Causality-weighted active learning for abnormal event identification based on the topic model

Yawen Fan
Shibao Zheng
Hua Yang
Chongyang Zhang
Hang Su
Shanghai Jiao Tong University
Department of Electronic Engineering
Institute of Image Communication and Information Processing
Shanghai Key Laboratory of Digital Media Processing and Transmissions
No. 800 Dongchuan Road
Shanghai 200240, China
E-mail: ywen.fan@gmail.com

Abstract. Abnormal event identification in crowded scenes is a fundamental task for video surveillance. However, it is still challenging for most current approaches because of the general insufficiency of labeled data for training, particularly for abnormal data. We propose a novel active-supervised joint topic model for learning activity and training sample collection. First, a multi-class topic model is constructed based on the initial training data. Then the remaining unlabeled data stream is surveyed. The system actively decides whether it can label a new sample by itself or if it has to ask a human annotator. After each query, the current model is incrementally updated. To alleviate class imbalance, causality-weighted method is applied to both likelihood and uncertainty sampling for active learning. Furthermore, a combination of a new measure termed query entropy and the overall classification accuracy is used for assessing the model performance. Experimental results on two real-world traffic videos for abnormal event identification tasks demonstrate the effectiveness of the proposed method. © 2012 Society of Photo-Optical Instrumentation Engineers (SPIE).
[DOI: 10.1117/1.OE.51.7.077204]

Subject terms: video surveillance; topic model; active learning; abnormal event identification.

Paper 120083 received Jan. 21, 2012; revised manuscript received Jun. 6, 2012; accepted for publication Jun. 6, 2012; published online Jul. 6, 2012.

1 Introduction
In video surveillance, automatic abnormal event identification is a fundamental vision task that has recently attracted considerable attention. However, it still remains a difficult problem due to several challenges faced in real-world scenarios, where scenes are cluttered with a multitude of objects and co-occurring activities. First, abnormal events are generally unknown in advance. In addition, both normality and abnormality are diverse. A key issue is the general insufficiency of labeled data for training and validation. This is particularly acute for anomalous data. Furthermore, it is difficult to model rare activities even when the training data are given, as they are visually subtle compared to typical activities. Our focus in this article is identifying such spatial-temporal rare and unexpected abnormal events.

For cases in which collecting sufficient abnormal event training data is impractical and unusual events cannot be defined in advance, traditional supervised learning will often be ineffective and unrealistic. As an alternative, extensive work has focused on unsupervised outlier detection.1–7 In these cases, only a normal event model is learned either automatically from unlabeled data or only based on typical instances. Then the event with low likelihood under the learned model is deemed abnormal. Although it has no requirements for supervision, this approach has crucial limitations in unconstrained public scenes. In addition to limited accuracy, it is unable to distinguish identified abnormal events. Zhang et al.8 proposed a semi-supervised adapted Hidden Markov model (HMM) to overcome the scarcity of training samples for abnormal events. In their approach, both normal event and abnormal event models are derived from a general usual event model in an on-line manner. However, their approach is a binary class output and still intractable for complex activity modeling in crowded public scenes. Hospedales et al.9 proposed a weakly supervised joint topic model (WS-JTM) to address the sparse issues. Different from the traditional supervised learning, this approach permits modeling abnormal behaviors from as few as one sample by using shared common topics. This batch learned multi-class model remains static once trained; thus, there is still the problem of adapting to unexpected abnormal events. Furthermore, exhaustive labeling of a large dataset would be prohibitively expensive.

Given the substantial human effort required to gather effective training samples, labeled training data collection has become a thriving research area in itself.8,11 A promising way to efficiently build up training datasets for model learning is by using an active learning strategy.12 Loy et al.13 presented a stream-based active learning approach to incorporate human feedback for on-line unusual events detection. Hospedales et al.14 proposed a pool-based active learning model to jointly discover new categories and learn to classify them. In these methods, a likelihood criterion selects the most unlikely points, while an uncertainty criterion selects points closest to the decision boundary. But the most confident points, particularly for the abnormal class, are ignored.

Therefore, a novel active-supervised joint topic model is proposed to overcome the above limitations from the two aspects of abnormal activity modeling and abnormal sample collection. The flow chart of the proposed framework is shown in Fig. 1. In the modeling perspective, WS-JTM9 is adopted to build the initial model with some labeled samples, and then the model is incrementally updated during
the learning process. In the sample collection perspective, a novel stream-based active learning algorithm is proposed. In contrast to traditional techniques, the main features of the proposed active learning strategy are outlined as follows: (1) The main goal of the active learning is to automatically select new training samples and decide which annotator should be queried (human or machine itself). (2) Causality-weighted criterion is employed, with more emphasis given to the temporal regions surrounding abnormal events to collect more rare samples. Although topic model is based on the independent and identically distributed (i.i.d) assumption \(^{15}\), the temporal dependence among the training samples is taken into account during the sample collection process. (3) A performance factor is computed based on the combination of a new measure termed query entropy (QE) and the overall classification accuracy on labeled samples. This factor is then used for dynamically updating the query threshold during the active learning process.

The remainder of this paper is organized as follows. Section 2 describes the feature detection and representation employed in this study. Section 3 introduces the topic model structure for the activity modeling task. A novel active learning approach for model training and sample collection is introduced in Sec. 4. In Sec. 5, the effectiveness of the proposed approach is demonstrated through the experiments on two real-world traffic sequences. Conclusions are introduced in Sec. 6, and future work is also described.

2 Video Feature Detection and Representation

In this section, low-level motion features are detected for video representation, since they are usually more reliable in crowded scenes. First, the scene is spatially divided into a \( M \times N \) grid with cells of size \( H \times H \), and the motion direction is quantized into eight orientations. Thus a visual codebook \( V = \{ v_f \}_{f=1}^{N_v} \) of \( N_v = M \times N \times 8 \) visual words is built. The combination of motion direction and position is discriminative enough across activities. To create visual documents, video sequence is temporally segmented into non-overlapping video clips. Pixel-wise optical flow between consecutive frames is computed using TV-L1 algorithm.\(^\text{16}\) Flow vectors with small magnitude are neglected to remove noise. The remaining flow vectors from each clip are mapped into one of the visual words. Finally, its corresponding visual document is composed by the words accumulated over its frames. Activities will be represented by co-occurring visual words.

3 Activity Modeling

3.1 Active-Supervised Joint Topic Model

Topic models have been applied to abnormal event detection in crowded videos and have shown promising results. In this work, the WS-JTM\(^\text{9}\) is employed to construct the initial model. This model (see Fig. 2) can jointly learn models for all the classes (containing normal and abnormal) using
where $X$ denotes the training dataset, and $Z^{-j,i}$ denotes all activities excluding $z_{j,i}$; $n_{z,v}$ denotes the counts of word $x_{j,i}$ associated with activity $z_{j,i}$; $n_{j,z}$ denotes the counts of activity $z_{j,i}$ in clip $j$. In contrast to the fully weakly supervised method, the class labels $c$ of the remaining unlabeled clips are obtained in active learning mode to save human effort, that will be discussed below.

### 3.2 Incremental Model Updating

In contrast to the batch-learned model, which remains static once trained, in this work the model is trained in a stream-based active learning mode, and becomes more accurate via incremental updating. In this study, the incremental learning is performed by using only new samples selected at current iteration for efficiency. But it still retains the accuracy of the model.\(^1\) Model parameters and hidden variables are updated iteratively using the current model and labeled samples. First, run the Gibbs sampler on the new labeled clip for a fixed number of iterations given the model parameters,

$$z_{j,i} \sim p(z_{j,i}|Z^{-j,i}, X, c_{j}, M^{j-1}),$$  
(2)

where $(X, c_{j})$ denotes the current clip and its label. $M^{j-1}$ represents the model parameters learned from the previous $j-1$ samples. After sampling, the per-clip topic parameter $\theta_{j}$ can be yielded:

$$\hat{\theta}_{j,k} = \frac{n_{j,k} + \alpha_{k}}{\sum_{k}(n_{j,k} + \alpha_{k})},$$  
(3)

where $n_{j,k}$ denotes the counts of activity $k$ in the document $X_{j}$. The incremental model updating is then achieved by updating the sufficient statistic with those provided by the new document $X_{j}$. Finally, the topic-word parameter $\phi_{k,v}$ can be estimated:

$$\hat{\phi}_{k,v} = \frac{n_{k,v} + \beta}{\sum_{v}(n_{k,v} + \beta)},$$  
(4)

where $n_{k,v}$ denotes the total counts of word $v$ associated with activity $k$.

### 4 Causality-Weighted Active Learning

#### 4.1 Problem Formulation

Formally, this work considers active learning in a stream-based setting. The goal is actively learning multi-class topic models from a limited amount of weakly labeled dataset and long unlabeled data stream. The problem can be stated as follows:

1. With limited amount of reliable labeled dataset $D_{l} = \{X_{1}, \ldots, X_{l}, \ldots\}$ from normal and some abnormal classes, an initial model is trained.

2. Given an unlabeled data stream $D_{u} = \{X_{1}, \ldots, X_{l}, \ldots\}$, which includes instances from both normal and abnormal classes, we aim to produce a function that can efficiently determine whether current sample should be labeled and which annotator should be queried:
Fan et al.: Causality-weighted active learning for abnormal event identification…

Fig. 3 Flow chart of the stream-based active learning.

- Expert manually label;
- Machine automatically label.

3. Incorporate the new labeled sample into training dataset and update the model.

The important aspect of an active learning algorithm is to identify the significance of samples in the training data stream. From imbalance dataset point of view, missing an abnormal sample is much worse than missing a normal sample. Hence, a causality-weighted method is adopted to introduce a bias toward abnormal classes. The significance of the sample is evaluated by causality-weighted likelihood and uncertainty. Thus, the sample can be classified into one of three types,

1. Confidently unknown sample: sample with low likelihood;
2. Ambiguous sample: sample with high likelihood and high uncertainty;
3. Confidently identified class sample: sample with high likelihood and low uncertainty.

Previous active learning methods often prefer the most informative sample in terms of low likelihood or high uncertainty, and discard the most representative sample in terms of high likelihood and low uncertainty. But given the class imbalance as well as the expectation that more data is almost always advantageous, the representative samples for abnormal classes should also be collected for training.

To recap, the main stream-based active learning loop (see Fig. 3 and Algorithm 1) consists of using the current model to generate candidate samples. Then if the sample is judged unknown or uncertain, it will be labeled by an expert; if the sample is confidently determined as one of the identified abnormal classes, its label predicted by machine will be taken as new ground truth. Finally, the model is incrementally updated by the new labeled sample. There is an assumption that the label predicted by machine is credible. But if it is not true, self-training will reinforce the error. Furthermore, the threshold of likelihood and uncertainty will affect the degree of this risk. In the extreme case, if the likelihood threshold is large enough, no samples will be labeled by machine and used to retrain the model. At the other extreme, if the likelihood threshold equals the minimum while the uncertainty threshold is large enough, all samples are labeled by machine.

Algorithm 1 Stream-based active learning for abnormal event identification.

Require: Unlabeled data stream \( \mathcal{U} = (X_1, X_2, \ldots, X_t, \ldots) \), an initial model \( \mathcal{F} \), trained with a small set of labeled samples \( \mathcal{L} \).

Ensure: A model \( \mathcal{F} \) trained with labeled samples \( \mathcal{L} \).

1: for \( t = 0 \) to the data stream run out do
2: Receive \( X_t \);
3: Compute query criteria \( q_{lik}^t \) and \( q_{unc}^t \);
4: if \( q_{lik}^t < T_{lik} \) then {confidently unknown sample}
5: Query \( X_t \) label, and add \( (X_t, c_t) \) to \( \mathcal{L} \);
6: Update model with \( ((X_t, c_t) \) and threshold \( T_{lik} \);
7: else if \( q_{unc}^t > T_{unc} \) then {ambiguous sample}
8: Query \( X_t \) label, and add \( (X_t, c_t) \) to \( \mathcal{L} \);
9: Update model with \( (X_t, c_t) \) and threshold \( T_{unc} \);
10: else if the predicted label \( c_t \in c_{anom} \) then {confidently identified abnormal class sample}
11: Add \( (X_t, c_t) \) to \( \mathcal{L} \);
12: Update model with \( (X_t, c_t) \) and threshold \( T_{lik} \);
13: end if
14: end for
15: Identify testing samples \( \mathcal{F} \) with model \( \mathcal{F} \);

4.2 Causality-Weighted Method

Activities occurring in scenes are implicitly temporally ordered. The occurrence of one abnormal event may arouse subsequent abnormal events (e.g., a fire engine’s interruption followed by slight traffic chaos at the junction). Furthermore, unlike in text analysis, a common situation in visual scenes is that visual documents are generated by segmenting the sequence into non-overlapping short length clips. So, one complete abnormal event would happen in several successive clips. And video clip is the basic processing unit for activities modeling. These two facts taken together, training clips are non-i.i.d; indeed, temporal dependencies are important clues for collecting sparse abnormal samples. Furthermore, causality is a proper concept in describing such time dependencies between events.17
Due to the fact that anomalies are a small fraction of the data as well as the expectation that more data are usually advantageous, our active learning algorithm exploits a causality-weighted method in order to increase the likelihood of sampling interesting anomalies in future iterations. The main idea is that the query score is biased to the time region around the abnormal clip. Therefore, the query score \( q_i^{\text{query}} \) at time \( t \) is defined as follows:

\[
q_i^{\text{query}} = \psi(X_t) \cdot (1 + \lambda \cdot \omega \delta(c_{t-1}, C_{\text{abm}})),
\]

where \( X_t \) is the current clip, and \( \psi(X_t) \) represents the query score of \( X_t \) according to some base query strategies, such as likelihood or uncertainty criterion. The second term weights the score by the previous clip’s label, \( \lambda \) is a parameter that controls the relative importance of the causality term. Specifically, the second term is set to 1, if \( c_{t-1} \) represents normal class with more emphasis given to the temporal region surrounding unusual event classes to address the problem of sparseness. \( \omega \) denotes the causality degree, which is quantified by the average similarity between the sample \( X_t \) and the other samples in class \( c_{t-1} \).

\[
w = \operatorname{sim}(X_t, d_{c_{t-1}}) = \frac{1}{N} \sum_{X_s \in d_{c_{t-1}}} f(X_s, X_t),
\]

where \( d_{c_{t-1}} \) denotes the training dataset of abnormal class \( c_{t-1} \), and \( N \) is the size of \( d_{c_{t-1}} \). In obtaining the pairwise similarity \( f(X_s, X_t) \), the traditional Hellinger distance is adopted, that is,

\[
f(X_s, X_t) = 1 - \sqrt{1 - \sum_k \sqrt{\theta_i^s(k) \cdot \theta_i^t(k)}},
\]

where \( \theta_i^s \) is the topic distribution of \( X_s \), and \( \theta_i^t \) is the predicted topic distribution of \( X_t \) based on the current class \( c_{t-1} \) model.

### 4.3 Likelihood Criterion

For unlabeled clip \( X_t = \{x_i\}_{i=1}^{N_t} \), its marginal likelihood for each model (normal and abnormal) is defined as,

\[
p(X_t | c, \alpha, \phi) = \int \sum_{Z} p(X_t, Z | \theta, \phi) p(\theta | \alpha) d\theta.
\]

Due to the intractable of sum over correlated \( Z \), the importance sampler proposed in Ref. 9 is adopted to approximate the marginal likelihood,

\[
p(X_t | c) \approx \frac{1}{S} \sum_{S} p(X_t, Z^S | \alpha, \phi) q(Z^S | c),
\]

where the conditioning on the parameters is dropped for clarity. To approximate the optimal proposal \( q_o(Z | c) \), mean field estimator with minimal Kullback-Leibler divergence is developed by iterating,

\[
q_i(z_j | c) \propto \left(\alpha_c^j + \sum_{j \neq i} q_i(z_j | c) \phi_{z_j, z_i} \right).
\]

The adaptation of likelihood threshold will be discussed in Sec. 4.5.

### 4.4 Uncertainty Criterion

For the sample with high likelihood that does not pass the likelihood criterion, it is further evaluated by the uncertainty criterion. In order to increase the accuracy of the classifier, the items with high uncertainty score are also selected for human labeling. Based on the method presented in Ref. 18, we calculate the uncertainty score using the Query By Committee (QBC) algorithm. The committee members here are \( N \) models selected by drawing a set of parameters \( \Phi_n \) after the Gibbs sampling burn in during the training process. The label of the current sample \( X_t \) predicted by one committee member is defined as:

\[
c_n = \arg \max_{k \in \{1, \ldots, K\}} p(c = k | X_t, \theta, \Phi_n).
\]

For measuring the level of disagreements among committee members, the vote entropy of sample \( X_t \) is denoted as:

\[
q_i^u = - \sum_{k=1}^K \frac{V(c = k)}{M} \log \left( \frac{V(c = k)}{M} \right),
\]

where \( V(c = k) \) represents the deterministic votes that a label receives from committee members, and \( M \) is the total number of committee members. The higher the entropy, the more uncertain a point is.

### 4.5 Adapting Likelihood Threshold

Since the model is updated in every iteration, using a constant number of likelihood threshold is not suitable. Inspired by the multi-class entropy (MCE)\(^{14}\) which requires access to a pool of unlabeled dataset, we propose a new method defined as QE based upon the human feedback and take it as indication of sample collection performance,

\[
H = - \frac{n}{N_q} \sum_{k=1}^N \sum_{c \neq c_k} \delta(c_j, c_k) \log \frac{\sum_{c} \delta(c_j, c_k)}{N_q},
\]

where \( N_q \) is the number of classes observed so far, and \( N_q \) is the number of samples labeled so far. The intuition here is that in a rare-class scenario with extreme class imbalance, classifiers are typically at risk of bias toward the majority class. A classifier with a higher entropy on the query data shows more bias to rare-class.

In addition, we measure the model classification performance using the overall classification accuracy during the active learning process,

\[
p = \frac{\sum_{c} N_q \delta(c_k, c_j)}{N},
\]

where \( N \) is the total number of training samples, \( c_k \) is the label predicted by the model, and \( c_j \) is the true category labeled by an expert. The assumption is that the performance
of the model is improved along with the learning iteration. Then the overall performance at time $t$ is defined as:

$$\rho_t = \mu H_t + (1 - \mu) p_t,$$

(15)

where parameter $\mu$ is the mixture prior for the two terms, and the range of $\rho_t$ is from 0 to 1. The first term on the right side above rewards the collection of rare-class samples, and the second term rewards an increase of classification accuracy. Given this overall performance measure, we define an update for the future likelihood threshold,

$$T_{h_t+1} = (1 - \Delta \rho_t) T_{h_t}, \quad \text{where} \quad \Delta \rho_t = \rho_t - \rho_{t-1}.$$  

(16)

In early learning iterations, the model is far from accurate, so the query is more likely decided by the likelihood criterion. Practically, the threshold is updated at set intervals for robustness.

### 5 Experimental Results

#### 5.1 Dataset and Setting

In this section, the proposed algorithm was tested on two public traffic videos from the QMUL dataset\(^9\) (360 $\times$ 288, 25 fps, 1 h) and the MIT dataset\(^2\) (720 $\times$ 480, 30 fps, 1.5 h). The experiments were performed on Intel Core i7 to 2600 machine with 3.4 GHz CPU and 4 G RAM, and implemented in hybrid programming with C and Matlab. These datasets contain multiple complex motion patterns, and also exhibit different environmental effects. Similar to what was reported in Ref.\(^9\), for the QMUL dataset this study is also interested in the U-turn at the center of the scene and the near-collision situation in which horizontal flow vehicles drive into the junction before turning traffic finishes. For the MIT dataset, the targets are the right turn at the right zone and the left turn at the bottom zone.

The videos were spatially split into 10 $\times$ 10 cells and the motion direction was quantized into eight orientations. Therefore, the sizes of the visual codebooks were 8352 ($36 \times 29 \times 8$) and 27648 ($72 \times 48 \times 8$), respectively. Each video was temporally segmented into 3-s-long clips. We manually labeled the clips into typical and unusual classes. The ground truth was used as feedback returned to model when querying for labels during the learning process. It was also employed for comparison during testing phase. The clips were partitioned into training and testing sets; the number of clips are detailed in Tables 1 and 2. Optical flow between consecutive frames was computed and the magnitude less than a threshold (0.8 for the QMUL dataset and 0.5 for the MIT dataset) was ignored. For active and fully supervised learning models, we used 20 normal topics, and one topic per abnormal class. For the unsupervised learning model, the topic number was 20.

#### 5.2 Analysis of Model Learning

The influence of likelihood threshold on the number of queried samples to be appended to the training dataset is shown in Fig. 4. Basically, the total number of queried samples increases along with the increase of the threshold. The higher

![Fig. 4 Influence of likelihood threshold on the number of queried samples.](image-url)
the threshold is, the more number of samples will be selected by likelihood criterion, while the ratio of samples selected by uncertainty criterion is lower. In the experiments, the uncertainty threshold is fixed.

Figure 5 shows the model overall performance $\rho_t$ changes over time. As expected, the general trend is a gradual increase. The model performance is improved with the increase of the number of labeled points.

As can be seen from Fig. 6, causality-weighted method shows better performance in collecting abnormal event samples, including known and unknown classes. This indicates that the proposed algorithm effectively guides the model toward abnormal event sample collection.

5.3 Activities Modeling

In this subsection, the model learned for different class behaviors is analyzed. As can be seen from Figs. 7 and 8, the topic mixture proportions $\theta_j$ over 22 topics of the example clips are represented by bars. The dominant topics (larger than 0.05) are illustrated by the most likely visual words. Different colors indicate typical and rare activities, respectively.

Figure 7 shows the topic distributions of six video clips chosen from the QMUL training dataset. Figure 7(a) explains traffic moving in a vertical direction; Fig. 7(b) represents turning traffic with some vertical traffic; Fig. 7(c) and 7(d) represent rightward and leftward traffic flows, respectively; Fig. 7(e) represents U-turn activity; Fig. 7(f) shows near-collision activity. As observed in Fig. 7(e), the learned U-turn topic (magenta color) is disambiguated from ongoing typical activities. While clearly in Fig. 7(f), the learned near-collision topic (yellow color) accounts for a large proportion, since the near-collision motion flow is a combination of partial horizontal and turn traffic. Thus, the learned model captures the essence of U-turn activity, but near-collision behavior is not yet perfectly modeled in turn affecting the performance of identification.

For the MIT dataset, eight video clips are chosen and shown in Fig. 8. Specifically, Fig. 8(a)–8(f) represent typical horizontal and vertical traffic flows with and without interleaved turning traffic. Figure 8(g) and 8(h) represent rare left and right turn traffic flows, respectively. Compared with the QMUL dataset, the traffic phases of the MIT dataset are less distinctive visually. As a consequence, learning typical
activities becomes harder. However, the model can explain the rare activities well, as shown in Fig. 8(g) and 8(h).

5.4 Performance of Abnormal Event Detection

The performance of the proposed approach on detecting abnormal events in testing dataset is compared with unsupervised and supervised learning approaches. For these three approaches, initial models were trained on the same dataset of size 100. Then for the proposed approach, $N$ new training samples were actively selected following the procedure in Sec. 4. For both unsupervised and supervised learning approaches, the first $N$ clips were sequentially selected from the unlabeled data stream. We experimented with varying training set size. The abnormal event detection rate is defined as follows:

$$\text{Abnormal event detection rate} = \frac{\text{number of correctly detected abnormal class samples}}{\text{total number of abnormal class samples}}.$$ (17)

The accuracy of abnormal event detection is first evaluated by using area under the receiver operating characteristic curve (AUROC) against the size of training set. As shown in Fig. 9, with the same number of training samples, the proposed method outperforms the other two methods. The detection rate of unsupervised learning approach changes slightly with respect to different numbers of training samples. This indicates that training samples collected by our method are more effective. Receiver operating characteristic (ROC) curves were obtained by varying the false alarm rate. Moreover, the performance is also evaluated using AUROC against the likelihood threshold. As shown clearly in Fig. 10, performance is directly influenced by threshold.

To provide further insights on the performance difference, the ROC curves are plotted in Fig. 11. For the QMUL dataset, in case of 300 training samples [Fig. 11(a)], it is clear that the abnormal event detection rate of the proposed approach is always greater than the other two until the false alarm rate is bigger than 0.5. At that point, the performance has some decrease. However, the three methods are at par when the false alarm rate is bigger than 0.7. In case of 700 training samples, performance of the proposed approach is inferior than the unsupervised learning approach when the false alarm rate is bigger than 0.55 [Fig. 11(b)]. For the MIT dataset, the proposed approach can always produce superior performance than the other two [Fig. 11(c) and 11(d)].
5.5 Performance of Event Classification

In this section we compare the classification performance of the proposed method against the fully supervised learning method. In each case results are quantified in terms of the average classification accuracy for each class. This ensures that errors of each type are weighted equally, although the test data is imbalanced. The average classification accuracy $T_{\text{average}}$ is defined as:

$$T_{\text{average}} = \frac{1}{N} \sum_{i} \frac{\text{number of correctly identified class } i \text{ samples}}{\text{total number of class } i \text{ samples}}.$$  

As illustrated in Fig. 12, the average classification accuracy improvement in this task is impressive especially with
Fig. 10 Abnormal event detection accuracy given varying likelihood threshold.

Fig. 11 ROC curves using different methods. (Quantity in brackets indicates AUROC.)
small size of training data. For example, as observed in Fig. 12(a), we need only 350 training samples collected by the proposed approach to attain 70.26% of accuracy, almost the same as the supervised approach accuracy (69.37%) that used up to 550 training samples.

We thus also evaluated full classification performance using normalized confusion matrix. Figure 13 shows the classification confusion matrices in case of 350 (QMUL) and 450 (MIT) training samples. Since the intra-class diversity of typical event was taken into account and more abnormal samples were collected, our method results in high true positive for both normal and most abnormal event identification. But it is obvious that the true positive of the near-collision event is not improved. There may be two reasons. First, near-collision event samples are very intra-class diverse. Secondly, the model for this class is far from perfect. However, we note that the proposed approach still performs significantly better than the supervised algorithm.

Fig. 12 Average classification accuracy with a uniform prior distribution.

Fig. 13 Classification confusion matrices with a uniform prior distribution. (a)-(b) QMUL: 350 training samples; (c)-(d) MIT: 450 training samples.
6 Conclusion
Abnormal event identification is difficult in crowded outdoor environments, especially in situations where abnormal events are rare and unpredictable. In this paper, a new approach is presented for training sample collection and model learning. The active learning framework not only overcomes the limitation of exhaustive human labeling, but also makes the model more robust and effective. The experimental results on two real-world traffic sequences show that the proposed approach achieves diverse training dataset collection, especially for abnormal events. In addition, the approach can provide competitive performance with the supervised approach while saving the human labor.

In future work, more experiments on different datasets will be carried out to evaluate the performance of the proposed approach. We will also further improve the performance by refining the model learning step, especially for some rare behaviors (e.g., near-collision). In addition, the optimal proportion between the typical samples and the abnormal samples in the training dataset should also be explored.

Acknowledgments
This work was supported by NSFC (No. 61171172, No. 61102099, No. 61001147) and STCSM (No. 10231203102).

References

Yawen Fan received her BS and MS degrees in electronic engineering from Hohai University, Nanjing, China in 2003 and 2006, respectively. She is currently pursuing a PhD degree at the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai, China. Her current research interests include computer vision, pattern recognition, and machine learning.

Shibao Zheng received his BS and MS degrees from Xidian University, Xi’an, China in 1983 and 1986, respectively. From 1986 to 1999, he was an expert for the National Project in HDTV. Now he is a professor in the Department of Electronic Engineering, Shanghai Jiao Tong University. He has done a lot of research work and made great progress in the field of intelligent video analysis and video surveillance systems. His research interests include DTV, intelligent video surveillance, and network multimedia.

Hua Yang received her BS and MS degrees from Haerbin Engineering University, China, in 1998 and 2001, respectively. She has received the doctor degree in EE department from Shanghai Jiao Tong University, Shanghai, China. She is now an associate professor at Shanghai Jiao Tong University. Her research interests are intelligent video surveillance and video analysis and understanding.

Chongyang Zhang received his PhD degree from Shanghai Jiao Tong University, Shanghai, China, in 2008. He is now an associate professor at the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai, China. His research interests include digital video coding, video communication, intelligent video surveillance, and multimedia signal processing.

Hang Su student member of IEEE, received his BS and MS degrees from the Department of Electronic Engineering at Shanghai Jiao Tong University in 2007 and 2009, respectively. He is currently a PhD candidate in Department of Electronic Engineering at Shanghai Jiao Tong University. He is currently working on computer vision applications in video perception systems. He has gained fruitful experience in video surveillance, crowd analysis, and human motion analysis.