Anomaly Localization in Topic-based Analysis of Surveillance Videos

Deepak Pathak
IIT Kanpur
Dept. of Computer Science
deepakp@iitk.ac.in

Abhijit Sharang
IIT Kanpur
Dept. of Computer Science
abhisg@iitk.ac.in

Amitabha Mukerjee
IIT Kanpur
Dept. of Computer Science
amit@iitk.ac.in

Abstract

Topic-models for video analysis have been used for unsupervised identification of normal activity in videos, thereby enabling the detection of anomalous actions. However, while intervals containing anomalies are detected, it has not been possible to localize the anomalous activities in such models. This is a challenging problem as the abnormal content is usually a small fraction of the entire video data and hence distinctions in terms of likelihood are unlikely.

Here we propose a methodology to extend the topic based analysis with rich local descriptors incorporating quantized spatio-temporal gradient descriptors with image location and size information. The visual clips over this vocabulary are then represented in latent topic space using models like pLSA. Further, we introduce an algorithm to quantify the anomalous content in a video clip by projecting the learned topic space information. Using the algorithm, we detect whether the video clip is abnormal and if positive, localize the anomaly in spatio-temporal domain. We also contribute one real world surveillance video dataset for comprehensive evaluation of the proposed algorithm. Experiments are presented on the proposed and two other standard surveillance datasets.

1. Introduction

Analyzing surveillance videos to identify unusual or anomalous events is a challenging problem owing to the wide range of events that can be called anomalous. The dominant class of approaches is to a) model the common actions based on weak supervision in terms of a set of training videos that may not contain anomalies [3, 17], b) identify anomalies as spatio-temporal patterns that do not agree with the models. Models may be based on clustering trajectories (point-based) [4, 7, 11, 22], or on finding correlations among a collection of features (trajectory, spatio-temporal position, texture, size). Correlations may seek to preserve local context using markov fields [9] , or may seek to identify latent topics from the word-document data [6, 16, 20, 23]. Finally, anomaly detection is usually based on computing the likelihood of the test video-fragment, given the model.

(a) Traffic Junction Dataset [20]. The left image is usual while right is an anomaly - car stops after the stop-line.

(b) Highway Dataset. The left image is usual while right is an anomaly with a jaywalker crossing the road.

(c) AVSS Dataset [8]. The left image is usual while right is an anomaly with a vehicle abruptly crossing the road.

Figure 1: Sample frames from video datasets.

In this paper, we focus on topic models for situations where several actions are happening in the scene at the same time. Such situations have been investigated recently by Varadarajan *et al*. [20]. Two difficulties can arise in such a case. First, many actions are occurring simultaneously, but in any anomalous scene only one of the events may be unusual. Consequently, the likelihood of a video snippet containing an anomaly may not vary significantly from that of

a "normal" snippet; in Figure 2 we present the distribution curves for normalized log-likelihood for test clips, and find that the difference between that of 'normal' and 'anomalous' clips is very small. Consequently, any classifier will have to be contend with a high cost in terms of false classifications.

The major challenges involved in scenes involving many agents are frequent occlusions, and a large divergent set of behaviors. Many anomaly detection datasets involve either very sparse agents, or dense data (crowds) that are a glutinous whole. Figure 1 shows all the three datasets we evaluate our approach on.

The second challenge relates to the fact that in a topic model, it is often possible to identify the document containing the anomaly, without being able to localize the region of the image+time where it is occurring. The word-based neighbourhood analysis also resolves this problem, and we are able to mark regions in the video that contain anomalous actions.

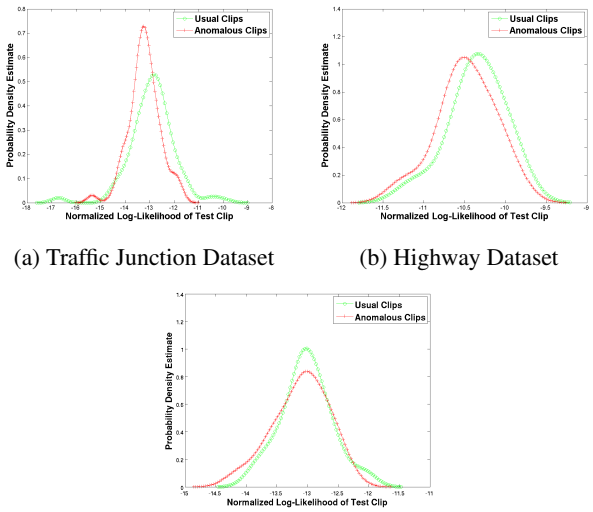
Our primary contribution to tackle these problems is to propose a mechanism where the topic-based identification of anomalous documents is combined with a classifier based on spatio-temporal quantized words. The mechanism has three steps: (a) We design the visual vocabulary by incorporating the location and blob size information with quantized spatio-temporal descriptors, which is particularly relevant to static camera scenes. The blob size helps in differentiating individuals from vehicles. The visual clips over this vocabulary are then represented in latent topic space using models like pLSA. (b) We propose an algorithm to quantify the anomalous content in a video clip by projecting the information learned from training on to the clip. (c) Based on the algorithm, we finally detect whether the video clip is abnormal or not and if so, further localize the anomaly in spatio-temporal domain. This mechanism is shown to provide an improvement of the area under the precision-recall curve(AUC) of anomaly detection over the results from the likelihood model proposed in [20]. We call this the 'projection model algorithm' and it is particularly robust to the amount of abnormal content in video clip. This not only enhances the detection accuracy but also provides spatio-temporal localization of the anomalous content.

This paper is organized further as follows. Section 2 discusses about the literary survey in general. Section 3 provides methodology i.e. modeling, detection and localization of anomalies. Results and experiments with respect to baseline are in Section 4 followed by concluding discussion in Section 5.

2. Related Work

Video abnormality detection broadly includes two major approaches. The first approach involves tracking objects in a series of frames, and then working in the trajectory space to identify the deviant points which are potential candidates for being anomalous [7, 11, 22]. Such methods yield good performance in general, but are not robust to occlusions which cause inaccuracy in tracking. This impedes their extent to surveillance videos with natural traffic scenarios. In the second approach, the intrinsic patterns of the scene are captured using feature descriptors, which are then used to model the behaviour. However, removal of tracking information in the latter approach leads to significant loss but this shows a more promising way to build generalizable real-life models. Niebles *et al.* [16] discuss the results obtained by applying topic models for the task of action recognition and classification. They model the features in terms of visual words, and use pLSA-LDA models to predict the correct actions in an unsupervised sense. Wang *et al.* [21] propose a hierarchical variant of LDA for connecting low level visual features, atomic actions and interactions. This model achieves multi-fold objectives of discovering typical actions in videos, segmenting long video sequences into different actions, and segmenting motions into different activities. Li *et al.* [12] use hierarchical pLSA for generating global behavioural correlations in a wide area scene. This model also aids in detecting anomalous activities in the scene subsequently. Mehran *et al.* [15] present a social force model for dense crowd behaviour, and their model is primarily driven by optical flow based ideas for detecting abnormal panicky crowd situations. They detect and localize anomalous behaviour, but the approach does not extend to sparse and structured anomalous space.

Works which attempt to detect unusual events in the videos begin by building a model for the normal events occurring in the videos. Mahadevan *et al.* [14], which aims at



(a) Traffic Junction Dataset (b) Highway Dataset

(c) AVSS Dataset

Figure 2: Density plots for Normalized Log-Likelihood for test 'usual' and 'anomalous' clips.

anomaly detection, and its extension [13] to localization of the anomaly rely on dynamic texture models for building a joint temporal and spatial model for constructing saliency measure for the events occurring in the video. A rare or unusual event is expected to possess temporal and spatial saliency values which are significantly deviant from the expected saliency values.

Roshtkhari *et al*. [18] also build a joint model for spatial and temporal dominant behaviour by constructing spatio-temporal volumes centred around every pixel. The features are clustered using a fuzzy C-means clustering algorithm. Anomalous events are detected based on the distances of the words occurring in the events to these centres. Varadarajan *et al*. [20] utilise topic modelling for understanding the usual events occurring in the video. It is assumed that in a domain, the set of usual events is fixed and can be mined from the distribution of the visual words and the video clips in the domain. A video clip which has the occurrence of anomalous events would then be expected to have a low likelihood over the learnt model.

3. Methodology

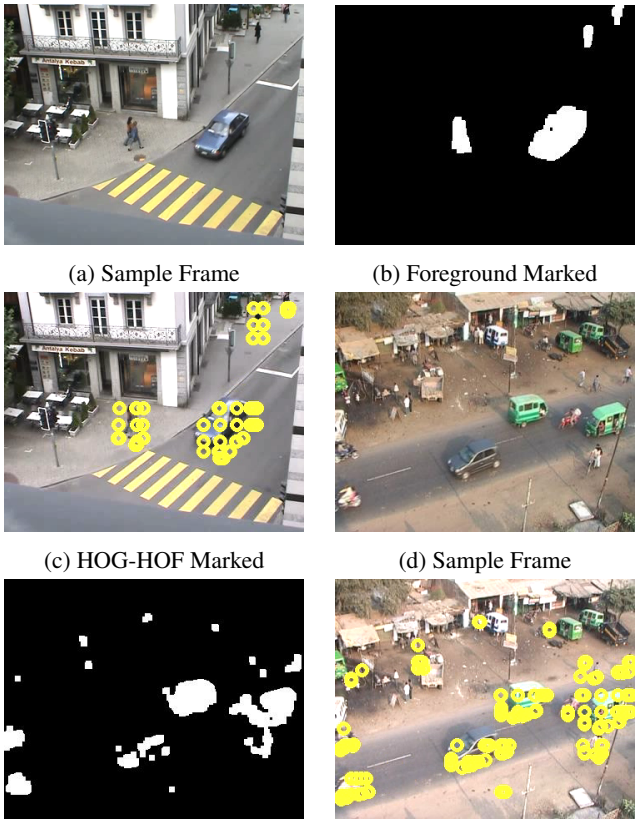
Our aim is to model the usual events, given a surveillance video, so as to detect the anomalous events in an unsupervised setting. The two major issues addressed in the methodology are robustness to the quantity of abnormal events in a video clip, and localization of anomaly in space-time domain. We now discuss the proposed three tier framework for video abnormality analysis - modeling, detection and localization.

3.1. Unsupervised Modelling

We use topic model to parametrically learn the informative content of surveillance video. Initially, videos are divided into clips and the length of clip (l seconds) defines the granularity for time interval of abnormality detection. These clips are analogous to *documents* in language processing. It is crucial that training clips contain no or very less anomaly, so that the resultant normalized likelihood of usual topics is high in the learned model. We then extract context-based finite dimensional, discrete domain words from the video, called *visual words*. These documents (video clips) are then represented as histograms over the finite visual vocabulary. This forms the basis for topic-modelling in videos.

3.1.1 Formation of Words

The *visual words* should be rich and generalizable enough to capture possible behaviours or events in any video. The vocabulary of such words should also belong to a finite domain, so as to account for unseen test document clips. Improving upon the feature descriptors suggested in [20, 21], we incorporate HOG-HOF descriptors [10] with location and size information; each of them being quantized. The proposed visual words are three dimensional : spatial location, HOG-HOF cluster and parent blob size.



(a) Sample Frame (b) Foreground Marked (c) HOG-HOF Marked (d) Sample Frame (e) Foreground Marked (f) HOG-HOF Marked

Figure 3: Representative images for intermediate results. Figures (a)-(c) are from Traffic Junction Dataset and (d)-(f) from Highway Dataset.

The foreground segregation in image frame reduces the complexity of problem to a great extent with focus on dynamic actions which are relevant candidates for being anomalous. We consider visual words only at these foreground pixels. We use the ViBe foreground extraction technique suggested in [1]. This approach develops the background model by ignoring the insertion time attribute of pixel entering it. The main benefit of ViBe is to depict the object in foreground for sometime even after it has stopped moving, for example, a car parked in a no-parking area. This assures that such events do not go unscrutinised during abnormality analysis. We then use median filtering and repeated morphological filters for smoothening the image and removing the noise. Figure 3 depicts the result of this pre-processing on sample frames. Now we discuss how the information in individual dimension of these visual words is being assimilated.

Each video frame is divided into disjoint grids, and each

grid acts as the location attribute of visual word. Every location cell in a frame containing at least one foreground pixel is a candidate for defining a visual word. The location attribute is significant especially in static camera videos.

As second attribute, we use space time extension of HOG descriptor complemented with quantized optical flow features. This is HOG-HOF descriptor suggested in [10]. We find this descriptor around a pixel selected randomly from the foreground ones in each cell. The idea of the descriptor is to consider a spatio-temporal volume around the interest point, which is further divided into disjoint cells. Thereafter in each cell, gradient orientations are quantized into 4-bins (HOG) and optical flow into a 5-bin histogram (HOF). These histograms are normalized within cells and then concatenated for the complete volume. The spatial scale (σ^2) and temporal scale (τ^2) parameters used are 4 and 2 respectively. The total length of descriptor is 162 i.e. 72 dimensional HOG vector and 90 dimensional HOF vector. Overall, these orientation quantized bins capture the gradient-texture information, while optical flow histogram incorporates the motion content in the neighborhood. These descriptor values are then clustered using k-means algorithm. We randomly pick 200K HOG-HOF descriptors from training documents and quantize them into 20 centers.

Final attribute corresponds to the size of connected component of the foreground pixel in consideration. We find the 4-side-connected 'blob' of the pixel using contour detection algorithm [19]. Contours are then filled and the area under the contour is computed, which we quantize using a threshold into either large or small.

3.1.2 Construction of vocabulary and documents

The dimension of frames in video dataset is $288*360$. Each frame is divided into $20*20$ disjoint grids, leading to $15*18$ possible cells. Hence these many values are possible for the location attribute. The vocabulary is the domain of all possible values for visual word. Since last dimension 'size' has 2 quantizations, each word is a triplet accounting for $(15*18)*20*2 = 10800$ possible combinations.

3.1.3 Probabilistic Latent Semantic Analysis

Probabilistic topic modeling has a wide literature in statistical learning. Beginning from Latent Semantic Analysis, the probabilistic graph based model pLSA was suggested by [5], and subsequently a parametric fully generative model with dirichlet prior LDA was suggested in [2]. It is suggested that pLSA and LDA give similar results in capturing activity pattern [20], so we would discuss pLSA model for topic discovery.

Say we represent each word as $w \in W = \{w_1, w_2, \dots w_M\}$ and each document as $d \in D = \{d_1, d_2 \dots, d_N\}$, then we have a $N * M$ term-frequency matrix $\mathbf{N}$ where $\mathbf{N}(i,j)$ is the frequency of w_j in d_i. LSA tries to factorize this matrix into lower vector space by estimating the SVD of N considering only significant diagonal terms. pLSA is probabilistic version of LSA to represent a document as a probability distribution over the space of latent factors called topics say $z \in Z = \{z_1, z_2 \dots, z_K\}$. The conditional independence assumption in the pLSA model is that given the topic z, the variable w and d are independent. The joint distribution of word and topic space, respecting this independence assumption is given by

$$P(d,w) = P(d)P(w|d) = P(d)\sum_{z\in Z} P(w,z|d)$$

$$= P(d)\sum_{z\in Z} P(w|z,d)P(z|d)$$

$$= P(d)\sum_{z\in Z} P(w|z)P(z|d)$$

The parameters of the model are estimated using EM algorithm as suggested in [5]. The likelihood estimate of the document matrix is shown as follows

$$\mathcal{L}(\theta;\mathbf{N}) = \sum_{d\in D}\sum_{w\in W} n(d,w)\log(P(d,w))$$

where $n(d,w)$ is the frequency of word w in document d. The limitation of pLSA model is that it is not fully-generative for the testing data where documents are unseen. Thus, we estimate the distribution $P(z|d)$ and $P(w|z)$ from the training data and change the EM algorithm for test set to estimate $P(z|d)$ using the $P(w|z)$ distribution estimate from the training set. Thus we use following form of Likelihood function:

$$\mathcal{L}(\theta;\mathbf{N}) = \sum_{d\in D}\sum_{w\in W} n(d,w)\log\left(P(d)\sum_{z\in Z} P(w|z)P(z|d)\right)$$

3.2. Anomaly Detection: Projection Model Algorithm

Although we have the overall likelihood values for any document obtained as a result of topic modeling, using them for abnormality detection is not a robust approach. This is sensitive to the amount of anomaly, i.e. the number of anomalous words, present in the document. In general, the abnormal event in any clip is confined to a small spatio-temporal region, thus leading to very few anomalous words in the clip relative to the total number of words present in it. Due to this, there is not much difference between the likelihood of an anomalous and that of a usual test clip. So, we propose an algorithm for individual evaluation of visual words present in the test video document. We call this *projection model algorithm* due to the fact that for every word we mine the information projected from the nearest training documents in topic space. The details of the algorithm are as follows.

1. The likelihood of documents in topic space i.e. $P(z|d)$ is given by the pLSA model. Using this, we can represent every document d_x as a distribution over topic space as $(\theta_1^x, \theta_2^x \ldots, \theta_k^x)$. Let D_{train} be the set of all such topic vectors for the training documents.

2. Given a new test document d_{test}, represent it in terms of a topic vector. In the topic space, find the nearest m training documents $d_i \in D_{train}$ using the Bhattacharyya metric. The Bhattacharyya distance between two documents d_x and d_y is defined as follows -

$$D_B(d_x, d_y) = -\log\left(\sum_{i=1}^{K}\sqrt{\theta_i^x \theta_i^y}\right)$$

3. Let the word histogram of document d_x be H_x. Then stack (i.e. add frequency of each bin) the histogram of all these m nearest train documents into a combined histogram H_0.

4. Now observe every word w_{test} that occurs in the test document d_{test} (the words that do not occur are ignored right away) at least once. If the frequency of the bin corresponding to w_{test} in H_0 is more than a certain threshold, call it usual i.e.

   If $(H_0(w_{test}) \geq th_{cur})$ then w_{test} is usual word

5. Consider the eight neighbors of w_{test} in the grid image in all possible spatial directions i.e. up,down,left etc. Let their set be $N(w_{test})$. Now if $H_0(w) \geq th_{nbr}$ is true for at least l neighbors $w \in N(w_{test})$, then the word w_{test} will be called usual.
   Note that l is any integer from 1 to 8.

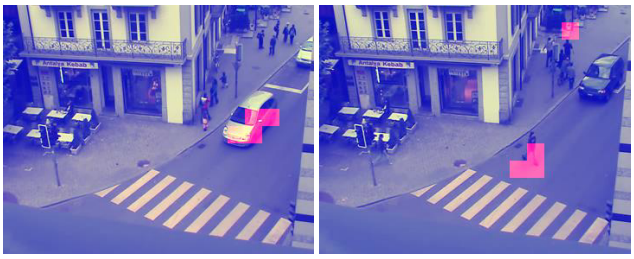
6. If steps 4 and 5 do not hold for w_{test}, then call it an anomalous word.

In the above algorithm, if the training data does not contain any abnormal event then we keep $th_{cur} = 1$. Thus, final parameters to optimize are $\{m, th_{nbr}, l\}$. Through experiments, we observed that keeping the value of m to be around one fourth of total training documents and keeping value of l as 3 gives decent performance.

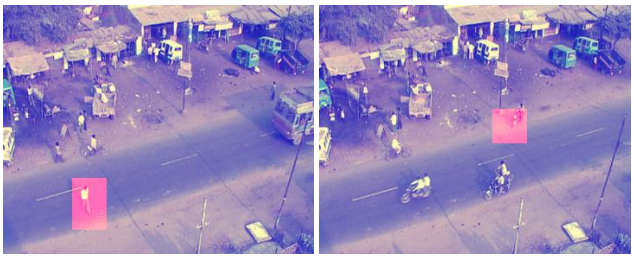
Now, the test document clip d_{test} will be called abnormal if the number of anomalous words in it is more than a threshold. We vary this threshold and present the precision-recall curve in the results section.

3.3. Anomaly Localization

We get the localization of anomaly as a direct bi-product of our projection model algorithm. If the overall test document d_{test} is being called abnormal, then we mark the anomalous flagged words in clip d_{test}. This is possible because words contain the spatial location information in them as their attribute. In implementation, we just need to do book-keeping of frame numbers while creating word histograms of each document.



(a) Traffic Junction Dataset



(b) Highway Dataset

(c) AVSS Traffic Dataset

Figure 4: Anomalous frames identified and anomalous words localised by the algorithm. Currently, in our implementation we highlight the anomalous event in test documents as shown above.

4. Experiments and Results

We perform experimentation on datasets created in real world setting shown in Figure 1:

Traffic Junction Dataset: It consists of a single video of 45 minutes duration shot from a camera perched at the top of a building at a traffic junction. This was released in [20] Anomalous events occurring in the video have been marked in a separate file which states the start time of the anomalous event, the end time of the anomalous event and the kind of anomaly. There are four kinds of anomalous actions in the video.

Highway Dataset: We contribute this 6 and a half minute video depicting the traffic scene of a highway in real world scenario. Frame rate is $25fps$. We provide the temporal labels for abnormal events for evaluation of results. The anomaly description is described in text for localizing the

Dataset	Likelihood Model	Projection Model
Traffic-Junction Dataset	54.47 (54.04)	**65.15** (**85.19**)
Highway Dataset	67.30 (71.08)	**81.40** (**84.36**)
AVSS Traffic Dataset	68.11 (68.63)	**75.49** (**75.53**)

Table 1: Results for Anomaly Detection. The reported values are AUC of precision-recall curve with average precision in parenthesis.

abnormal event. This dataset is made publicly available[1].

AVSS traffic Dataset: This video was released as a part of i-LIDS vehicle detection challenge in AVSS 2007 [8]. It consists of a single long shot of traffic on a sub-urban street over the day, with varying lighting and weather conditions, and trembling camera which makes the task of robust foreground extraction challenging. Unlike the first two datasets, the anomalous events have not been provided separately. So, we manually find and label the anomalous events.

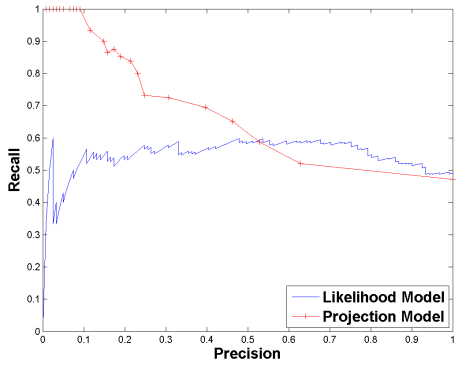
The experimentation was conducted by keeping the number of actions in the video to be 20, which served as the number of topics in the document. We experiment with length $l = 4$ to $l = 10$ seconds, presenting final results on contiguous document clips of 4 sec duration. Anomalous video clips were separated from the rest of the video clips for testing. From the remaining set, 75% of the clips were used for training and the remaining 25% of the clips were included in the test data along with the anomalous ones. Figure 4 shows the detection and localization results of proposed *projection model algorithm*.

For anomaly detection i.e. predicting whether a clip contains anomaly or not, the baseline model was fixed as the normalized likelihood model with the same feature set i.e. vocabulary as our proposed algorithm. Baseline results are obtained by varying the log-normalized likelihood as the threshold. The comparison with baseline model, as seen in Figure 5 for all datasets, suggests that *projection model algorithm* improves upon detection in addition to localization of anomalies. The quantitative results for detection accuracy are reported in Table 1. We measure the 0/1 accuracy for localization of anomalies by manual inspection of test clips. Since the baseline model does not deal with anomaly localization, we do not provide a direct comparison of the two models with respect to anomaly localization. Our algorithm localizes **55.88%** of abnormal events in Traffic Junction Dataset, **63.15%** in Highway Dataset, and **59.23%** in AVSS Traffic dataset correctly.

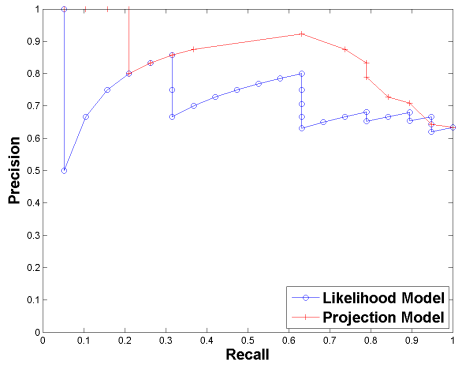
5. Conclusion

In this paper, we propose an improved methodology for anomaly detection and localization in situations with a large number of agents actively pursuing divergent goals. With a weakly supervised input, the *projection model algorithm* improves upon the baseline likelihood model in detection accuracy, and additionally localizes the abnormal actions in space-time. The system uses object-based models, including object size, which are identified via modern techniques for foreground modelling and low-level feature description.

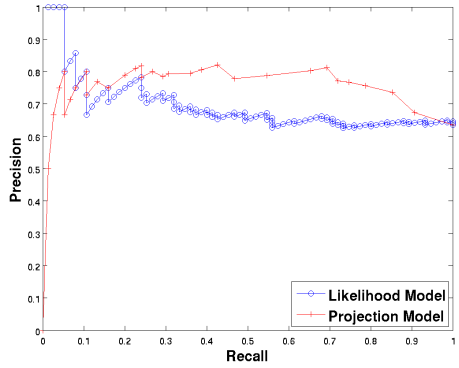
[1] http://www.cse.iitk.ac.in/users/vision/traffic-datasets/dataset3/dataset3.html



(a) Traffic Junction Dataset



(b) Highway Dataset



(c) AVSS Traffic Dataset

Figure 5: Precision-Recall curves for anomaly detection. Anomalous clips were considered as positive and the non-anomalous clips as negative examples.

This is efficient in situations with sparse anomalous set in structured behaviour scenarios, which is mostly the case with surveillance videos. We also contribute a real world surveillance video dataset with marked temporal anomalous events.

One possible direction for future work would be to investigate hierarchical topic models instead of pLSA to obtain finer granularity in activity pattern analysis. The proposed pipeline has independent components connected together, and thus improving upon any of the steps would increase the overall accuracy. In future, we plan to automatically infer semantic tags for objects in natural language using the commentary for localized abnormal events.

References

[1] O. Barnich and M. Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *Image Processing, IEEE Transactions on*, 20(6):1709–1724, 2011. 3

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. 4

[3] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009. 1

[4] H. Guo, X. Wu, N. Li, R. Fu, G. Liang, and W. Feng. Anomaly detection and localization in crowded scenes using short-term trajectories. In *Robotics and Biomimetics (ROBIO), 2013 IEEE International Conference on*, pages 245–249. IEEE, 2013. 1

[5] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999. 4

[6] T. M. Hospedales, J. Li, S. Gong, and T. Xiang. Identifying rare and subtle behaviors: A weakly supervised joint topic model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(12):2451–2464, 2011. 1

[7] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank. Semantic-based surveillance video retrieval. *Image Processing, IEEE Transactions on*, 16(4):1168–1181, 2007. 1, 2

[8] i-Lids dataset for AVSS 2007 available at : `http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html`. 1, 6

[9] J. Kim and K. Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2921–2928. IEEE, 2009. 1

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 3, 4

[11] T.-L. Le, M. Thonnat, A. Boucher, and F. Brémond. A query language combining object features and semantic events for surveillance video retrieval. In *Advances in Multimedia Modeling*, pages 307–317. Springer, 2008. 1, 2

[12] J. Li, S. Gong, and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. In *BMVC*, volume 3231, page 3232, 2008. 2

[13] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1, 2013. 3

[14] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981. IEEE, 2010. 2

[15] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE, 2009. 2

[16] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008. 1, 2

[17] O. P. Popoola and K. Wang. Video-based abnormal human behavior recognitiona review. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(6):865–878, 2012. 1

[18] M. J. Roshtkhari and M. D. Levine. Online dominant and anomalous behavior detection in videos. *Computer Vision and Image Understanding*, 2013. 3

[19] S. Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985. 4

[20] J. Varadarajan and J.-M. Odobez. Topic models for scene analysis and abnormality detection. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1338–1345. IEEE, 2009. 1, 2, 3, 4, 5

[21] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):539–555, 2009. 2, 3

[22] T. Zhang, S. Liu, C. Xu, and H. Lu. Mining semantic context information for intelligent video surveillance of traffic scenes. 2013. 1, 2

[23] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3441–3448. IEEE, 2011. 1