# Topic Models for Scene Analysis and Abnormality Detection*

Jagannadan Varadarajan[1,2], Jean-Marc Odobez[1]
[1]Idiap Research Institute, 1920, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne,1015,Lausanne, Switzerland
{vjagann, odobez}@idiap.ch

## Abstract

*Automatic analysis and understanding of common activities and detection of deviant behaviors is a challenging task in computer vision. This is particularly true in surveillance data, where busy traffic scenes are rich with multifarious activities many of them occurring simultaneously. In this paper, we address these issues with an unsupervised learning approach relying on probabilistic Latent Semantic Analysis (pLSA) applied to a rich set visual features including motion and size activities for discovering relevant activity patterns occurring in such scenes. We then show how the discovered patterns can directly be used to segment the scene into regions with clear semantic activity content. Furthermore, we introduce novel abnormality detection measures within the scope of the adopted modeling approach, and investigate in detail their performance with respect to various issues. Experiments on 45 minutes of video captured from a busy traffic scene and involving abnormal events are conducted.*

## 1. Introduction

Increasing needs for security applications have motivated the advancement of research in the area of visual surveillance systems in recent days. Because of the overwhelming amount of data from these surveillance systems, unsupervised methods with ideally no manual labeling are preferred. In this paper we address the problem of learning common patterns of activities occurring in a busy traffic scene. In such a scene, it is not easy to extract trajectories of individual objects due to frequent occlusions. Still, one would like to obtain dominant activity patterns occurring in the scene, segment the scene based on activities happening at each location, and detect abnormal events. In order to achieve this we use low level visual features extracted from the video as input to the pLSA topic model to find the spatio-temporal correlations among these features. We improve the method proposed by Wang *et al*. in [8] by adding object size information along with location and motion information, and present a novel way to segment the scene using the patterns discovered with pLSA directly as features. The obtained scene segments have semantic interpretation at each level of cluster size.

Using topic models like pLSA allows us to use different abnormality measures based on the interpretation of the model. We investigate the use of abnormality measures that include likelihood measure coming from the fitting of the learned topics to the test document, or the likelihood of measured topic distribution compared to those observed in the training data. As a novelty in this context, we explore the use of document reconstruction errors relying on distribution distances like Kullback-Leibler divergence or Bhattacharyya distance. Using 45 minutes of video captured from a busy traffic scene and involving abnormal activities (vehicles parked at wrong locations, or people crossing outside the zebra crossing), we compare the performance of the different measures by plotting recall-precision curves and discuss the effects of document size, object size features, and number of topics on the detection results. Results show the need for appropriate normalization of abnormality measures and that our abnormality measure performs as well as the more traditional likelihood fitting measure.

The rest of the paper is organized as follows.In the next section we present a brief survey of related work. In section 3, input features to pLSA are described. In section 4 the dataset, discovered activity patterns and scene segmentation are detailed. In section 5, we present various abnormality measures and comparative results. The conclusions are given in section 6.

## 2. Related Work

Methods proposed for activity analysis can be broadly classified into two categories. Under the first category, objects are first detected, tracked and the object trajectories are

*The authors gratefully acknowledge the financial support from Swiss National Science Foundation (Project: 200020-122062) under which this work was carried out.



2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops
978-1-4244-4441-0/09/$25.00 ©2009 IEEE

used for further analysis. Examples of this approach can be seen in [7, 2, 12, 10]. While good results are reported using this approach in uncluttered scenarios, it is sensitive to occlusion and tracking errors. The problem becomes more pronounced in the case of crowded scenes as frequent occlusions make reliable tracking an impossibility.

Under the second category, motion and appearance features are extracted from the video stream without tracking or object detection. The features thus extracted are directly used to create models of activities, but loss of tracking information makes it very difficult to separate different activities happening simultaneously. Recently, Topic models have been successfully used in getting semantic activity patterns from low-level feature co-occurrences. Wang *et al*. [8] introduced the use of location and optical flow features along with Hierarchical Bayesian approach to model activities and interactions. Li *et al*. [4] used spatio-temporal features along with a hierarchical pLSA for learning global behavior correlations. The method we propose is along the ideas given in [8], but we use a richer feature set by including object size, and propose a method of using the model for scene segmentation.

How is abnormality defined? Qualitatively, an abnormal (rare, unusual) event can be simply defined as "an action done at an unusual location, at an unusual time". Quantitatively, abnormality is defined in [6] in two ways, 1) Events that are fundamentally different in appearance, and 2) unusual order of events, where many of the events could be normal. Machine learning approaches to abnormality detection define them simply as behaviors that cannot be explained by the learnt models. In cases where the scene model is learnt by clustering trajectories [9, 2] abnormality is defined as an outlier trajectory i.e, when an object trajectory's distance to every cluster exceeds the intraclass distance of every cluster [2]. When activities are modelled as a sequence of observations as in [6], based on a set of observations $y_o, ...y_{t-1}$ a prior for observation $y_t$ is formed. After observing $y_t$, the posterior distribution is evaluated. The distance between the prior and posterior distributions is used as a criteria to identify anomalies. Xiang *et al*. in [12] propose a run-time accumulative anomaly measure based on likelihood obtained from the learned model. Zhong *et al*. in [13] build a database of spatio-temporal patches using normal behavior and detect those patterns that cannot be composed from the database as being abnormal. The work mentioned so far extract explicit object information to identify abnormalities.

Among the methods using low level visual features, Wang *et al*. [8] use likelihood measure calculated from the learnt model. But as simple motion features are used, it does not model activities of static objects in the scene. In [4], abnormalities are detected using an un-normalized likelihood measure. It was shown to work only with a single type of abnormality. While un-normalized measure can give good results when the documents are more or less of same size, they are prone to errors due to variations in document size. In our work, we show that due to the large variability in the video content, simple un-normalized measure does not give good results. Therefore we investigate possible abnormality measures within our modelling framework in order to understand the various aspects influencing a particular measure. In our experiments, we found that the normalized log-likelihood measure and a novel abnormality measure based on the Bhattacharyya distance between the raw word distribution and the reconstructed one using the learned topic distribution gives good performance.

## 3. From Visual Features to Activity Patterns

Topic models have shown good performance in modeling complex scenarios with a simple data representation. They were initially proposed to automatically discover the main themes or topics from large corpus of text documents, where a topic refers to a set of consistently co-occurring words in the text documents. In video analysis, these topics correspond to the activities that are frequently occurring in the scene, where the meaning of an activity depends on the visual words which have been used to build the documents. In the following, we first present the visual features used to characterize our scene content and build our documents, and then describe the pLSA topic model.

### 3.1. Visual words

To discover global activity patterns using pLSA, we need to define our vocabulary (the set of visual words characterizing the scene content), and how we build our video documents. In our case, a visual activity is described by three types of features: location, motion, and size features.

**Location:** In surveillance videos, most of the activities are characteristic of the place where they occur. Thus, location has to be taken into account when building our vocabulary, and we quantize a pixel position into non-overlapping cells of $10 \times 10$. Therefore for a video of dimension, $280 \times 360$ we obtain a set of $28 \times 36$ cells.

**Motion:** To identify the relevant parts of the scene, we first perform background subtraction using the algorithm proposed in [11] and detect the foreground pixels. For each of them, we also compute its optical flow using the Lucas-Kanade algorithm. Foreground pixels are categorized into static pixels (static label) and moving pixels by thresholding the magnitude of the optical flow vectors. Moving pixels are further differentiated by quantizing their motion direction into four labels (left, right, up, down) according to the intervals $(\frac{-\pi}{2}, \frac{\pi}{2}], (\frac{\pi}{2}, \frac{3\pi}{2}], (\frac{3\pi}{2}, \frac{-3\pi}{2}], (\frac{-3\pi}{2}, \frac{-\pi}{2}]$. Thus, in total, we have 5 possible motion words.

**Size:** To further characterize foreground objects, we asso-

ciate with each foreground pixel the size of the connected component it belongs to. In our dataset we observe that the foreground blobs can be roughly classified into two categories based on foreground blob size. The first one consists of small blobs corresponding mainly to pedestrians and the second one consists of large blobs corresponding to vehicles or group of pedestrians. Therefore, we apply a simple K-Means clustering on the extracted blob sizes with $K = 2$, and use the cluster number as a size word describing roughly the size of objects in the scene.

**Vocabulary:** Our vocabulary could be defined as the cartesian product of the location, motion, and size word spaces, leading to a total of $28 \times 36 \times 5 \times 2 = 10080$ words. However, while knowing the joint feature (motion,size) for each location might be desirable (for instance to distinguish between cars and people on zebra crossings), this results in a high dimensional vocabulary. As pLSA models word co-occurrences across documents, we expect that topics will capture separately people activity or car activity at a given location since they don't occur simultaneously. In other words, *given* an activity and location, we expect the motion and size to be independent, and thus we can simply concatenate them and define the set of words for a cell $c$, denoted by $V_c$, to be the concatenation of the motion and size words[1], leading to a codebook of $28 \times 36 \times (5 + 2)$ words only. Thus, a word can be denoted by $w_{c,a}$, where $c$ is the location and $a$ one of the seven characteristic labels.

**Documents:** They are built by dividing the video into short video clips, and count for each clip or document $d$ the number of times $\mathrm{n}(d, w)$ a word $w$ occurs in it to obtain the document bag-of-words representation. Henceforth, we use the terms document and clip interchangeably.

### 3.2. pLSA

Probabilistic latent space models [3], [1], [8] have been used to capture co-occurrence information between elements in a collection of discrete data in order to discover the recurrent topics in the collection. In our context, we expect such analysis to discover the main scene activities, where an activity mainly consists of the recurrent observation of the same motion and size words in scene regions. In this paper, we used the pLSA [3] model which originates from a statistical view of LSA. Although pLSA is a non-fully generative model, its tractable likelihood maximization makes it an interesting alternative to fully generative models like LDA [1] with comparative performance.

pLSA is a statistical model that associates a latent variable $z \in \mathcal{Z} = \{z_1, \ldots, z_{N_A}\}$ with each observation (occurrence of a word in a document). These variables, usually called topics, are then used to build a joint probability model

[1]This means that when constructing documents, a pixel will provide two words for the cell it belongs to: a motion word and a size word.

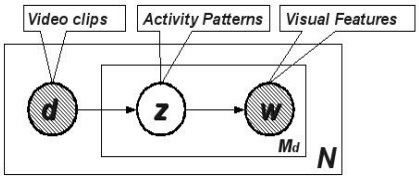

Figure 1. The Generative model of pLSA

over documents and words, defined as the mixture

$$P(w, d) = P(d)P(w|d) = P(d)\sum_{z=z_1}^{z_{N_A}} P(z|d)P(w|z). \quad (1)$$

pLSA introduces a conditional independence assumption, namely that the occurrence of a word $w$ is independent of the video document or clip $d$ it belongs to, given a topic $z$. The model in Eq. 1 is defined by the probability of a document $P(d)$, the conditional probabilities $P(w|z)$ which represent the probability of observing the word $w$ given the topic $z$, and by the document-specific conditional multinomial probabilities $P(z|d)$. The topic model decomposes the conditional probabilities of words in a document $P(w|d)$ as a convex combination of the topic specific word distributions $P(w|z)$, where the weights are given by the distribution of topics $P(z|d)$ in the document.

The parameters of the model are estimated using the maximum likelihood principle. More precisely, given a set of training documents $\mathcal{D}$, the log-likelihood of the model parameters $\Theta$ can be expressed by:

$$\mathcal{L}(\Theta|\mathcal{D}) = \sum_{d\in\mathcal{D}}\sum_{w} \mathrm{n}(d, w) \log\left(P(w|d)\right) \quad (2)$$

where the probability model is given by Eq. 1. The optimization is conducted using the Expectation-Maximization (EM) algorithm [3]. This estimation procedure allows to learn the topic distributions $P(w|z)$ representing the sought scene activities.

At test time, we are interested in estimating the weights $P(z|d)$ of the topics for a document $d$. This is achieved by running the EM algorithm keeping the learned model $P(w|z)$ fixed and maximizing the log likelihood of the words in the document:

$$L_d^u(P(z|d)) = \sum_{w} \mathrm{n}(d, w) \log\left(\sum_{z} P(z|d)P(w|z)\right) \quad (3)$$

## 4. Activity patterns and scene segmentation

To illustrate the pLSA modeling with the proposed features, we present some topics that were discovered by the approach and how it can be used to identify activities related to different object sizes or to segment the scene into different semantic regions.

Figure 2. Traffic Scene

### 4.1. Dataset

The approach can typically be used on outdoor video sequences. We applied it to videos capturing a portion of a busy traffic-controlled road junction. Sample frames are shown in Fig. 2. The scene has multiple activities that include people walking on the pavement or waiting for vehicles to cross over zebra crossings, and vehicles moving in and out of the scene in different directions. A video sequence of 45 minutes recorded at 25 Hz with frame size of $288 \times 360$ was captured and video clips of 5 seconds duration (125 frames) were defined as our documents. These clips were divided into a training dataset of 2210 video documents, and a test dataset of 320 clips.

### 4.2. Activity patterns

An activity like a vehicle moving on the road can be described by a set of motion and size features co-occurring over a sequence of locations. Similarly, a pedestrian standing at the foot path can be described by a co-occurring set of static pixels and size features. Thus, each activity pattern or a topic is a strongly co-occurring set of visual features represented by $p(w|z)$. To identify the set of locations which are mainly active for a given topic, we can marginalize the word distribution w.r.t. the words that occur at the same location. That is, we can plot the map defined for each cell $c$ by: $p(activity \in c|z) = \sum_{w \in V_c} p(w|z)$. Fig. 3(a)-(f) show the activity locations of selected topics highlighted.

**Size or Static related topics:** We can identify which of the extracted topics are more related to the activities of objects of small or large sizes by ranking the aspect according to the size probability obtained by marginalizing over the word 'small size' of every cell, i.e. by computing $p(size = small|z) = \sum_{w_{c,a}/a=small} p(w|z)$. For instance, Fig. 3(c)-(d) show the top two topics from 10 topics used to train pLSA involving small objects that correspond to pedestrians walking on the side-walk. A similar analysis can be done with static objects, and corresponding topics indicate pedestrians waiting to cross the road and cars waiting at the traffic light (Fig. 3(e) and (f)). Interestingly, note that the topic model was able to discover that during several parts of the junction traffic cycle, both pedestrian (bottom right) and cars (top right) needed to wait simultaneously.

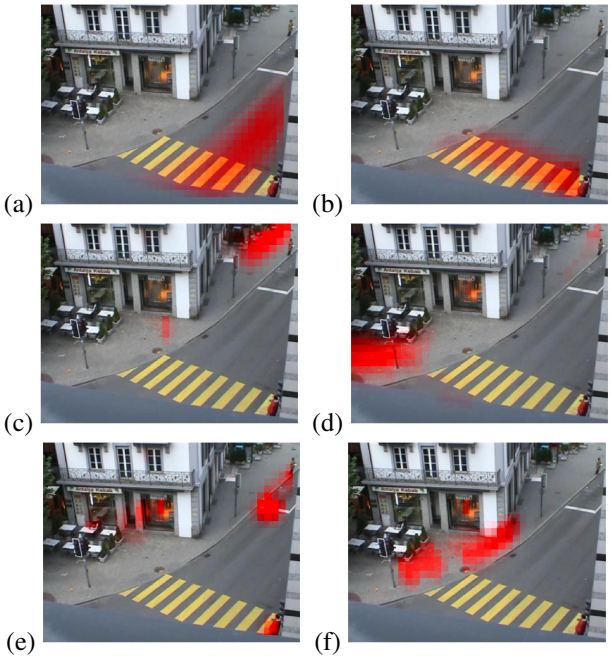


Figure 3. (a)-(b) Examples of common activity patterns (a) vehicles passing, (b) pedestrians crossing the road, (c)-(d) the first two topics involving small objects - pedestrians walking on the foot path, (e)-(f) the first two topics involving static pixels, (e) partially occluded vehicle waiting for signal (top right) with pedestrians waiting for signal at the bottom-right, (f) pedestrians waiting at the footpath for crossing the road.

### 4.3. Scene segmentation

Another way to investigate the learned topic is to segment the scene according to the extracted activities. Knowledge of the semantic scene regions could then provide context to the actions and thus help in understanding the intent of actions in a scene location. For example, in a typical traffic scene like in Figure 2, activities like pedestrians walking along the pavement, waiting at the zebra crossings are seen on the pedestrian side while vehicular movements are (in principle) only seen on the roads. An activity based segmentation achieves this by grouping parts of the scene into segments such that each segment corresponds to locations where similar semantic activities take place.

**Approach:** In [4], Li *et al*. represent the activity at a given location by the distribution of quantized spatio-temporal words that are observed at this location in the training data. In this paper, we propose instead to characterize a location by the set of *activities* that can occur at this pixel. This should lead to a less noisy representation, and implicitly incorporate temporal information as the activities model *observations which co-occur*, unlike raw feature distributions [4]. Activities at the cell location $c$ are represented by the topic distribution at this cell, denoted $P(z|c)$ and defined as $P(z|c) = P(z|V_c) \propto P(V_c|z) = \sum_{w \in V_c} P(w|z)$. In prac-

tice, we expect these distributions to smoothly evolve when the location $c$ moves along semantically similar regions (e.g. while moving along the same side of the road), and change abruptly when the location moves across some semantic border (e.g. moving from the road zone to the sidewalk region). Thus, clusters mainly correspond to smooth manifolds which can not be well represented using metric based clustering approaches like K-means. We used a spectral clustering algorithm [5] which have been shown to better capture such manifolds. It takes an input, an affinity matrix $A$ which in our case is given by

$$A_{c_i,c_j} = exp(\frac{-D^2_{Bhat}(P(z|c_i), P(z|c_j))}{2\sigma^2}) \quad (4)$$

where $D_{Bhat}$ denotes the Bhattacharyya distance used to compute the pairwise similarity between the two activity distributions at cell $c_i$ and $c_j$, and is defined by:

$$D_{Bhat}(P,Q) = \sqrt{1 - \sum_{x \in X} \sqrt{p(x).q(x)}}. \quad (5)$$

The scale $\sigma$ is taken to be the value that gives minimum cluster distortion [5].

**Results:** Figure 4 illustrates the results obtained when applying the algorithm with number of clusters equal to 2,3, 4 and 9, when the number of topics extracted with pLSA was 10. As can be seen, the results reveal that the number of clusters correspond to different level of details in interpreting the semantic activities in the scene. When $K = 2$, the algorithm segments the scene into regions of activity and no activity. When $K = 3$, the activity region is further divided into the pedestrian and vehicle regions. When $K = 4$, the road is split into the different sides of the road. When $K = 9$, further semantic regions like the region corresponding to zebra crossing, where both car and pedestrian motion can occur, or the different regions from where people come to cross the road (and wait) appear. Thus, we see that there is not a single valid value for $K$, but that each value lead to a scene segmentation with clear semantic interpretation.

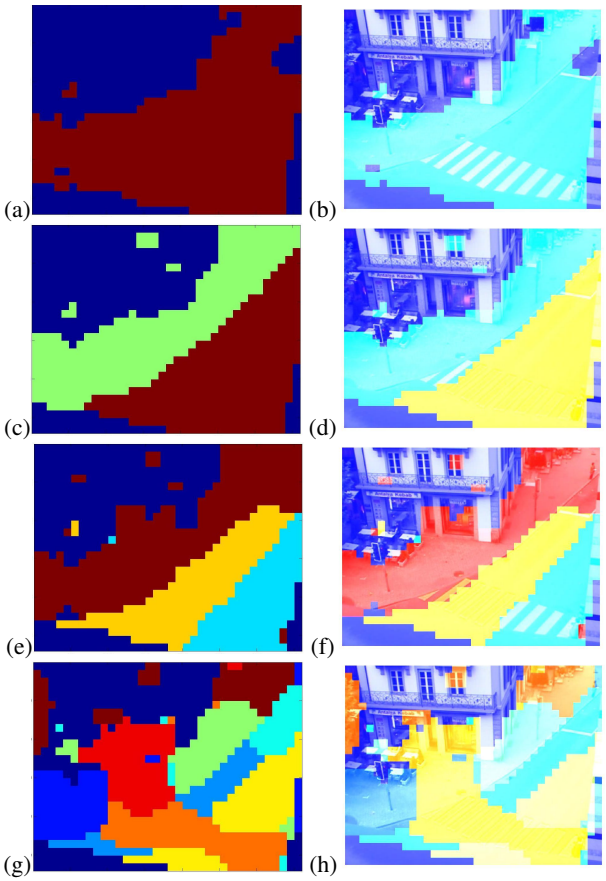

Figure 4. Semantic Scene Segmentation obtained from 10 pLSA Topics with the segments superimposed on the scene in the right: (a)-(b) 2 clusters (c)-(d) 3 clusters (e)-(f) 4 clusters, (g)-(h) 9 clusters

## 5. Abnormality detection

As discussed in the related work section, there exist several ways to define abnormality. In this Section, we present those that are pertaining to the topic model that we are using, and evaluate their performance on our dataset.

### 5.1. Abnormality measures

Modeling using a generative approach gives scope to use a variety of measures to identify unusual patterns in the data. But, little study has been done in comparing the different measures on the same task. Here, we present various possible measures that can be used based on the approach we consider, and evaluate the measures within the proposed framework to understand their merits and demerits.

**Fitting measures:** The estimation of the topic distribution $P(z|d)$ of a given clip is obtained by optimizing the log-likelihood function of Eq. 3. Thus, one natural way to consider if a clip is normal or abnormal is to use this log-likelihood measure $L^u_d(P(z|d))$ at the end of the fitting phase. If the activities happening within the clip corresponds to those observed in the training dataset, then the fitting should be able to find a suitable topic distribution explaining the bag-of-word representation of the clip. Thus, normal clips will generally provide high log-likelihood. On the other hand, if an abnormal activity is going on, none of the learned topic is able to explain the observed words of that activity, resulting in a low likelihood fit. In [4], this measure is used to find abnormal behavior correlations from a traffic scene.

The likelihood expression in Eq. 3 suffers from a severe drawback: it is not normalized and thus, whatever the quality of the fit, the measure is highly correlated with the doc-

ument size. To solve this issue, we can exploit the average log-likelihood of each word, by dividing $\mathrm{n}(d,w)$ by the number of words $\mathrm{n}_d = \sum_w \mathrm{n}(d,w)$ in Eq. 3, and get the normalized log-likelihood measure:

$$\begin{aligned} L_d^{nl}(P(z|d)) &= \sum_w \frac{\mathrm{n}(d,w)}{\mathrm{n}_d} \log \sum_z P(z|d)P(w|z) \\ &= \sum_w P_o(w|d) \log P_c(w|d) \end{aligned} \quad (6)$$

where $P_o(w|d) = \frac{\mathrm{n}(d,w)}{\mathrm{n}_d}$ is called the objective distribution as it is measured directly from the test document, and $P_c(w|d) = \sum_z P(z|d)P(w|z)$ is called the constrained distribution as it lies in the constrained simplex spanned by the topic distribution $P(w|z)$.

**Distribution reconstruction errors:** The goal of optimizing the likelihood function is to fit the constrained distribution to the objective distribution. Thus one possibility to evaluate the quality of the fitting is to measure the discrepancy between the two distributions. For instance, we could use the Kullback-Leibler divergence as the distance measure leading to:

$$\begin{aligned} L_d^{KL}(P(z|d)) &= KL[P_o(w|d)|P_c(w|d)] \\ &= -L_d^{nl}(P(z|d)) - H(P_o(w|d)) \end{aligned} \quad (7)$$

where $H(P_o(w|d))$ is the entropy of document $d$, which is a constant specific to each document. From this expression we note that the topic distribution $P(z|d)$ which maximizes the likelihood expression in Eq. 3 is actually the one that minimizes the KL distance $L^{KL}$. We can thus interpret the fitting as a document reconstruction process where the error in reconstruction is given by Eq. 7. Accordingly, we can use such reconstruction error measure as our abnormality measure. This also allows us to use other probability distances as abnormality measures. In this paper, we also used the Bhattacharyya distance given by Eq. 5 to compare $P_o$ and $P_c$, according to:

$$L_d^{Bh}(P(z|d)) = D_{Bhat}\left(P_o(w|d), P_c(w|d)\right) \quad (8)$$

**Scene topic abnormality measures:** Activities that can occur in a scene are characterized by the activity topics learned by the model. However, in general, not all combinations of activities are valid, that is, can occur simultaneously in the scene. This constraint can be taken into account by learning the allowed distribution of topics using a training dataset. In our case, this was done by fitting a Gaussian Mixture Model with $L$ mixtures to the topic distributions $P(z|d)$ extracted from the training document. Then, when considering a test document, we first estimate its topic distribution (by optimizing using Eq. 3), and then compute the likelihood of this distribution with the GMM to evaluate its validity. In this view, abnormal clips are outlier entities that have low likelihood of being generated by the $L$ GMM mixtures of the topic distribution.

## 5.2. Results and discussion

The different measures were applied to our test data, containing 140 normal activity documents, and 180 video clips corresponding to abnormal documents, where abnormality is defined as: people crossing the road at the wrong place (far away from zebra crossing), vehicle parked at the pedestrian path, or vehicles stopping ahead of the stop line while the stop sign is red. In the following experiments, unless stated otherwise, 20 topics were used to model the scene activities.

**Qualitative illustration.** The abnormality measures that we have defined allowed us to identify multiple instances of several abnormal events occurring both in isolation or simultaneously with other normal activities in the image. Fig. 5 shows the first video clips that were retrieved as abnormal using the normalized log-likelihood $L^{nl}$[2] measure. The object causing the abnormality is marked with red boxes for identification. Fig. 5(a)-(b) shows the event where a car is parked in the pedestrian foot path. In (b), additionally a pedestrian crosses the road in the wrong place. In Fig. 5(c)-(d) a car stops ahead of the stop line, and this stop is not due to stopped cars in front of it. In Fig. 5, (e)-(f) pedestrian were crossing the road away from the foot path.

**Quantitative evaluation.** Recall-Precision (RP) curves were considered to quantitatively assess the performance of the approach and compare the abnormality measures. Fig. 6 shows the RP curves for the Likelihood, Normalized Likelihood, KL-Divergence, Bhattacharyya distance and Topic-GMM likelihood abnormality measures. We first note that the performance obtained with the Topic-GMM likelihood abnormality measure has the least detection rate. Indeed, this approach only considers as normal the specific combination of activities observed in the training documents. Thus, as we only have around 35 minutes of video in the training data, there is very little chance to observe all common activity combinations, and thus this abnormality measure quickly performs randomly, i.e. according to the odds in the test set. As expected, the unnormalized likelihood measure too does not achieve good performance. The reconstruction error measure obtained from KL-divergence and Bhattacharya distance show better performance, but still not as good as the normalized likelihood measure, which achieves the best performance with good detection rates (with a precision of almost 1 for a recall of 50%).

Document size normalization. An analysis of the detection errors made using the likelihood measure, the KL-

[2]The Adaptive Bhattacharyya measure that we will describe below produced the same results.

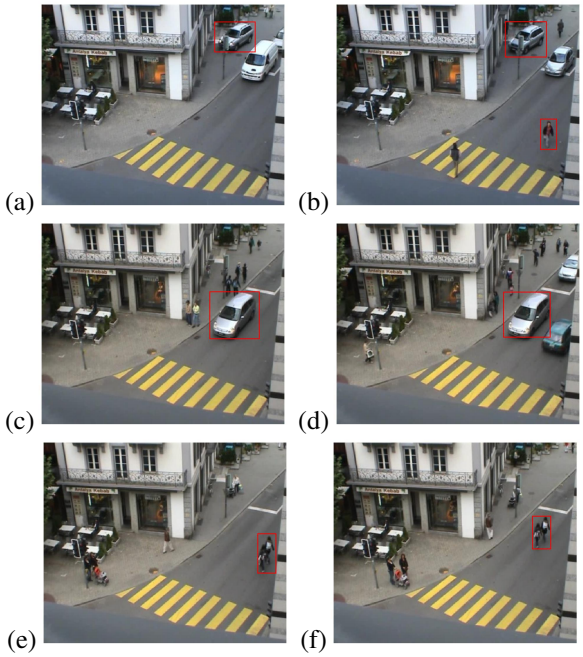


Figure 5. The top abnormal events retrieved using the normalized likelihood measure or the Adaptive Bhattacharyya distance measure. Note that, for illustration purposes, several abnormal documents corresponding to the same already displayed events have been omitted. (a)-(b) shows the event where a car is parked in the pedestrian foot path. (b) pedestrian crossing the road in the wrong place, (c)-(d) a car stopping ahead of the stop line. (e)-(f) pedestrian crossing the road away from the foot path.

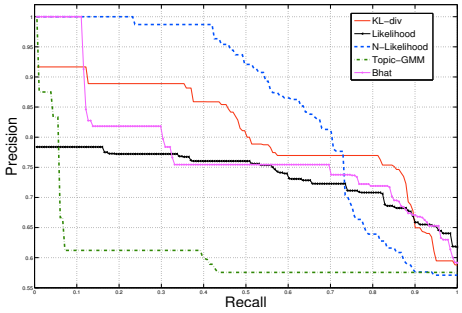


Figure 6. RP curves for Likelihood, Normalized Likelihood, KL-Divergence, Bhattacharyya distance and GMM likelihood.

divergence, and the Bhattacharyya distance, reveal that they are affected by document size or entropy. This is illustrated in Fig. 7, where we plot the Bhattacharya distance error measure as a function of the document size. As can be seen, smaller documents (with low entropy) tend to have higher reconstruction error, while larger documents (with high entropy) tend to have lower error. The normalized log-likelihood measure directly alleviates this effect by using the average word log-likelihood as abnormality measure. The KL-divergence measure can be normalized by removing the document specific entropy term. When this is done,

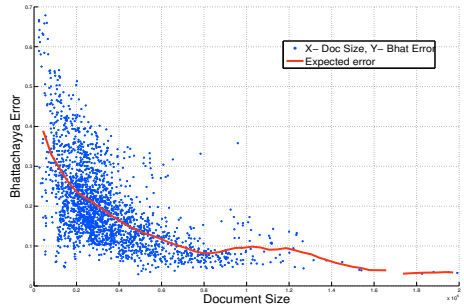


Figure 7. Bhattacharyya abnormality error measure vs Document Size: Scatter plot showing the relation between the Document size (number of words) and Bhattacharyya distance abnormality measure. The superimposed curve in red shows the expected Bhattacharyya error for a given size computed from the training data.

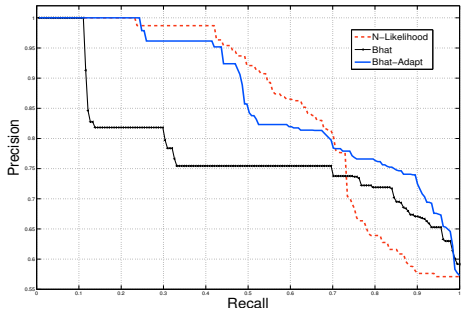


Figure 8. RP curves for Normalized Likelihood, Bhattacharyya distance and Adaptive Bhattacharyya distance.

we are left with the cross entropy term $H(P_o|P_c)$ given by:

$$H(P_o|P_c) = \sum_w P_o(w|d) log P_c(w|d) \qquad (9)$$

which is simply the normalized log-likelihood measure.
In the case of Bhattacharyya distance, such a direct normalization is not possible. Therefore we treat this bias by performing an adaptive normalization based on document size and learnt from the training data. For this, we construct a histogram of document size in the training set, and calculate for each bin the expected error measure for documents belonging to that bin (please see the red curve in Fig. 7). Then, for a test document, its reconstruction error using Bhattacharyya distance is normalized with the expected error according to its size before being compared with the abnormality threshold. Fig. 8 shows the results obtained after removal of the document size bias. As can be seen, this leads to a considerable improvement, and the Adaptive-Bhattacharyya abnormality measure performs now as well as the normalized log-likelihood measure, although with a different behavior. While the latter one performs better at medium recall, the Adaptive-Bhattacharyya succeeds to keep a precision significantly higher than random for very high recall.
Video Features. We also evaluate the effect of adding the size words in our description of activities, as compared to

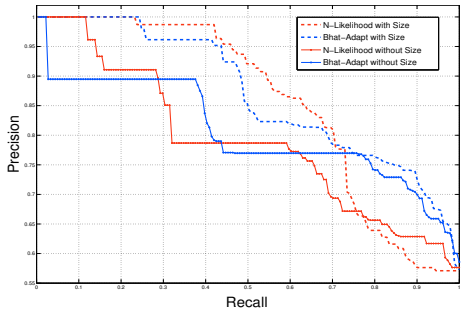


Figure 9. RP curves for Normalized Likelihood, Adaptive Bhattacharyya distance, with and without using the size words.

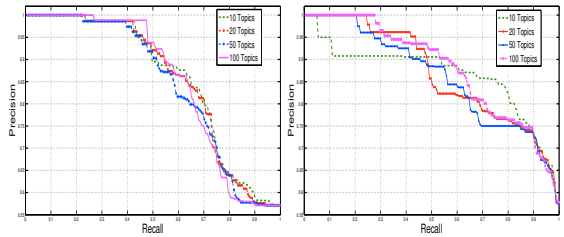


Figure 10. Effect of varying the number of topics: RP curves for Normalized Likelihood (left), and Adaptive Bhattacharyya distance (right) when using 10, 20, 50 and 100 Topics.

using just optical flow words as was done in [8]. This is displayed in Fig. 9, where the normalized log-likelihood and Adaptive Bhattacharyya distance abnormality RP curve are plotted with and without object size words. These curves show that the detection rates improve significantly when object size words are used as compared to just optical flow words.

Number of topics. Finally, Fig. 10 plots the RP curves for our two best measures when 10,20,50 and 100 topics were used (20 topics were used in the other curves) to model the different scene activities. As can be seen, the number of topics does not affect the results much. This is particularly true for the normalized likelihood measure, and when using more than 20 topics in the Adaptive Bhattacharyya case.

## 6. Conclusion

In this paper we have presented an unsupervised approach to activity analysis using pLSA. A novel scene segmentation based on the learned topics is proposed to localize and analyze the activity patterns, and results show that the obtained segmentation matches well with locations of semantic activities of the scene. A detailed investigation on various abnormality measures is presented. The results obtained from our experiments on a real dataset show that topic modeling approach is effective for abnormality deduction. They have highlighted the need for normalizing abnormality measures w.r.t. the document size, and, we believe, have provided greater insights into the merits and demerits of the abnormality measures, enabling one to choose the most appropriate method for the task. In the future, we would like to confirm our results with more datasets, and explore fusing the results from Normalized Likelihood and Adaptive Bhattacharyya measure to improve our results.

## References

[1] D. M. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Machine Learning Research*, (3):993–1022, 2003.

[2] A. Hervieu, P. Bouthemy, and J.-P. L. Cadre. A statistical video content recognition method using invariant features on object trajectories. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(11):1533–1543, 2008.

[3] T. Hofmann. Unsupervised learning by probability latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[4] J. Li, S. Gong, and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. In *British Machine Vision Conference*, 2008.

[5] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856, 2001.

[6] P.Cui, L. Sun, Z.Q.Liu, and S.Yang. A sequential montecarlo approach to anomaly detection in tracking visual events. *IEEE Workshop on Visual Surveillance*, pages 1–8, 2007.

[7] C. Stauffer and E. L.Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on PAMI*, 22:747–757, 2000.

[8] X. Wang, X. Ma, and E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. on PAMI*, 31(3):539–555, 2009.

[9] X. Wang, K. Tieu, and E. L. Grimson. Learning semantic scene models by trajectory analysis. *European Conference on Computer Vision*, 14(1):234–778, 2004.

[10] X. Wang, K. Tieu, and E. L. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Trans. on PAMI*, 1(1):893–908, 2009.

[11] J. Yao and J.-M. Odobez. Multi-layer background subtraction based on color and texture. In *IEEE International Conference in Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[12] E. E. Zelniniker, S. Gong, and T. Xiang. Global abnormal behaviour detection using a network of cctv cameras. In *The Eigth International Workshop on Visual Surveillance*, 2008.

[13] H. Zhong, S. Jianbo, and V. Mirko. Detecting unusual activity in video. *IEEE Conf. Computer Vision and Pattern Recognition*, 2(1):819–826, 2004.