamdi and Yoshihiko Gotoh. 2020. Graph-based topic models for trajectory clustering in crowd videos. *Machine Vision and Applications* 31, 5 (2020), 1–13.

[5] Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications* 6, 1 (2015), 147–153.

[6] Saad Ali and Mubarak Shah. 2007. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–6.

[7] Vahid Bastani, Lucio Marcenaro, and Carlo S Regazzoni. 2016. Online nonparametric bayesian activity mining and analysis from surveillance video. *IEEE Transactions on Image Processing* 25, 5 (2016), 2089–2102.

[8] Ben Benfold and Ian Reid. 2011. Stable multi-target tracking in real-time surveillance video. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3457–3464.

[9] David Blei, Lawrence Carin, and David Dunson. 2010. Probabilistic topic models. *IEEE Signal Processing Magazine* 27, 6 (2010), 55–65.

[10] David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

[11] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *International Conference on Machine learning*. ACM, 113–120.

[12] David M Blei and John D Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics* (2007), 17–35.

[13] David M Blei, Andrew Y Ng, and Michael I Jordan. 2002. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*. 601–608.

[14] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

[15] Jordan L Boyd-Graber and David M Blei. 2009. Syntactic topic models. In *Advances in Neural Information Processing Systems*. 185–192.

[16] Xiao-Qin Cao and Zhi-Qiang Liu. 2015. Type-2 fuzzy topic models for human action recognition. *IEEE Transactions on Fuzzy Systems* 23, 5 (2015), 1581–1593.

[17] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In *AAAI Conference on Artificial Intelligence*.

[18] Baitong Chen, Satoshi Tsutsui, Ying Ding, and Feicheng Ma. 2017. Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics* 11, 4 (2017), 1175–1189.

[19] Qian Chen, Ni Ai, Jie Liao, Xin Shao, Yufeng Liu, and Xiaohui Fan. 2017. Revealing topics and their evolution in biomedical literature using Bio-DTM: a case study of ginseng. *Chinese Medicine* 12, 1 (2017), 27.

[20] Shizhe Chen, Jia Chen, and Qin Jin. 2017. Generating video descriptions with topic guidance. In *ACM on International Conference on Multimedia Retrieval*. 5–13.

[21] Shizhe Chen, Qin Jin, Jia Chen, and Alexander G Hauptmann. 2019. Generating video descriptions with latent topic guidance. *IEEE Transaction Multimedia* 21, 9 (2019), 2407–2418.

[22] Yu Chen, Tom Diethe, and Peter Flach. 2016. ADLTM: A Topic Model for Discovery of Activities of Daily Living in a Smart Home. (2016).

[23] Yuhao Chen, Ming Yang, Chunxiang Wang, and Bing Wang. 2019. 3D Semantic Modelling With Label Correction For Extensive Outdoor Scene. In *IEEE Intelligent Vehicles Symposium*. IEEE, 1262–1267.

[24] Jen-Tzung Chien and Meng-Sung Wu. 2008. Adaptive Bayesian latent semantic analysis. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 1 (2008), 198–207.

[25] Ayesha Choudhary, Manish Pal, Subhashis Banerjee, and Santanu Chaudhury. 2008. Unusual activity analysis using video epitomes and pLSA. In *Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 390–397.

[26] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2634–2641.

[27] Sokemi Rene Emmanuel Datondji, Yohan Dupuis, Peggy Subirats, and Pascal Vasseur. 2016. A survey of vision-based traffic monitoring of road intersections. *IEEE Transactions on Intelligent Transportation Systems* 17, 10 (2016), 2681–2698.

[28] NA Deepak and UN Sinha. 2016. Analysis of Human Gait for Person Identification and Human Action Recognition. *Analysis* 4, 4 (2016).

[29] Mohamed Dermouche, Julien Velcin, Leila Khouas, and Sabine Loudcher. 2014. A joint model for topic-sentiment evolution over time. In *IEEE International Conference on Data Mining*. IEEE, 773–778.

[30] Adji B Dieng, Chong Wang, Jianfeng Gao, and John Paisley. 2016. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702* (2016).

[31] M Divya, K Thendral, and S Chitrakala. 2013. A Survey on Topic Modeling. *International Journal of Recent Advances in Engineering & Technology* 1 (2013), 57–61.

[32] Yinpeng Dong, Hang Su, Jun Zhu, and Bo Zhang. 2017. Improving interpretability of deep neural networks with semantic information. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4306–4314.

[33] Ke Du, Ying Shi, Bowen Lei, Jie Chen, and Mingjun Sun. 2016. A method of human action recognition based on spatio-temporal interest points and PLSA. In *2016 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration*. IEEE, 69–72.

[34] Liang Du, Haitao Lang, Ying-Li Tian, Chiu C Tan, Jie Wu, and Haibin Ling. 2016. Covert Video Classification by Codebook Growing Pattern. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 11–18.

[35] Paul Duckworth, Muhannad Al-Omari, James Charles, David C Hogg, and Anthony G Cohn. 2017. Latent Dirichlet Allocation for Unsupervised Activity Analysis on an Autonomous Mobile Robot.. In *AAAI*. 3819–3826.

[36] Yawen Fan, Quan Zhou, Wenjing Yue, and Weiping Zhu. 2017. A dynamic causal topic model for mining activities from complex videos. *Multimedia Tools and Applications* (2017), 1–16.

[37] Li Fei-Fei and Pietro Perona. 2005. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2. IEEE, 524–531.

[38] Jiangfan Feng and Amin Fu. 2018. Scene semantic recognition based on probability topic model. *Information* 9, 4 (2018), 97.

[39] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. 2013. Learning multimodal latent attributes. *IEEE transactions on pattern analysis and machine intelligence* 36, 2 (2013), 303–316.

[40] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. 2014. Learning multimodal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 2 (2014), 303–316.

[41] Laya Elsa George and Lokendra Birla. 2018. A Study of Topic Modeling Methods. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 109–113.

[42] Yogesh Girdhar, Walter Cho, Matthew Campbell, Jesus Pineda, Elizabeth Clarke, and Hanumant Singh. 2016. Anomaly detection in unstructured environments using bayesian nonparametric scene modeling. In *IEEE International Conference on Robotics and Automation*. IEEE, 2651–2656.

[43] Shaogang Gong and Tao Xiang. 2011. *Visual analysis of behaviour: from pixels to semantics*. Springer Science & Business Media.

[44] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. 2007. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence* 29, 12 (2007), 2247–2253.

[45] Tom Griffiths. 2002. Gibbs sampling in the generative model of latent dirichlet allocation. (2002).

[46] Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *Artificial Intelligence and Statistics*. 163–170.

[47] Adrien Guille and Edmundo-Pavel Soriano-Morales. 2016. TOM: A library for topic modeling and browsing.. In *EGC*. 451–456.

[48] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. 2014. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM international conference on Multimedia*. 17–26.

[49] Avishai Hendel, Daphna Weinshall, and Shmuel Peleg. 2010. Identifying surprising events in videos using bayesian topic models. In *Asian Conference on Computer Vision*. Springer, 448–459.

[50] Yan Heng, Zhifeng Gao, Yuan Jiang, and Xuqi Chen. 2018. Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach. *Journal of Retailing and Consumer Services* 42 (2018), 161–168.

[51] Joao F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2015. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (2015), 583–596.

[52] Geoffrey E Hinton and Ruslan R Salakhutdinov. 2009. Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems*. 1607–1614.

[53] Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*. 856–864.

[54] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *International ACM Conference on Research and Development in Information Retrieval*. ACM, 50–57.

[55] Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 42, 1 (2001), 177–196.

[56] Timothy Hospedales, Shaogang Gong, and Tao Xiang. 2009. A markov clustering topic model for mining behaviour in video. In *International Conference on Computer Vision*. IEEE, 1165–1172.

[57] Timothy Hospedales, Shaogang Gong, and Tao Xiang. 2012. Video behaviour mining using a dynamic topic model. *International Journal of Computer Vision* 98, 3 (2012), 303–323.

[58] Timothy M Hospedales, Jian Li, Shaogang Gong, and Tao Xiang. 2011. Identifying rare and subtle behaviors: A weakly supervised joint topic model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 12 (2011), 2451–2464.

[59] Sujuan Hou, Ling Chen, Dacheng Tao, Shangbo Zhou, Wenjie Liu, and Yuanjie Zheng. 2017. Multi-layer multi-view topic model for classifying advertising video. *Pattern Recognition* 68 (2017), 66–81.

[60] Chun-Hao Huang, Edmond Boyer, Nassir Navab, and Slobodan Ilic. 2014. Human shape and pose tracking using keyframes. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3446–3453.

[61] Michael C Hughes. 2010. Supervised topic models for video activity recognition. *Unpublished manuscript* (2010).

[62] Jing Huo, Yang Gao, Yinghuan Shi, and Hujun Yin. 2018. Cross-modal metric learning for auc optimization. *IEEE transactions on neural networks and learning systems* 29, 10 (2018), 4844–4856.

[63] Tâm Huynh, Mario Fritz, and Bernt Schiele. 2008. Discovery of activity patterns using topic models. In *International Conference on Ubiquitous Computing*. ACM, 10–19.

[64] Olga Isupova, Danil Kuzin, and Lyudmila Mihaylova. 2015. Abnormal behaviour detection in video using topic modeling. In *USES Conference Proceedings*. The University of Sheffield.

[65] Olga Isupova, Danil Kuzin, and Lyudmila Mihaylova. 2018. Learning methods for dynamic topic modeling in automated behavior analysis. *IEEE Transactions on Neural Networks and Learning Systems* 29, 9 (2018), 3980–3993.

[66] Olga Isupova, Lyudmila Mihaylova, Danil Kuzin, Garik Markarian, and Francois Septier. 2015. An expectation maximisation algorithm for behaviour analysis in video. In *International Conference on Information Fusion*. IEEE, 126–133.

[67] Rahul Radhakrishnan Iyer, Sanjeel Parekh, Vikas Mohandoss, Anush Ramsurat, Bhiksha Raj, and Rita Singh. 2016. Content-based video indexing and retrieval using corr-lda. *arXiv preprint arXiv:1602.08581* (2016).

[68] Anil K Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31, 8 (2010), 651–666.

[69] V Jelisavčić, Bojan Furlan, Jelica Protić, and Veljko Milutinović. 2012. Topic models and advanced algorithms for profiling of knowledge in scientific papers. In *International Convention MIPRO*. IEEE, 1030–1035.

[70] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78, 11 (2019), 15169–15211.

[71] Hawook Jeong, Youngjoon Yoo, Kwang Moo Yi, and Jin Young Choi. 2014. Two-stage online inference model for traffic pattern analysis and anomaly detection. *Machine vision and applications* 25, 6 (2014), 1501–1517.

[72] Longlong Jing and Yingli Tian. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[73] Peiguang Jing, Yuting Su, Liqiang Nie, Xu Bai, Jing Liu, and Meng Wang. 2017. Low-rank multi-view embedding learning for micro-video popularity prediction. *IEEE Transactions on Knowledge and Data Engineering* 30, 8 (2017), 1519–1532.

[74] Michael I Jordan. 2010. Bayesian nonparametric learning: Expressive priors for intelligent systems. *Heuristics, probability and causality: A tribute to Judea Pearl* 11 (2010), 167–185.

[75] Arnold Kalmbach, Maia Hoeberechts, Alexandra Branzan Albu, Hervé Glotin, Sébastien Paris, and Yogesh Girdhar. 2016. Learning deep-sea substrate types with visual topic models. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1–9.

[76] Zenun Kastrati, Ali Shariq Imran, and Arianit Kurti. 2019. Integrating word embeddings and document topics with deep learning in a video classification framework. *Pattern Recognition Letters* 128 (2019), 85–92.

[77] Hirokatsu Kataoka, Yoshimitsu Aoki, Kenji Iwata, and Yutaka Satoh. 2015. Evaluation of vision-based human activity recognition in dense trajectory framework. In *International Symposium on Visual Computing*. Springer, 634–646.

[78] Hirokatsu Kataokai, Kenji Iwata, Yutaka Satoh, Masaki Hayashi, Yoshimitsu Aok, and Slobodan Ilic. 2016. Dominant Codewords Selection with Topic Model for Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 65–72.

[79] Angelos Katharopoulos, Despoina Paschalidou, Christos Diou, and Anastasios Delopoulos. 2016. Fast supervised lda for discovering micro-events in large-scale video datasets. In *ACM international conference on Multimedia*. 332–336.

[80] Sayed Hossein Khatoonabadi and Ivan V Bajic. 2013. Video object tracking in the compressed domain using spatio-temporal Markov random fields. *IEEE Transactions on Image Processing* 22, 1 (2013), 300–313.

[81] Jaechul Kim and Kristen Grauman. 2009. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2921–2928.

[82] Patrik Ehrencrona Kjellin and Yan Liu. 2016. A Survey On Interactivity in Topic Models. *International Journal of Advanced Computer Science and Applications* 7, 4 (2016), 456–461.

[83] Santhosh Kelathodi Kumaran, Adyasha Chakravarty, Debi Prosad Dogra, and Partha Pratim Roy. 2019. Likelihood learning in modified Dirichlet Process Mixture Model for video analysis. *Pattern Recognition Letters* 128 (2019), 211–219.

[84] Lakhdar Laib, Mohand Said Allili, and Samy Ait-Aoudia. 2019. A probabilistic topic model for event-based image classification and multi-label annotation. *Signal Processing: Image Communication* 76 (2019), 283–294.

[85] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Conference of the European Chapter of the Association for Computational Linguistics*. 530–539.

[86] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[87] Sangno Lee, Jeff Baker, Jaeki Song, and James C Wetherbe. 2010. An empirical comparison of four text mining methods. In *International Conference on System Sciences*. IEEE, 1–10.

[88] Haojie Li, Lijuan Liu, Fuming Sun, Yu Bao, and Chenxin Liu. 2016. Multi-level feature representations for video semantic concept detection. *Neurocomputing* 172 (2016), 64–70.

[89] Jian Li, Shaogang Gong, and Tao Xiang. 2008. Global Behaviour Inference using Probabilistic Latent Semantic Analysis.. In *British Machine Vision Conference*, Vol. 3231. 3232.

[90] Jian Li, Shaogang Gong, and Tao Xiang. 2012. Learning behavioural context. *International journal of computer vision* 97, 3 (2012), 276–304.

[91] Jian Li, Timothy M Hospedales, Shaogang Gong, and Tao Xiang. 2010. Learning rare behaviours. In *Asian Conference on Computer Vision*. Springer, 293–307.

[92] Li-Jia Li and Li Fei-Fei. 2007. What, where and who? classifying events by scene and object recognition. In *International conference on computer vision*. IEEE, 1–8.

[93] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. 2015. Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 3 (2015), 367–386.

[94] Wentong Liao, Bodo Rosenhahn, and Machael Yang. 2015. Video Event Recognition by Combining HDP and Gaussian Process. In *IEEE International Conference on Computer Vision Workshops*. 19–27.

[95] Chih-Ching Lin, Shwu-Huey Yen, and Ching-Ting Tu. 2017. Visual object tracking via LDA. In *International Conference on Applied System Innovation*. IEEE, 315–318.

[96] Lu Lu, Zhan Yi-Ju, Jiang Qing, and Cai Qing-Ling. 2017. Recognizing human actions by two-level Beta process hidden Markov model. *Multimedia Systems* 23, 2 (2017), 183–194.

[97] Wenhan Luo, Björn Stenger, Xiaowei Zhao, and Tae-Kyun Kim. 2015. Automatic topic discovery for multi-object tracking. In *AAAI Conference on Artificial Intelligence*.

[98] Guangyi Lv, Tong Xu, Enhong Chen, Qi Liu, and Yi Zheng. 2016. Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding. In *AAAI Conference on Artificial Intelligence*.

[99] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision*. 2794–2802.

[100] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *IEEE international conference on computer vision*. 2630–2640.

[101] Arjan Mieremet, Ivo Alberink, Bart Hoogeboom, and Derk Vrijdag. 2018. Probability intervals of speed estimations from video images: The Markov Chain Monte Carlo approach. *Forensic science international* 288 (2018), 29–35.

[102] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[103] Samaneh Moghaddam and Martin Ester. 2012. On the design of LDA models for aspect-based opinion mining. In *ACM International Conference on Information and Knowledge Management*. ACM, 803–812.

[104] Brendan Tran Morris and Mohan Trivedi. 2013. Understanding vehicular traffic behavior from video: a survey of unsupervised approaches. *Journal of Electronic Imaging* 22, 4 (2013), 041113.

[105] T Nathan Mundhenk, Daniel Ho, and Barry Y Chen. 2018. Improvements to context based self-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9339–9348.

[106] Shi-Yong Neo, Yuanyuan Ran, Hai-Kiat Goh, Yantao Zheng, Tat-Seng Chua, and Jintao Li. 2007. The use of topic evolution to help users browse and find answers in news video corpus. In *ACM International Conference on Multimedia*. ACM, 198–207.

[107] David Newman, Padhraic Smyth, Max Welling, and Arthur U Asuncion. 2008. Distributed inference for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*. 1081–1088.

[108] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. 2008. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision* 79, 3 (2008), 299–318.

[109] Zhenxing Niu, Gang Hua, Le Wang, and Xinbo Gao. 2018. Knowledge-Based Topic Model for Unsupervised Object Discovery and Localization. *IEEE Transactions on Image Processing* 27, 1 (2018), 50–63.

[110] Aytug Onan, Serdar Korukoglu, and Hasan Bulut. 2016. LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis. *International Journal of Computer Linguistics Applications* 7, 1 (2016), 101–119.

[111] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *IEEE conference on computer vision and pattern recognition*. 4594–4602.

[112] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[113] Deepak Pathak, Abhijit Sharang, and Amitabha Mukerjee. 2015. Anomaly localization in topic-based analysis of surveillance videos. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 389–395.

[114] Anastasia Podosinnikova, Francis Bach, and Simon Lacoste-Julien. 2015. Rethinking lda: moment matching for discrete ica. In *Advances in Neural Information Processing Systems*. 514–522.

[115] Oluwatoyin P Popoola and Kejun Wang. 2012. Video-based abnormal human behavior recognition—A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 865–878.

[116] Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image and Vision Computing* 28, 6 (2010), 976–990.

[117] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *ACM International Conference on Knowledge Discovery and Data Mining*. ACM, 569–577.

[118] Anderson Rocha, Walter Scheirer, Terrance Boult, and Siome Goldenstein. 2011. Vision of the unseen: Current trends and challenges in digital image and video forensics. *ACM Computing Surveys (CSUR)* 43, 4 (2011), 26.

[119] Filipe Rodrigues, Mariana Lourenco, Bernardete Ribeiro, and Francisco C Pereira. 2017. Learning supervised topic models for classification and regression from crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 12 (2017), 2409–2422.

[120] Mikel Rodriguez, Josef Sivic, Ivan Laptev, and Jean-Yves Audibert. 2011. Data-driven crowd analysis in videos. In *IEEE International Conference on Computer vision*. IEEE, 1235–1242.

[121] Sergio Rodríguez-Pérez and Raul Montoliu. 2013. Bag-of-words and topic modeling-based sport video analysis. In *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 189–196.

[122] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In *IEEE conference on computer vision and pattern recognition*. IEEE, 1194–1201.

[123] Xin Rong. 2014. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738* (2014).

[124] Lukas Rybok, Simon Friedberger, Uwe D Hanebeck, and Rainer Stiefelhagen. 2011. The kit robo-kitchen data set for the evaluation of view-based activity recognition systems. In *IEEE-RAS International Conference on Humanoid Robots*. IEEE, 128–133.

[125] Imran Saleemi, Khurram Shafique, and Mubarak Shah. 2009. Probabilistic modeling of scene dynamics for applications in visual surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 8 (2009), 1472–1485.

[126] Juan C SanMiguel, Andrea Cavallaro, and José M Martínez. 2012. Adaptive online performance evaluation of video trackers. *IEEE Transactions on Image Processing* 21, 5 (2012), 2812–2823.

[127] Kelathodi Kumaran Santhosh, Debi Prosad Dogra, and Partha Pratim Roy. 2018. Temporal unknown incremental clustering model for analysis of traffic surveillance videos. *IEEE Transactions on Intelligent Transportation Systems* 20, 5 (2018), 1762–1773.

[128] Kelathodi Kumaran Santhosh, Debi Prosad Dogra, Partha Pratim Roy, and Bidyut Baran Chaudhuri. 2019. Trajectory-Based Scene Understanding Using Dirichlet Process Mixture Model. *IEEE Transactions on Cybernetics* (2019).

[129] Christian Schuldt, Ivan Laptev, and Barbara Caputo. 2004. Recognizing human actions: a local SVM approach. In *International Conference on Pattern Recognition*, Vol. 3. IEEE, 32–36.

[130] Matthew W Segar, Kershaw V Patel, Colby Ayers, Mujeeb Basit, WH Wilson Tang, Duwayne Willett, Jarett Berry, Justin L Grodin, and Ambarish Pandey. 2020. Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis. *European journal of heart failure* 22, 1 (2020), 148–158.

[131] Giulia Slavic, Damian Campo, Mohamad Baydoun, Pablo Marin, David Martin, Lucio Marcenaro, and Carlo Regazzoni. 2020. Anomaly detection in video data based on probabilistic latent space models. In *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*. IEEE, 1–8.

[132] Angela A Sodemann, Matthew P Ross, and Brett J Borghetti. 2012. A review of anomaly detection in automated surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 1257–1272.

[133] Berkan Solmaz, Brian E Moore, and Mubarak Shah. 2012. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 10 (2012), 2064–2070.

[134] Khurram Soomro, Amir Roshan Zamir, and M Shah. 2012. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision* 2, 11 (2012).

[135] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 952–961.

[136] Xing Sun, Nelson HC Yung, Edmund Y Lam, and Hayden K-H So. 2016. Unsupervised tracking with a low computational cost using the doubly stochastic Dirichlet process mixture model. *Electronic Imaging* 2016, 14 (2016), 1–8.

[137] Yee Whye Teh and Michael I Jordan. 2010. Hierarchical Bayesian nonparametric models with applications. *Bayesian nonparametrics* 1 (2010), 158–207.

[138] Yee W Teh, David Newman, and Max Welling. 2007. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*. 1353–1360.

[139] Nguyen Anh Tu, Thien Huynh-The, Kifayat Ullah Khan, and Young-Koo Lee. 2018. ML-HDP: A Hierarchical Bayesian Nonparametric Model for Recognizing Human Actions in Video. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 3 (2018), 800–814.

[140] Jagannadan Varadarajan, Rémi Emonet, and Jean-Marc Odobez. 2013. A sequential topic model for mining recurrent activities from long term video logs. *International journal of computer vision* 103, 1 (2013), 100–126.

[141] Jagannadan Varadarajan and Jean-Marc Odobez. 2009. Topic models for scene analysis and abnormality detection. In *IEEE International Conference on Computer Vision Workshops*. IEEE, 1338–1345.

[142] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *ACM International Conference on Knowledge Discovery and Data Mining*. ACM, 448–456.

[143] Chong Wang, John Paisley, and David Blei. 2011. Online variational inference for the hierarchical Dirichlet process. In *International Conference on Artificial Intelligence and Statistics*. 752–760.

[144] He Wang and Carol O'Sullivan. 2016. Globally continuous and non-Markovian crowd activity analysis from videos. In *European Conference on Computer Vision*. Springer, 527–544.

[145] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*. 3551–3558.

[146] Hongxing Wang, Gangqiang Zhao, and Junsong Yuan. 2014. Visual pattern discovery in image and video data: a brief survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4, 1 (2014), 24–37.

[147] Jinqiao Wang, Wei Fu, Hanqing Lu, and Songde Ma. 2014. Bilayer sparse topic model for scene analysis in imbalanced surveillance videos. *IEEE Transactions on Image Processing* 23, 12 (2014), 5198–5208.

[148] Jun Wang, Limin Xia, Xiangjie Hu, and Yongliang Xiao. 2019. Abnormal event detection with semi-supervised sparse topic model. *Neural Computing and Applications* 31, 5 (2019), 1607–1617.

[149] Le Wang, Gang Hua, Rahul Sukthankar, Jianru Xue, and Nanning Zheng. 2014. Video object discovery and co-segmentation with extremely weak supervision. In *European Conference on Computer Vision*. Springer, 640–655.

[150] Tingwei Wang and Chuancai Liu. 2013. Human action recognition using supervised pLSA. *International Journal of Signal Processing, Image Processing and Pattern Recognition* 6, 4 (2013), 403–414.

[151] Wei Wang, Payam Mamaani Barnaghi, and Andrzej Bargiela. 2010. Probabilistic topic models for learning terminological ontologies. *IEEE Transactions on Knowledge and Data Engineering* 22, 7 (2010), 1028–1040.

[152] Xiaogang Wang, Keng Teck Ma, Gee-Wah Ng, and W Eric L Grimson. 2011. Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *International journal of computer vision* 95, 3 (2011), 287–312.

[153] Xiaogang Wang, Xiaoxu Ma, and W Eric L Grimson. 2009. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 3 (2009), 539–555.

[154] Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-Markov continuous-time model of topical trends. In *ACM International Conference on Knowledge Discovery and Data Mining*. ACM, 424–433.

[155] Yinying Wang, Alex J Bowers, and David J Fikis. 2017. Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of EAQ articles from 1965 to 2014. *Educational Administration Quarterly* 53, 2 (2017), 289–323.

[156] Zheng Wang, Jie Zhou, Jing Ma, Jingjing Li, Jiangbo Ai, and Yang Yang. 2020. Discovering attractive segments in the user-generated video streams. *Information Processing & Management* 57, 1 (2020), 102130.

[157] Daniel Weinland, Remi Ronfard, and Edmond Boyer. 2006. Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding* 104, 2-3 (2006), 249–257.

[158] FP Wheeler. 1998. Bayesian Forecasting and Dynamic Models (2nd edn). *Journal of the Operational Research Society* 49, 2 (1998), 179–180.

[159] Sinead Williamson, Chong Wang, Katherine Heller, and David Blei. 2010. The IBP compound Dirichlet process and its application to focused topic modeling. (2010).

[160] Chenxia Wu, Jiemi Zhang, Ozan Sener, Bart Selman, Silvio Savarese, and Ashutosh Saxena. 2018. Watch-n-patch: unsupervised learning of actions and relations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 2 (2018), 467–481.

[161] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.

[162] Xun Xu, Timothy M Hospedales, and Shaogang Gong. 2017. Discovery of shared semantic spaces for multiscene video query and summarization. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 6 (2017), 1353–1367.

[163] Junyu Xuan, Jie Lu, Guangquan Zhang, and Xiangfeng Luo. 2015. Topic model for graph mining. *IEEE Transactions on Cybernetics* 45, 12 (2015), 2792–2803.

[164] Jianfei Xue and Koji Eguchi. 2018. Sequential Bayesian nonparametric multimodal topic models for video data analysis. *IEICE Transactions on Information and Systems* 101, 4 (2018), 1079–1087.

[165] Jianfei Xue and Koji Eguchi. 2019. Supervised Nonparametric Multimodal Topic Models for Multi-class Video Classification. *ITE Transactions on Media Technology and Applications* 7, 2 (2019), 80–91.

[166] Michael Ying Yang, Wentong Liao, Yanpeng Cao, and Bodo Rosenhahn. 2018. Video Event Recognition and Anomaly Detection by Combining Gaussian Process and Hierarchical Dirichlet Process Models. *Photogrammetric Engineering & Remote Sensing* 84, 4 (2018), 203–214.

[167] Shuang Yang, Chunfeng Yuan, Weiming Hu, and Xinmiao Ding. 2014. A hierarchical model based on latent dirichlet allocation for action recognition. In *International Conference on Pattern Recognition*. IEEE, 2613–2618.

[168] Shuang Yang, Chunfeng Yuan, Baoxin Wu, Weiming Hu, and Fangshi Wang. 2015. Multi-feature max-margin hierarchical Bayesian model for action recognition. In *IEEE conference on computer vision and pattern recognition*. 1610–1618.

[169] Yang Yang, Jingen Liu, and Mubarak Shah. 2009. Video scene understanding using multi-scale analysis. In *IEEE International Conference on Computer Vision*. IEEE, 1669–1676.

[170] Litao Yu, Zi Huang, Jiewei Cao, and Heng Tao Shen. 2016. Scalable video event retrieval by visual state binary embedding. *IEEE Transactions on Multimedia* 18, 8 (2016), 1590–1603.

[171] Niange Yu, Xiaolin Hu, Binheng Song, Jian Yang, and Jianwei Zhang. 2018. Topic-oriented image captioning based on order-embedding. *IEEE Transactions on Image Processing* 28, 6 (2018), 2743–2754.

[172] Yin Yuan, Haomian Zheng, Zhu Li, and David Zhang. 2010. Video action recognition with spatio-temporal graph embedding and spline modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2422–2425.

[173] Yun Zhai and Mubarak Shah. 2006. Video scene segmentation using Markov chain Monte Carlo. *IEEE Transactions on Multimedia* 8, 4 (2006), 686–697.

[174] Jianguo Zhang and Shaogang Gong. 2010. Action categorization by structural probabilistic latent semantic analysis. *Computer Vision and Image Understanding* 114, 8 (2010), 857–864.

[175] Bin Zhao, Wei Xu, Genlin Ji, and Chao Tan. 2015. Discovering Topic Evolution Topology in a Microblog Corpus. In *International Conference on Advanced Cloud and Big Data*. IEEE, 7–14.

[176] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan. 2013. Relevance topic model for unstructured social group activity recognition. In *Advances in neural information processing systems*. 2580–2588.

[177] Liang Zhao, Lin Shang, Yang Gao, Yubin Yang, and Xiuyi Jia. 2013. Video behavior analysis using topic models and rough sets [applications notes]. *IEEE Computational Intelligence Magazine* 8, 1 (2013), 56–67.

[178] Zhicheng Zhao, Yifan Song, and Fei Su. 2016. Specific video identification via joint learning of latent semantic concept, scene and temporal structure. *Neurocomputing* 208 (2016), 378–386.

[179] Yin Zheng, Yu-Jin Zhang, and Hugo Larochelle. 2014. Topic modeling of multimodal data: an autoregressive approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1370–1377.

[180] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. 2011. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3441–3448.

[181] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. 2012. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2871–2878.

[182] Houkui Zhou, Huimin Yu, Roland Hu, Guangqun Zhang, Junguo Hu, and Tao He. 2019. Analyzing multiple types of behaviors from traffic videos via nonparametric topic model. *Journal of Visual Communication and Image Representation* 64 (2019), 102649.

[183] Qiqi Zhu, Yanfei Zhong, Liangpei Zhang, and Deren Li. 2017. Scene Classification Based on the Fully Sparse Semantic Topic Model. *IEEE Transactions on Geoscience and Remote Sensing* 55, 10 (2017), 5525–5538.

[184] Xudong Zhu and Hui Li. 2012. Unsupervised human action categorization using latent Dirichlet Markov clustering. In *International Conference on Intelligent Networking and Collaborative Systems*. IEEE, 347–352.

[185] Xudong Zhu and Zhijing Liu. 2011. Human behavior clustering for anomaly detection. *Frontiers of Computer Science in china* 5, 3 (2011), 279.

[186] Jialing Zou, Qixiang Ye, Yanting Cui, David Doermann, and Jianbin Jiao. 2014. A belief based correlated topic model for trajectory clustering in crowded video scenes. In *International Conference on Pattern Recognition*. IEEE, 2543–2548.

[187] Jialing Zou, Qixiang Ye, Yanting Cui, Fang Wan, Kun Fu, and Jianbin Jiao. 2016. Collective motion pattern inference via locally consistent latent dirichlet allocation. *Neurocomputing* 184 (2016), 221–231.