Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Multiple Hierarchical Dirichlet Processes for Anomaly Detection in Traffic

Vagia Kaltsa[a,b,**], Alexia Briassouli[a], Ioannis Kompatsiaris[a], Michael G.Strintzis[b]

[a]*Centre for Research and Technology Hellas (CERTH), 6th km Charilaou-Thermi, Thessaloniki, Greece*
[b]*Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece*

ABSTRACT

This work introduces an unsupervised approach to scene analysis and anomaly detection in traffic video data, as captured from static surveillance cameras. A hybrid local-global scheme is introduced, so as to capture both local and global information, by extracting features in superpixel-generated spatiotemporal volumes, which are then merged into regions with dynamically varying boundaries. The resulting regions' shapes vary according to the underlying motion in the scene, as captured by the superpixels. Representative descriptors are then calculated in these regions, and multiple local Hierarchical Dirichlet Process (HDP) models are deployed in them, one for each region, for the unsupervised characterization of normal and "abnormal" events. The extraction of meaningful descriptors in these regions, instead of the whole frame, increases the resolution of the algorithm, while avoiding noise induced artifacts, and thus resulting in the successful detection of a wide range of "anomalies", both in the local and global scales. Experiments on benchmark datasets containing various scenarios in traffic scenes prove our method's efficacy and generality, leading to higher accuracy than the current State of the Art (SoA), and at a lower computational cost. Systematic quantitative experimental results and comparisons are provided on benchmark datasets, setting up a valuable baseline for future comparisons and improvements.

1. Introduction

Video analysis from surveillance systems is becoming increasingly important, as it can be used to discover and even avert abnormal and potentially dangerous events, such as accidents, natural disasters, terrorist acts etc. The manual analysis of surveillance videos is extremely time consuming and cumbersome, making the automatic detection of anomalous occurrences in videos of a long duration a necessity. The automated analysis of videos of crowded scenes, such as traffic environments, is particularly useful due to the complexity of such environments and the difficulties encountered in monitoring them, even for human observers. In this work, we address the problem of anomaly detection and scene understanding in real world traffic scenarios. This constitutes a difficult task, especially in real world traffic surveillance applications, as "anomalous" events are not known beforehand. Actually the purpose - and greatest challenge - of such monitoring is precisely to detect "abnormal" events that are not explicitly defined. In practice, anomalous events happen at unknown time instants, at unknown spatial locations and are not known, so there is no prior description of their characteristics or relevant descriptors. Moreover, an "anomaly" pattern in one video sequence may often be part of the "normal" pattern of another. This is common in traffic videos, where a particular vehicle's motion places its own restrictions on the rest of the scene. In order to address these issues, we define as "anomalies" the events that have a low probability of occurrence based on earlier observations. In structured traffic video data, being examined by this work, this can be interpreted as the set of all possible traffic violations. In addition to the unknown nature of the events themselves, difficulties arise due to scene diversity, occlusions, changes of illumination, adverse weather conditions, and sensor noise. Finally, computational efficiency is an additional factor, as it has to be kept within reasonable limits: in most applications, it is essential to detect suspicious events in time, so that security personnel can handle the situation efficiently.

In this work, we propose a novel framework for scene un-

**Corresponding author:
e-mail: vagiakal@iti.gr (Vagia Kaltsa)

derstanding and anomaly detection in traffic scenes, based on the deployment of multiple topic models in a hybrid, local-global setup. A common problem in this area arises due to the switching between local and global anomalies throughout the video. Most methods dealing with anomaly detection in different kinds of environments ((Bertini et al., 2012), (Boiman and Irani, 2007), (Calderara et al., 2011), and (Li et al., 2014)) focus exclusively on local statistics to infer about a frame's "irregularity", thus neglecting global patterns that may arise. However, this constitutes a very common situation in the examined traffic scenarios where many "anomalies" often emerge as a result of correlation patterns involving the entire frame. In order to capture these correlations, several works based on topic modeling have been proposed ((Wang et al., 2009), (Hospedales et al., 2012), (Isupova et al., 2016), (Hospedales et al., 2011), (Varadarajan et al., 2012), (Varadarajan et al., 2013), (Song et al., 2014)), (Kuettel et al., 2010)) examining traffic related scenarios. Nevertheless, in these works the extracted traffic models describe frame dyamics as a whole, restricting their algorithm's resolution, as important local information may be lost in the large bulk of data. So, even though hybrid approaches have been used in the past in anomaly detection tasks ((Kim and Grauman, 2009)), works examining traffic scenes based on topic modeling, often miss the local aspect.

In order to tackle this issue, we adopt a novel hybrid, topic-model based approach combining both local and global information. The scene is first divided into SLIC (Simple Linear Iterative Clustering) (Achanta et al., 2012) superpixels with homogeneous appearance characteristics, which are then clustered into regions with dynamically varying boundaries. Descriptors are efficiently extracted in these regions according to varying spatiotemporal tubes, as obtained from the combined action of superpixels clustering and tracking information in a proposed scheme. Subsequently, Multiple Hierarchical Dirichlet Processes (HDP) (Y.W. Teh and Blei, 2007) are deployed, one for each region separately, capturing **local** feature correlations and creating meaningful local topics. The resulting topic probability distributions are then used to **globally** reconstruct the frames under examination, and each reconstructed video subsequence's "abnormality" is assessed by a classifier based on its cosine difference from the original subsequence. Thus, videos can be characterized as normal or "abnormal", while preserving local information and at the same time smoothing out the influence of local noise, achieving a balance between the local and global.

The contribution of our work lies in the investigation of multiple topic models deployed in the scene, in an original framework, resulting in an effective fast algorithm capable of being used in real life surveillance scenarios. Superpixels and clustering have been used in the past for anomaly detection ((Cui et al., 2007), (Kwon and Lee, 2015)), however this work proposes a different deployment scheme. Particularly, the benefits of superpixels' structure are exploited in two different ways: a) they define region boundaries which vary over time according to the level of activity in them, and therefore visual events, and b) they constitute the structural elements of varying spatiotemporal tubes used for feature extraction, instead of the classic supervoxels commonly used in the literature. Furthermore, it is the first time that multiple HDP models are deployed and their combination is used as an inference about a frame's "abnormality", in contrast to many SoA works which propose variants, to address different categories of problems ((Wulsin et al., 2012), (Yakhnenko and Honavar)). Our contribution can be summarized as follows:

1. Dynamically varying spatiotemporal volumes are built based on the combined action of superpixels clustering and tracking, resulting in the extraction of informative descriptors.
2. Multiple local topic models (HDPs) are applied for the first time in the extracted regions of interest, to capture small scale interactions.
3. The outcomes of multiple HDPs are used in a frame reconstruction process, to infer about each frame's irregularity, thus integrating local and global information in a novel, hybrid framework.

As the experimental results show, the proposed method efficiently captures both local and global anomalies in real world challenging traffic scenes, outperforming state of the art algorithms in accuracy, and at a lower computational cost.

2. State of the art

Methods dealing with scene analysis and understanding from surveillance camera videos can be divided into two main categories.

The first one is comprised of methods based on trajectory extraction and processing. A work in this category is that of Saleemi et al. (Saleemi et al., 2009) where object trajectories are modeled using kernel density estimators to specify their multivariate nonparametric probability density function (pdf). A unified Markov Chain Monte Carlo (MCMC) sampling-based scheme is then used to generate the most likely paths. Anomalies are detected based on the estimated pdf of the next state by comparing the actual measurements of objects with the predicted tracks. A different approach is followed in (Jiang et al., 2011), where three different levels of semantics are considered after tracking all moving objects in the video. Rules of normal events are automatically extracted at each level and anomalies are defined as the events deviating from these rules. A two-stage inference model is used by (Jeong et al., 2014) for modeling trajectory patterns in a probabilistic framework. In (Yang et al., 2013) sub-trajectories are modeled by a sequence learning model, and multi-instance learning is applied in order to detect anomalies, while a tracking scheme involving pedestrian detection is proposed in (Yuan et al., 2015), for anomaly detection in crowded scenes based on spatio-temporal variations. Finally, in the work of (Piciarelli et al., 2008), trajectory clustering and a single class Support Vector Machine (SVM) are used for the identification of anomalous trajectories in surveillance video sequences, while a Dual Hierarchical Dirichlet Process (Dual-HDP) is proposed in (Wang et al., 2008) for trajectory and region modeling. These methods show some promising results, but are most applicable in scenarios

where trajectories are well defined and clearly visible. Their usability becomes limited in the case of more realistic traffic scenes where the motions are often highly dense, complex, with many occlusions and local noise. In those cases, extracted trajectories are susceptible to errors due to occlusions and local noise, while initial step in some of these approaches, object detection, may also be challenging and contain errors.

The second category, is defined by the recent trend of applying statistical models directly on raw image data, such as pixel location or other low level features. Non-parametric methods, including Gaussian Mixture Models, Dynamic Bayesian Networks or Probabilistic Topic Models, are used in the literature to capture spatiotemporal changes or find correlations among features, as in the case of topic models.

Wang et al. (Wang et al., 2009) use hierarchical Bayesian models to model scenes in two layers: atomic activities, represented as distributions over visual words, and interactions, represented as distribution over atomic activities. Typical and abnormal activities are then defined in a probabilistic framework, according to the Bayesian models. Despite promising results, that method relies on a single model for the whole frame, unlike the framework proposed in this work, while features are extracted around each moving pixel, instead of extracting informative pixel regions, thus facilitating the existence of noise artifacts. Furthermore, scalability issues arise due to its high computational cost, as the learning model requires about 12 hours to process a 1.5 hour video.

A new dynamic topic model, the Markov Clustering Topic Model (MCTM), is introduced by Hospedales et al. (Hospedales et al., 2012) while two new learning algorithms for the same model are proposed in (Isupova et al., 2016). The suggested model clusters visual events into activities and subsequently clusters activities into global behaviors, in order to identify salient events. Latent Dirichlet Allocation (LDA) is used in this case, which is less powerful than HDP used in our work, as it requires the prior specification of the number of events in the video sequences. Moreover, the use of a single Markov-chain to describe the whole video dynamics, assumes that all scenes are described by a global rule, making the method likely to miss local abnormalities and events.

A weakly supervised topic model is described in (Hospedales et al., 2011), in order to identify rare and subtle events in the scene. This work requires the use of clips of each class for training purposes, including training samples of anomalous events. However, in practical applications like traffic surveillance, there is a vast diversity of potentially abnormal events, which makes it impossible to find adequate instances of each class.

In another approach for abnormal event detection, Varadarajan et al. (Varadarajan et al., 2012) introduce a novel model that exploits temporal relationships between activity pairs, as well as global rules that are dominant in the scene, to effectively capture the scene's dynamical context. In (Varadarajan et al., 2013), temporal information is embedded in a topic based model called Probabilistic Latent Sequential Motifs (PLSM) to efficiently discover motifs of activities that occur concurrently in the scene while swarm intelligence was deployed by (Kaltsa et al., 2014) for anomaly detection in crowded scenarios. However, these works ignore the interactions of motifs, making them inappropriate for anomaly detection in traffic videos, where interactions may play a leading role in the determination of normal and "abnormal" activity.

Clustering of groups of pixels is used in the work of (Cui et al., 2007), where blobs corresponding to visual events are created. Blob-level features are then used in a sequential Monte Carlo approach in order to detect abnormalities. However, in this case "anomaly" is examined in each frame as a whole, making it harder to detect anomalies occurring only in a small part of a dense scene (global approach). Additionally, this blob based method, as derived from the connected components, seems to be inadequate in dense scenarios. Furthermore, the "superpixels" in (Cui et al., 2007) are essentially blocks of pixels which have a fixed square shape, whereas the superpixels used in our work are formed based on appearance similarity, following the SLIC method of (Achanta et al., 2012). Additionally, the features of (Cui et al., 2007) are derived based on statistics of pixel-frame differences to define regions of interest, instead of more accurate optical flow estimates, making that method susceptible to noise.

A graph representation and a MCMC model is used by (Kwon and Lee, 2015) for event summarization and rare event detection. However, as stated, that method is only suitable for events caused by several objects, missing "anomalies" caused by a single object (e.g. a police car driving in the wrong way). This constitutes a quite important limitation, as it may miss significant "irregularities" often occurring in traffic scenes.

An interesting work in the field of "anomaly" detection in traffic scenes is that of (Yuan et al., 2017), where the problem is examined from the driver's perspective, using moving camera, instead of surveillance camera data. By definition, (Yuan et al., 2017) addresses a different problem than our work, as ego-centric camera data poses challenges related to the varying camera viewpoint, while also providing different types of information about the traffic scene, compared to static camera surveillance videos.

Another recent work (Cheng et al., 2015) uses hierarchical feature representations and a Gaussian Process Regression (GPR) framework to build a low-level and a high-level codebook respectively. Anomalies are then detected, after the integration of local and global anomaly detectors. A two-level motion pattern and Latent Dirichlet Allocation (LDA) were adopted in (Song et al., 2014) for scene understanding, which, as mentioned above, has limited applicability as it requires prior knowledge of the number of events in the scene. In (Kuettel et al., 2010), Dependent Dirichlet Process and Hidden Markov Model (DDP-HMM) are used in order to capture spatiotemporal dependencies in traffic scenarios. The main drawback of these methods is their usually high computational cost, due to the large volume of data they manage, in combination with the complexity of their models.

It is worth mentioning that our topic based framework, proposed in this work, is significantly differentiated from the previously described topic related works in the following two ways.

1. We extract features in local regions of interest which are comprised of superpixels with a similar appearance and

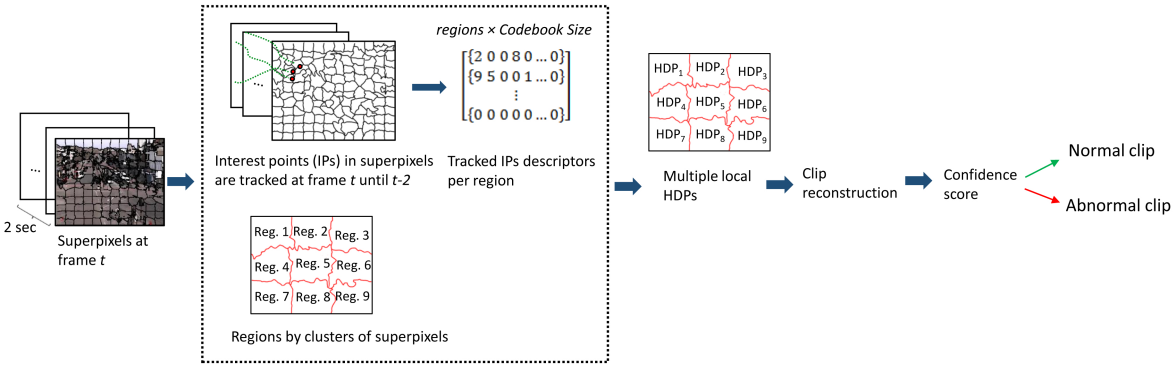


Fig. 1. Overview of the proposed method. At each time instant t, a frame is divided into superpixels, and interest points are extracted in them. They are then tracked over previous frames during a 2 sec time window, so as to retain information about their temporal evolution and spatiotemporal correlations. Regions are formed in each frame by grouping the superpixels: in this work, each frame comprised of 3×3 regions. Codewords are formed for the optical flow orientation and location of the interest points. Multiple HDPs are then deployed in each region, to efficiently characterize the topics in them. The regions are reconstructed based on these HDPs, and a confidence score for each region determines if it contains abnormal events.

a significant level of motion, instead of using the whole set of moving pixels present in the frame. This way, more accurate region modeling is achieved, while local noise artifacts are efficiently avoided.

2. We deploy multiple HDPs in these frame regions, instead of a single global model proposed in the aforementioned works, resulting in the enhancement of our algorithm's resolution.

These features, in combination with the low computational cost, due to its modeling simplicity, make our method most suitable to be used for anomaly detection tasks in structured traffic video sequences.

The paper is organised as follows: Sec. 3 describes the problem formulation and presents methods for video representation and deployment of local HDPs; Sec. 4 presents the anomaly detection and localization framework; and a detailed experimental evaluation is discussed in Sec. 5. Conclusions are summarized in Sec. 6.

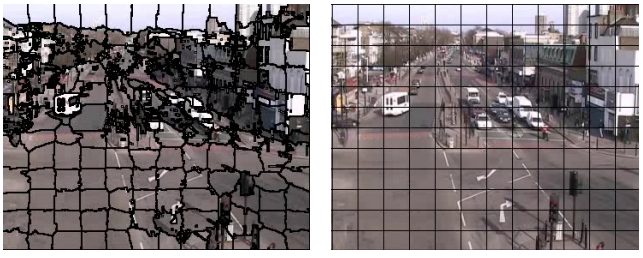
3. Problem Formulation

In this work we address the problem of spatiotemporal anomaly detection and localization in traffic video data. This is a particularly challenging problem due to the variety of scenarios and the complexity of the scenes, often containing a wide range of different objects (people, small/big vehicles etc.) whose density may dramatically change at a moment's notice.

In our approach each frame is divided into superpixels and regions are formed as clusters of these superpixels. Thus, regions with varying boundaries are generated, avoiding the division of moving parts of objects into different areas, and retaining valuable information about object shape and motion. Spatiotemporal Interest Points (STIPs) are then sparsely extracted and tracked in these volumes. Their motion history from the temporal window in which they are tracked is used as input in local topic models, which are applied to capture separate semantic topics for each region. The use of many topic models in each frame - instead of one - allows the efficient capture of elusive details. The use of local topic models also overcomes issues originating from local sources of noise, such as incomplete or erroneous trajectories. An overview of the proposed method is presented in Fig. 1.

3.1. Video Representation

In order to extract a meaningful representation of the video sequence, each frame is divided into 12×12 superpixels, using the implementation of (Achanta et al., 2012). In this manner pixels are grouped into regions with a homogeneous appearance, which have a high likelihood of corresponding to entire objects. Thus they avoid breaking up objects into different cells, as is often the case when regular grids are applied to video frames or images. An instance of this can be seen clearly in Fig. 2(a), where superpixel-based segmentation of the QMUL dataset results in accurate object localization, whereas when a regular grid is applied, objects can be broken up, such as the white truck in Fig.2(b). Our intuitive expectation that the use of superpixel-based segmentation of video frames will result in more accurate event detection is further proven by our experiments, where it is shown that their incorporation contributes to the improvement of our algorithm's performance. Superpixels have been used in the past for image segmentation or video representation (Li et al., 2012), (Chang et al., 2013), and (Li et al., 2013), to segment along object outlines. In our approach, superpixels are grouped into regions of interest, as detailed below, resulting in an area of varying size, that captures both local and global information.



(a) (b)

Fig. 2. An instance of the QMUL dataset when: (a) superpixels are used (b) rigid cells are used In both cases a grid of 12×12 is applied.

In order to capture global event characteristics, while retaining local information, we cluster the superpixels, dividing the frames into $m \times n$ regions. We empirically found that the division of frames into 3×3 regions leads to the best tradeoff between the successful capture of local scene details and containing a sufficient percentage of the overall scenes motion energy. The setting of this parameter is examined in more detail in Sec. 5.1. Different numbers of superpixels were also clustered into regions, from which it was found that a grid consisting of regions containing the same number of superpixels ($4 \times 4 = 16$ superpixels) leads to satisfactory results in all cases examined in this work. As the size of the superpixels themselves changes, the resulting regions have a varying size that is adaptive to the scene characteristics, and are therefore expected to lead to improved results.

After the division of the frame in a fixed number of regions, the motion in each superpixel is estimated using the Farneback optical flow (Farnebäck, 2003), and the amount of activity in it is assessed as follows: a superpixel is considered active if 40% of the pixels in it are characterized by motion whose optical flow value is greater than an empirically determined threshold. Interest points are then extracted in the active superpixels: the position of the pixel in the 95^{th} percentile motion magnitude of all pixels, belonging in the same superpixel, is selected. Pixels with the highest motion magnitude are excluded in order to eliminate possible noise caused by outliers. The adopted threshold of 95% was experimentally seen to be flexible, with other values leading to roughly the same results. The utilization of this method, proved to contribute to more meaningful interest points and hence, to more representative descriptors.

Once interest points are extracted, they are tracked using the KLT tracker (Shi and Tomasi, 1994). Each frame is then characterized by visual words, derived from the spatial location of interest points and the orientation of their neighbourhood pixels' optical flow values. More specifically, all possible quantized locations and orientations are represented by a fixed codebook: spatial information is retained in the positions of all superpixels' centers, and the orientation of their optical flow value is quantized into 8 bins. Thus, the corresponding codebook comprises of the Cartesian product of all possible locations and orientations. Finally, a codeword indicating no motion is also added, making the codebook's size equal to $\{(12 \times 12) \times 8 + 1\}$, for our grid of 12×12 superpixels. Consequently, each interest point is defined by the location of the superpixel it belongs to, and the dominant orientation of its "active" neighbourhood within a region of interest (ROI) of 12×12 pixels around it.

In order to account for temporal variations, which are central in the detection of abnormal events, the video sequence is divided into overlapping subsequences of 2 seconds in length, overlapping by half their size (1 sec). Regions comprised of superpixels are then described as a histogram over the finite aforementioned codebook, with no zero elements corresponding to the visual words in their interest points' trajectory history. In this way, informative descriptors are extracted from varying spatiotemporal volumes, as constructed from the combined action of superpixels clustering and tracking. Local HDPs are subsequently deployed, for each region separately to effectively describe its dynamics.

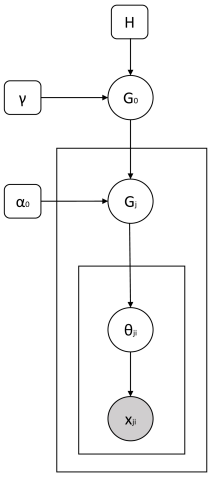


Fig. 3. Graphical model representation of HDP model as suggested in (Y.W. Teh and Blei, 2007). Each node denotes a random variable, with shading suggesting an observed variable. Rectangles denote a replication of the model within the rectangle.

3.2. Local Hierarchical Dirichlet Processes

Probabilistic Topic Models (PTM) were first introduced in text mining (Hofmann, 1999) to efficiently process large collection of data by capturing their underlying latent structure. In order to achieve this, topics are extracted in a generative manner from the data corpus, clustering similar words present in the documents. Subsequently, each document is characterized by a probability distribution of the topics in it, which results in a low-dimensional semantic descriptor. In video analysis, documents correspond to video clips (subsequences), whose length is equal to the size of the temporal window used. Thus, the terms "clip" and "document" will be used interchangeably throughout the rest of the paper. Topics can be interpreted as the recurring actions, arising as histograms over the extracted visual vocabulary, described in Sec. 3.1.

The general idea behind the use of topic models in video surveillance is to depict data as a distribution over high level concepts, without any previous knowledge about what these concepts are. The use of a PTM is well suited for anomaly detection in traffic scenes, as it can allow the unsupervised extraction of the rules, usually governing traffic scenes. In our work, Hierarchical Dirichlet Process (HDP) was deployed for topic modeling, in order to avoid the manual determination of the number of topics, which is required by the widely used Latent Dirichlet Allocation (LDA) (D.M. Blei and Jordan, 2003). HDP constitutes an hierarchical extension of the Dirichlet Process (DP), which is defined as a distribution over a random probability measure using a nonparametric prior, allowing the mixture models to share components. The generative model, proposed in (Y.W. Teh and Blei, 2007), is depicted in Fig. 3, where it can be seen that Dirichlet Processes are employed on two levels. A global random probability measure G_0, drawn from a Dirichlet process $DP(\gamma, H)$, is used as a prior distribution over the whole corpus. The base distribution H is a probability measure used

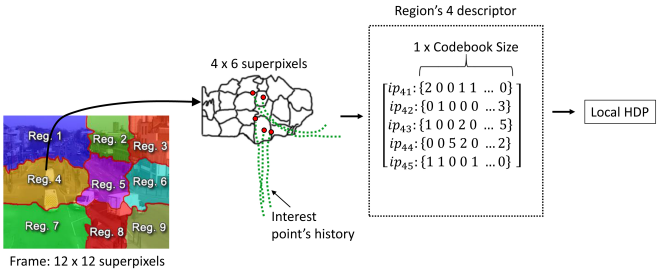


Fig. 4. Example of local HDP extraction for region 4. Each of the 9 regions comprise of a scene-related number of neighbour superpixels, defined at the beginning of the process. For each clip of temporal duration equal to the window size, the interest point history is captured, and a descriptor is calculated for it. Local HDPs are implemented separately in each region, using their descriptors as an input.

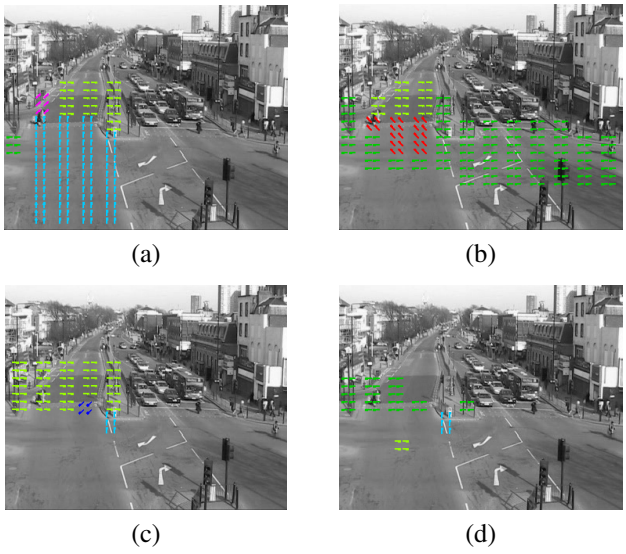


Fig. 5. Instances of topics for region 4 from training data: each image corresponds to a topic instance for region 4 with vectors representing the codewords contained in them. Colors are used to better visualize the direction of the particular codeword. It is clear that the topics are also formed by codewords corresponding to pixels outside of region 4, as they are based on each interest point's history over the past 2 sec. This allows our method to retain information about temporal correlations of the frame features.

by a DP to draw distributions around it, while γ stands for a positive scaling parameter.

$$G_0 \sim DP(\gamma, H) \tag{1}$$

For each document j in the corpus D, a random measure G_j, is then drawn from a second Dirichlet Process with base distribution G_0 and scaling parameter α_0.

$$G_j|G_0 \sim DP(\alpha_0, G_0), \text{ for each j} \tag{2}$$

In using G_0 as a sample measure to create G_j, we implicitly assume that each G_j has support at the same locations as G_0. This means that the whole corpus shares the same set of topics, with only their proportion being differentiated in each document. Hyperparameters γ, α_0 both control the concentration of the word distribution and therefore influence the sparsity of the topics and their percentage in the corpus. Subsequently, a specific topic ϑ_{ji} for document j is sampled from G_j, and a word x_{ji} is then drawn from a multinomial distribution of the selected topic ϑ_{ji}. The overall process can be summarised by the following conditional distributions:

$$\theta_{ji}|G_j \sim G_j \tag{3}$$

$$x_{ji}|\vartheta_{ji} \sim Mult(\theta_{ji}) \tag{4}$$

As for most Bayesian non parametric models, exact posterior inference for the HDP is intractable. Thus, a variety of approximate inference techniques have been proposed, including variational inference and Markov Chain Monte Carlo (MCMC) sampling. The general idea refers to the approximation of the posterior distribution over the latent variables. In our work, the Split-Merge MCMC algorithm, an enriched version of the Gibbs sampler proposed in (Wang and Blei, 2012), was preferred. It operates at the top level of HDP and its main advantage is its potential to lead to a faster convergence, compared to traditional Gibbs sampling.

In our method, multiple HDP models, referred to here as "local HDPs", are generated concurrently, in order to describe the whole frame's dynamics. This goes beyond the current State of the Art (SoA) topic-related works ((Hospedales et al., 2011),(Hospedales et al., 2012),(Kuettel et al., 2010),(Varadarajan et al., 2012),(Hospedales et al., 2009) and (Wang et al., 2009)), which rely exclusively on a single model for the entire frame. The motivation for choosing multiple local HDPs is based on the observation that an anomaly is often described with very few words, relatively to the total number of words in the clip. This results in often missing anomalies occurring locally, and in a limited space, especially in high density scenarios.

Thus, we develop a local-global approach, to detect anomalies with spatiotemporal accuracy. Each frame is firstly divided into 12×12 superpixels, which are spatially clustered to form a 3×3 grid of superpixel-based regions. Then, the interest points in each superpixel-based region at frame t, and their history from their tracking over a 2 sec temporal window, are used to form the descriptor of the region. Interest points with a lifespan less than that of the window size, are also taken into account only if they exist for at least 20 frames, thus eliminating possible outlier noise. The resulting region's descriptor is subsequently used as input for the deployment of the corresponding local HDP. The procedure is depicted in Fig. 4, while examples of topics generated by a local HDP for a specific region are shown in Fig. 5.

The use of interest points' history for the extraction of region's descriptors, often results in visual words based on interest points found outside the boundaries of the superpixel-based region at frame t. This is because an interest point's location in the timespan $t : t-2$ may be part of a different region (e.g. Fig. 5-a, 5-b). In this way, the correlation of features found in different neighbour regions may also be efficiently captured. The proposed framework for local HDP extraction results in the capture of combined filtered local and global information. The experimental results prove that the specific scheme leads

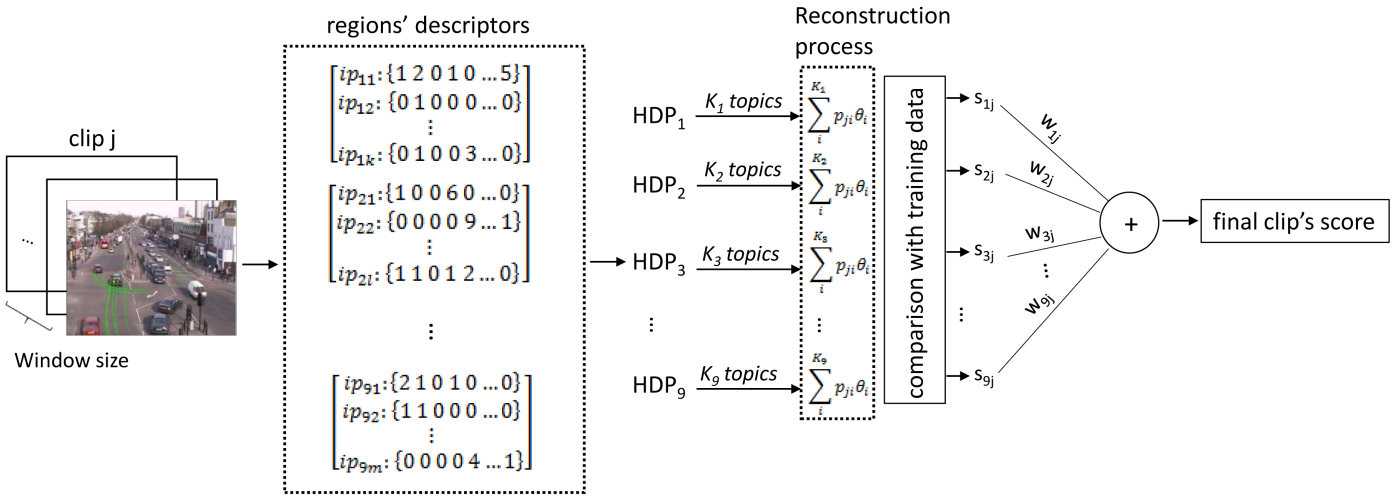


Fig. 6. Inference about a clip's j score. Region descriptors are formed by capturing the trajectory history of all interest points in their area. Subsequently, local HDPs are extracted and their inference is used for the reconstruction of each region. The resulting reconstructed regions are compared with the clips from the training corpus, and a confidence score s_{rj} is calculated for each region r in clip j. The final clip score is then calculated as the weighted sum of all local scores, to determine if a frame is normal or if it contains abnormal events.

to a highly reliable algorithm, as the use of multiple models in the same frame contributes to the capture of more scene details, necessary for an efficient classifier. At the same time, the anomaly localization process is facilitated, as each local HDP provides inference for a particular region.

4. Anomaly Detection and Localisation

The deployment of local HDPs results in the extraction of semantically meaningful topics, defined as histograms over our fixed vocabulary that characterize the underlying motion of the regions. Each region is then described within a spatiotemporal window by a probability distribution over the extracted local topics. Anomaly detection takes place by a reconstruction process of the clip: successful reconstruction of a region by the extracted topic models implies that they describe it well, whereas inaccurate reconstruction implies that an anomaly is present. Thus, each region, is firstly reconstructed according to its corresponding local HDP inference.

If a region is described by a set of local topics $\Theta : \{\theta_1, \theta_2, ...\theta_{K_n}\}$, where each topic θ_i corresponds to a histogram over the fixed codebook's words, the reconstruction of the region can be calculated by the following equation:

$$r_j = \sum_{i}^{K_n} p_{ji}\theta_i \tag{5}$$

where r_j denotes the reconstruction of a specific region r in clip j, p_{ji} stands for the probability of topic i in clip j, and K_n constitutes the total number of topics, as found from local HDPs.

After the division of each dataset to training and testing subsets, the reconstruction of the regions of all clips follows. The cosine distance measure is then used to define a confidence score indicating the region's regularity. This distance measure is chosen as it is indicated for highdimensional positive spaces and has proven to be a useful measure for histogram comparison (Nguyen and Bai, 2011). Furthermore, cosine distances entail relatively low computational cost. Thus, if r_1, r_2 represent two reconstruction instances of the same region, but for different clips, as obtained by Eq. 5, their corresponding distance can be calculated by:

$$D_{r_1,r_2} = 1 - \frac{\sum_{i=1}^{m} r_1(i)r_2(i)}{\sqrt{\sum_{i=1}^{m} r_1(i)^2 r_2(i)^2}} \tag{6}$$

where m indicates the dimensions of r_1 and r_2, which in our case are both equal to the codebook size.

We then define the confidence score of region r in clip j as the minimum distance from the training data, given by:

$$s_{rj} = min\{D_{r_j,r_{tr}}\} \tag{7}$$

where s_{rj} denotes the confidence score of region r for clip j, and $D_{r_j,r_{tr}}$, represent the distances between each reconstructed region r_j, and all the respective reconstructed regions found in the training data, r_{tr}.

With the method proposed, we can determine each clip's "abnormality" by examining each region separately and inferring about its "irregularity". It follows that the localization of the anomaly constitutes an automatic process, as the algorithm is region-based. For more detailed location coordinates of the anomaly, the distribution of the more rare words can be used, as they contain spatial information.

In order to aquire an inference score for the clip as a whole and not for each region separately , we propose a meta processing algorithm that combines the outcomes of all regions. In brief, the algorithm uses statistical measures describing each region r in order to decide the respective weight w_{rj} for each region's confidence score concerning clip j. The final score for each clip then is computed by the weighted sum of all regions' scores for this clip. The method is summarized in Algorithm 1,

Algorithm 1 Meta processing algorithm

1: $w_{rj} = 0$: Initialize all regions' weights for clip j.
2: **for** each region r **do**
3: calculate *hist*: histogram of region's scores as derived from Eq. 7.
4: check clip's score percentile *perc* in the corpus.
5: **if** $perc \geq 98^{th}$ percentile **then**
6: check the frequency *freq* of the corresponding bin in *hist*.
7: **if** $freq \leq$ 2% of the total number of clips **then** $w_{rj} = 1$
8: **end if**
9: **end if**
10: **end for**
11: $score = \sum_r w_{rj} s_{rj}$, where s_{rj} denotes the score of clip j in region r.

where each clip's score is given by Eq. 8.

$$s_{clip} = \sum_r w_{rj} s_{rj} \tag{8}$$

with w_{rj} corresponding to the weight of region r in clip j, and s_{rj} is the region's score, calculated by Eq. 7. The overall process of the extraction of a clip's final score is depicted in Fig. 6.

Anomalous clips are defined as those that have a total score value greater than 3σ, with σ representing the standard deviation of all scores in the corpus. The proposed method is found to work efficiently in challenging traffic datasets, surpassing SoA methods, and detecting even the most challenging anomalous events. At the same time, its computational cost is kept low, making the scheme promising even for future use in real-time applications.

5. Experiments

In order to evaluate the effectiveness of our method, we have applied it on three benchmark datasets of traffic, where different kinds of anomalies were detected and localized. These include the QMUL junction dataset (Russell), the Idiap traffic dataset (Varadarajan and Odobez, 2009) and the Uturn dataset (Benezeth et al., 2009). In all cases, our algorithm's speed and accuracy were calculated and compared with the SoA, demonstrating improved performance.

5.1. Parameter Estimation

The Hierarchical Dirichlet Processes used in our work do not require the prior determination of a significant number of parameters. However, a few basic ones are set a priori in order to determine, for example, the length of the video subsequences to be analyzed. The size of the regions to be examined, both in space and time, is determined experimentally, so as to achieve a balance between local and global information, as it is discussed in more detail below. Because of the static nature of surveillance cameras, all these parameters need to be tuned and set only once, and hence do not affect the general applicability of the method.

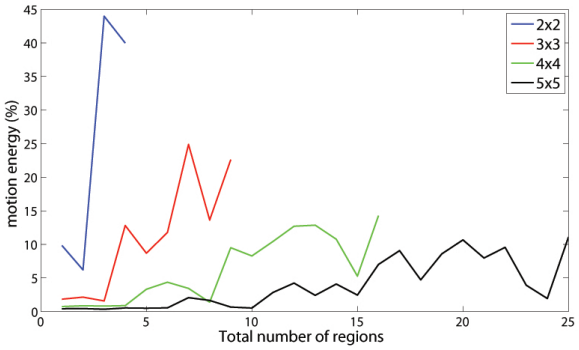


Fig. 7. Investigation of the division of frame in a grid of 2×2, 3×3, 4×4 and 5×5 regions respectively. The horizontal axis depicts the total number of regions per frame, while the vertical axis represents the mean motion energy of each region, as a percentage of the total energy of the frame.

In this work, and as mentioned in the previous Sections, documents defined as clips are created in order to exploit temporal information. The length of each clip should be large enough to contain sufficient information but, at the same time, small enough to avoid the loss of valuable details in long temporal segments. In our experiments, a window size of $2sec$ with overlap of half its size ($1sec$) was found to meet these requirements.

The spatial division of the frame to 9 regions in a 3×3 grid, was experimentally found be a suitable choice, as it divides the frame in small regions where details can be successfully captured but without losing the correlations present on a greater scale. Numerical comparisons of the effect of the varying region sizes is depicted in Figure 7, where it is demonstrated how a balance is achieved between global scene characteristics and local information. In particular, we computed the percentage of the total frame's "motion energy", defined as the sum of the optical flow magnitudes for each region, in the cases where the video frames are divided into 2×2, 3×3, 4×4 and 5×5 regions. In the first case, an extracted region contained $40 - 45\%$ of the total frame's "motion energy", leading to a large concentration of the motion-related information in the scene, which makes it prone to missing valuable local information. On the other hand, the division of the frame into 4×4 and 5×5, resulted in regions containing 10% or less of the "motion energy". These smaller regions are thus more susceptible to the influence of local noise, while neglecting correlations present on a larger scale. Finally, the choice of a 3×3 division of each frame resulted in regions that contained up to 25% of the "motion energy", which is a good tradeoff between the successful capture of local scene details and containing a sufficient percentage of the overall scenes motion energy. The definition of this parameter is global and thus applicable to all cases.

For the HDP models, the topic Dirichlet parameter h, is set to $h = 0.5$, while the parameters of the prior distributions, γ, α_0 were given gamma priors, $\gamma \sim Gamma(0.1, 1)$ and $\alpha_0 \sim Gamma(0.1, 1)$. In general, the use of small values for the concentration parameters γ, α_0 favors topic sparsity, and results in the creation of more topics. The algorithm was run for

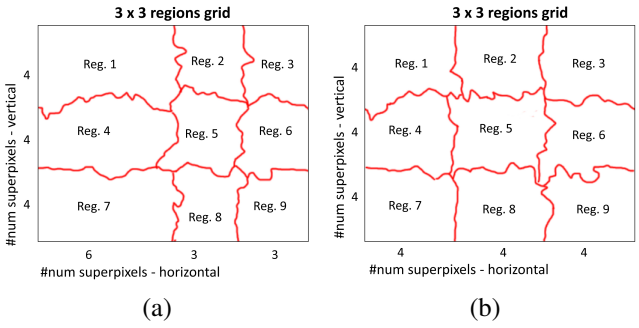


Fig. 8. Examples of two regions' grids, after the division of frame in 12×12 superpixels. In all cases a 3×3 grid of regions is created, with only the number of superpixels in each region, being differentiated. (a) Grid used for the QMUL dataset, (b) symmetrical grid with all regions containing the same number of superpixels. This grid was used for Idiap and Uturn dataset.

200 iterations, giving meaningful and descriptive topics at a low computational cost. Examples of topics for the QMUL dataset can be seen in Fig. 5. All HDPs generating exclusively a single topic, were ignored, as they represent regions with negligible motion.

5.2. Evaluation Criteria

To evaluate our method, we utilize the widely used in the literature, Area Under the Curve (AUC) measure. This is derived from the true positive and false positive rates, with larger areas suggesting better overall performance. Furthermore the Equal Error Rate (EER) derived from this curve, is also used. It corresponds to the error rate of a system when false positives (detections of anomalies in a normal situation) are equal to false negatives (missed anomaly detections). This is achieved by adjusting the threshold for accepting/rejecting a change until equal errors are achieved. The lower the EER, the higher the accuracy of the system.

The area under the precision-recall curve, is also used for the comparison of our method with SoA for the Idiap dataset, as is commonly applied in the literature.

The evaluation and comparison of anomaly detection methods on traffic datasets is hindered by the lack of quantitative results in many SoA works. The literature on anomaly detection in the benchmark traffic datasets provides almost exclusively qualitative results, while video segmentation and scene analysis constitute the main scope of many of these works.

5.3. QMUL

The QMUL junction dataset (Russell) consists of 90000 frames depicting the traffic scene of a busy junction. It constitutes a highly challenging dataset for anomaly detection, due to the wide range of densities it contains, including many scenes with a high density of vehicles. It also comprises of complex motions, with numerous object occlusions, and traffic patterns often interrupted. This is due to the nature of the scene being recorded, which contains a central zone where cars stop temporarily until oncoming traffic stops. Anomalous events like U-turns are difficult to detect, as part of their trajectory is common to other activities (like crossing the road or turning), and moreover, their pattern is often interrupted. As a result, they are often missed by the SoA literature.

Table 1. Comparisons with SoA for anomaly detection in QMUL

Method	AUC	EER	precision-recall area
local kNN	43.6	54.3	-
Dense STC	68.7	36.4	-
Sparse STC	64.5	42.7	-
Sparse IBC	61.8	42.7	-
Loy et al.	77.4	28.1	-
Consistent GPR	85.4	23.8	-
MCTM	-	-	27.87
EM	-	-	30.51
model proposed	**86.57**	**21.69**	**45.63**

To apply our method, the grid of 12×12 superpixels was divided in 9 regions, as shown in Fig. 8-a. The first 12000 frames were used for training our models, while the remaining frames were used for evaluation. In total 412 clips, each with duration of 2 sec, were formed for training purposes and 2689 clips were used to be tested for anomaly detection.

Our method was able to efficiently detect and localize a number of different anomalies, as shown in Fig. 9. The area outlined with red color corresponds to the region where an anomaly occurs, while green lines indicate the tracking history of the interest points found in the particular region. Comparisons with the SoA, provided by (Cheng et al., 2015), are depicted in Table 1. It can be seen that the method proposed outperforms the SoA, as it results in higher AUC, precision-recall curve and a lower EER. The method is compared against 8 other SoA works, which use different approaches to detect spatiotemporal anomalies in this dataset. These include: the use of local nearest neighbour distances for the extraction of scores indicating anomalous events (Saligrama and Chen, 2012), the utilization of a probabilistic framework to infer about sparse or dense spatiotemporal patches modeled by an hierarchical codebook (Javan Roshtkhari and Levine, 2013), the use of "inference by composition" (IBC) to determine about "irregularities" found in sparse patches (Boiman and Irani, 2007), the deployment of Gaussian Process Regression (GPR) to identify unsual spatiotemporal configurations in (C. C. Loy and S.Gong, 2009) and (Cheng et al., 2015), and finally the Markov Clustering Topic Model (MCTM) and its variation (EM) provided by (Hospedales et al., 2012) and (Isupova et al., 2016) respectively.

It is noteworthy, that it surpasses the "consistent GPR" method of (Cheng et al., 2015) by 1.17 in AUC and at a lower computational cost. This is evident in Table 5, where it can be seen that our algorithm is about 6 times faster than "consistent GPR" in the inference process, as it is able to process 5.04 frames per second, in contrast to 0.82 frames per second denoted in (Cheng et al., 2015). The ROC curve for the QMUL dataset is depicted in Fig. 11-a.

Our methodology's evaluation in terms of AUC and EER is presented in Table 2. The accuracy of the proposed method is compared with cases where: 1) a single HDP model is used instead of the multiple local HDPs proposed in this work, 2)

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

Fig. 9. Anomalies detected and localized in QMUL dataset. Red color depicts the region that the anomaly was localized, while green lines correspond to the tracking history of the interest points found in the area. The anomalies concern: (a) Uturn, (b) Jay walking, (c) wrong direction, and (d) traffic break.

Table 2. Investigation of the proposed methodology in QMUL dataset

	single HDP	without superpixels	without weights	method proposed
AUC	54.78	77.07	82.18	**86.57**
EER	44.58	29.43	24.10	**21.69**

cells of a fixed size are used instead of superpixels, and 3) our meta processing algorithm is not taken into account, with equal weights ($w = 1$) assigned to all regions' scores, to assess if a clip is normal or not. The experimental results show that the proposed framework leads to improved accuracy, supporting the use of multiple HDPs with varying cell sizes and weights.

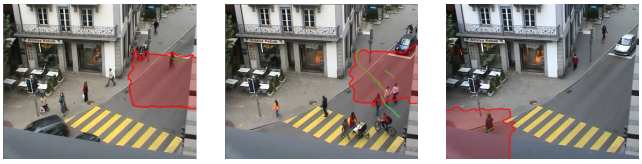
5.4. Idiap

The Idiap traffic junction dataset (Varadarajan and Odobez, 2009) consists of a 45 minute video depicting a road scene containing multiple activities such as: pedestrians walking on the pavement or waiting to cross the road, vehicles moving in and out the scene in different directions. Anomalous events are comprised of jaywalking or people crossing the road away from the pedestrian zone, while an occurence of a vehicle entering the pedestrian area is also present. About 18000 frames were used for training, with the remaining frames kept for testing. In total, 714 clips of 2 sec were created to form the training data, while 1911 clips were used as test data.

Comparisons for this dataset are based on the precision recall area, and are depicted in Table 3, where the AUC for our method is also presented. The proposed algorithm is compared with the Markov Clustering Topic Model of (Hospedales et al., 2009) and its variation provided by (Isupova et al., 2016). Our method achieves a score of 60.41, which is significantly higher than the 37.59 currently found in the literature. We also tested multiple HDPs without using superpixels, with the same number of cells of a constant size to form a region. This resulted in lower performance, confirming the advantage of using superpixels in our algorithm. The ROC curve for this dataset is provided in Fig. 11 (b), while examples of anomalies being correctly detected and localized by our algorithm are shown in Fig. 10.

Table 3. Comparisons for Idiap dataset

	MCTM	EM	model without superpixels	model proposed
precision-recall area	36.43	37.59	48.49	**60.41**
AUC	-	-	83.82	**90.74**



(a) (b) (c)

Fig. 10. Anomalies detected and localized in Idiap dataset. Red color depicts the region that the anomaly was localized, while green lines correspond to the tracking history of the interest points found in the area. All the anomalies concern jay walking.

5.5. Uturn

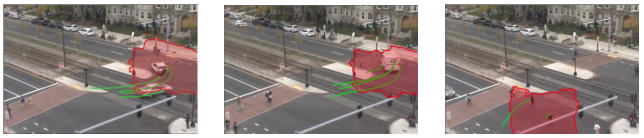
The Uturn dataset (Benezeth et al., 2009) shows normal traffic in a crossroad and some cars making illegal U-turns (defined as "anomalies"). The dataset comprises of 6117 frames of 360×240 pixels. The scenes are quite sparse and the dataset is of a limited size, so there is not much training data. The training set was composed of 1500 frames, while the remaining frames were used for testing, with 97 and 303 clips created in each case, respectively.

Examples of "anomalies" correctly detected in the Uturn dataset can be seen in Fig. 12. The evaluation is summarized in Table 4, where it is compared with the following 5 methods: the hierarchical mixture of dynamic textures method and its variants as proposed in (Li et al., 2014), the local statistical aggregates approach of (Saligrama and Chen, 2012) and the swarm intelligence framework of (Kaltsa et al., 2015). As it can be seen, proposed method achieves very high performance, with an AUC of 94.46, which is comparable to the highest 95.3 in our previous work (Kaltsa et al., 2015). The lack of more training documents, an important factor for the deployment and success of topic models, contributed to the current performance of our algorithm, which is expected to be even further improved with

Table 4. Comparisons for Uturn dataset

	H-MDT-spat	H-MDT-temp	H-MDT	local stat. aggr.	swarm intelligence	method proposed
AUC	83.9	92.9	95.2	94.7	**95.3**	94.46

more training. It is remarkable to note, that our method also revealed 2 jaywalking-related "anomalies" (Fig. 12-c), which are not reported or found in the rest of the SoA ((Benezeth et al., 2009),(Li et al., 2014),and (Kaltsa et al., 2015)), where only Uturns are defined as "anomalies".



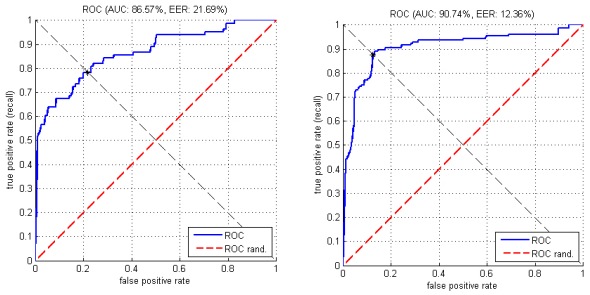
(a) (b) (c)

Fig. 12. Anomalies detected and localised in Uturn dataset. Red color depicts the region that the anomaly was localized, while green lines correspond to the tracking history of the interest points found in the area. Anomalies concern: (a)-(b) Uturn, (c) Jay walking

5.6. Computational Cost

The computational cost consitutes an important factor for the evaluation of an algorithm, responsible for processing of surveillance videos. Whether the algorithm is applied online or offline, it has to run in a reasonable timeframe despite the large volume of data. Although, our algorithm has not been optimized for better performance, our experimental results show that its computational cost remains quite low. Currently, our algorithm works offline, however there is the potential to extend it to an online mode. All experiments were run in C++, on a 16 GB RAM computer with a 3.5 GHz CPU.

Table 5 compares our algorithm's computational cost with the "consistent GPR" method for the QMUL dataset (Cheng et al., 2015). During the training phase, our algorithm is about 1.2 times faster, while in the most important test phase, it achieves a speed up of about 6 times. The omission of superpixel extraction significantly improves its speed, but leads to the deterioration of the algorithm's performance, as Table 2 shows. In Table 6 our method's computational cost is also provided for all datasets, with and without the use of superpixels. It is clear, that the process of superpixel extraction is the main step that delays our algorithm's performance, as its omission results in a 3 times speed improvement, from about 5 frames per second to about 14 frames per second.



(a) (b)

Fig. 11. ROC curves for anomaly detection in (a) QMUL junstion dataset, (b) Idiap traffic dataset.

Table 5. Comparison of Computational cost in QMUL: frames per second

Method	Learning	Inferring
Consistent GPR (Cheng et al., 2015)	4.07	0.82
model proposed	5.00	5.04
model without superpixels	14.23	13.67

6. Conclusion

In this work we propose a novel framework for anomaly detection in different traffic scenarios, given data recorded from static surveillance cameras. Informative regions are formed by clusters of superpixels, in which interest points are detected. The interest points are then tracked over time and their location and flow orientation are encoded and used by multiple HDPs which are applied in each region separately, detecting the topics in them. This results in the robust capture of both local and global information about the scene, by maintaining local feature correlation information over space and time. Its remarkable performance in 3 different benchmark traffic datasets proves the method's generality and its applicability in real life situations. Especially, its significantly high performance in the QMUL dataset, where different kinds of "anomalies" are successfully detected, demonstrates that the proposed algorithm can be effectively used for challenging traffic videos with many occlusions, local scale variations and complex correlated motion patterns. This fact, in combination with its low computational cost, make our algorithm very appropriate for a variety of surveillance applications.

Acknowledgments

This work was funded by the European Unions Horizon 2020 Research and Innovation Programme, within the project "beAWARE: Enhancing decision support and management services in extreme weather climate events".

Table 6. Computational cost: frames per second

Dataset	model proposed	model without superpixels
QMUL	5.04	13.67
Idiap	4.66	11.98
Uturn	5.00	14.07

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Ssstrunk, S., 2012. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 2274–2282.

Benezeth, Y., Jodoin, P.M., Saligrama, V., Rosenberger, C., 2009. Abnormal events detection based on spatio-temporal co-occurences, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2458–2465.

Bertini, M., Bimbo, A.D., Seidenari, L., 2012. Multi-scale and real-time non-parametric approach for anomaly detection and localization. Computer Vision and Image Understanding 116, 320 – 329. Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

Boiman, O., Irani, M., 2007. Detecting irregularities in images and in video. International Journal of Computer Vision 74, 17–31.

C. C. Loy, T.X., S.Gong, 2009. Modelling multi-object activity by gaussian processes, in: in Proc. Brit. Mach. Vis. Conf.

Calderara, S., Heinemann, U., Prati, A., Cucchiara, R., Tishby, N., 2011. Detecting anomalies in peoples trajectories using spectral graph analysis. Computer Vision and Image Understanding 115, 1099 – 1111.

Chang, J., Wei, D., Fisher, III, J.W., 2013. A video representation using temporal superpixels, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Cheng, K.W., Chen, Y.T., Fang, W.H., 2015. Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation. IEEE Transactions on Image Processing 24, 5288–5301.

Cui, P., Sun, L.F., Liu, Z.Q., Yang, S.Q., 2007. A sequential monte carlo approach to anomaly detection in tracking visual events, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

D.M. Blei, A.N., Jordan, M., 2003. Latent dirichlet allocation. J. Machine Learning Research 3, 993–1022.

Farnebäck, G., 2003. Two-Frame Motion Estimation Based on Polynomial Expansion. Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 363–370.

Hofmann, T., 1999. Probabilistic latent semantic indexing, in: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA. pp. 50–57.

Hospedales, T., Gong, S., Xiang, T., 2009. A markov clustering topic model for mining behaviour in video, in: 2009 IEEE 12th International Conference on Computer Vision, pp. 1165–1172.

Hospedales, T., Gong, S., Xiang, T., 2012. Video behaviour mining using a dynamic topic model. International Journal of Computer Vision 98, 303–323.

Hospedales, T.M., Li, J., Gong, S., Xiang, T., 2011. Identifying rare and subtle behaviors: A weakly supervised joint topic model. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 2451–2464.

Isupova, O., Kuzin, D., Mihaylova, L., 2016. Learning methods for dynamic topic modeling in automated behaviour analysis. ArXiv e-prints .

Javan Roshtkhari, M., Levine, M.D., 2013. Online dominant and anomalous behavior detection in videos, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Jeong, H., Yoo, Y., Yi, K.M., Choi, J.Y., 2014. Two-stage online inference model for traffic pattern analysis and anomaly detection. Machine Vision and Applications 25, 1501–1517.

Jiang, F., Yuan, J., Tsaftaris, S.A., Katsaggelos, A.K., 2011. Anomalous video event detection using spatiotemporal context. Computer Vision and Image Understanding 115, 323 – 333. doi:https://doi.org/10.1016/j.cviu.2010.10.008. special issue on Feature-Oriented Image and Video Computing for Extracting Contexts and Semantics.

Kaltsa, V., Briassouli, A., Kompatsiaris, I., Hadjileontiadis, L.J., Strintzis, M.G., 2015. Swarm intelligence for detecting interesting events in crowded environments. IEEE Transactions on Image Processing 24, 2153–2166.

Kaltsa, V., Briassouli, A., Kompatsiaris, I., Strintzis, M.G., 2014. Swarm-based motion features for anomaly detection in crowds, in: 2014 IEEE International Conference on Image Processing (ICIP), pp. 2353–2357.

Kim, J., Grauman, K., 2009. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2928.

Kuettel, D., Breitenstein, M.D., Gool, L.V., Ferrari, V., 2010. What's going on? discovering spatio-temporal dependencies in dynamic scenes, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1951–1958.

Kwon, J., Lee, K.M., 2015. A unified framework for event summarization and rare event detection from multiple views. IEEE Transactions on Pattern Analysis and Machine Intelligence 37, 1737–1750.

Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M., 2013. Video segmentation by tracking many figure-ground segments, in: The IEEE International Conference on Computer Vision (ICCV).

Li, W., Mahadevan, V., Vasconcelos, N., 2014. Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 18–32.

Li, Z., Wu, X.M., Chang, S.F., 2012. Segmentation using superpixels: A bipartite graph partitioning approach, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 789–796.

Nguyen, H.V., Bai, L., 2011. Cosine Similarity Metric Learning for Face Verification. Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 709–720.

Piciarelli, C., Micheloni, C., Foresti, G.L., 2008. Trajectory-based anomalous event detection. IEEE Transactions on Circuits and Systems for Video Technology 18, 1544–1554.

Russell, D., . URL: http://www.eecs.qmul.ac.uk/~sgg/QMUL_Junction_Datasets/Junction/Junction.html.

Saleemi, I., Shafique, K., Shah, M., 2009. Probabilistic modeling of scene dynamics for applications in visual surveillance. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 1472–1485.

Saligrama, V., Chen, Z., 2012. Video anomaly detection based on local statistical aggregates, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2112–2119.

Shi, J., Tomasi, C., 1994. Good features to track, in: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 593–600.

Song, L., Jiang, F., Shi, Z., Molina, R., Katsaggelos, A.K., 2014. Toward dynamic scene understanding by hierarchical motion pattern mining. IEEE Transactions on Intelligent Transportation Systems 15, 1273–1285.

Varadarajan, J., Emonet, R., Odobez, J.M., 2012. Bridging the past, present and future: Modeling scene activities from event relationships and global rules, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2096–2103.

Varadarajan, J., Emonet, R., Odobez, J.M., 2013. A sequential topic model for mining recurrent activities from long term video logs. International Journal of Computer Vision 103, 100–126.

Varadarajan, J., Odobez, J.M., 2009. Topic models for scene analysis and abnormality detection, in: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 1338–1345.

Wang, C., Blei, D., 2012. A split-merge mcmc algorithm for the hierarchical dirichlet process. arXiv:1201.1657v1 [stat.ML] .

Wang, X., Ma, K.T., Ng, G.W., Grimson, W.E.L., 2008. Trajectory analysis and semantic region modeling using a nonparametric bayesian model, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Wang, X., Ma, X., Grimson, W.E.L., 2009. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 539–555.

Wulsin, D., Jensen, S., Litt, B., 2012. A Hierarchical Dirichlet Process Model with Multiple Levels of Clustering for Human EEG Seizure Modeling. ArXiv e-prints .

Yakhnenko, O., Honavar, V., . Multi-Modal Hierarchical Dirichlet Process Model for Predicting Image Annotation and Image-Object Label Correspondence. pp. 283–293.

Yang, W., Gao, Y., Cao, L., 2013. Trasmil: A local anomaly detection framework based on trajectory segmentation and multi-instance learning. Computer Vision and Image Understanding 117, 1273 – 1286.

Yuan, Y., Fang, J., Wang, Q., 2015. Online anomaly detection in crowd scenes via structure analysis. IEEE Transactions on Cybernetics 45, 548–561.

Yuan, Y., Wang, D., Wang, Q., 2017. Anomaly detection in traffic scenes via spatial-aware motion reconstruction. IEEE Transactions on Intelligent Transportation Systems 18, 1198–1209.

Y.W. Teh, M.I. Jordan, M.B., Blei, D., 2007. Hierarchical dirichlet processes.

J. Am. Statistical Assoc. 101, 1566–1581.