Stream-based Active Unusual Event Detection

Chen Change Loy, Tao Xiang and Shaogang Gong

School of EECS, Queen Mary University of London, United Kingdom
{ccloy,txiang,sgg}@eecs.qmul.ac.uk

Abstract. We present a new active learning approach to incorporate human feedback for on-line unusual event detection. In contrast to most existing unsupervised methods that perform passive mining for unusual events, our approach automatically requests supervision for critical points to resolve ambiguities of interest, leading to more robust and accurate detection on subtle unusual events. The active learning strategy is formulated as a stream-based solution, i.e. it makes decision on-the-fly on whether to query for labels. It adaptively combines multiple active learning criteria to achieve (i) quick discovery of unknown event classes and (ii) refinement of classification boundary. Experimental results on busy public space videos show that with minimal human supervision, our approach outperforms existing supervised and unsupervised learning strategies in identifying unusual events. In addition, better performance is achieved by using adaptive multi-criteria approach compared to existing single criterion and multi-criteria active learning strategies.

1 Introduction

Video surveillance data is typically characterised by highly imbalanced class distribution, i.e. most of the samples corresponding to normal event classes whilst the remaining unusual event (rare or abnormal events that should be examined further) classes only constituent a small percentage of the entire dataset. In addition, normal patterns are often known *a-priori*, whilst the unusual events are unforeseeable. Consequently, most unusual event detection methods [1–4] employ outlier detection strategy, in which a model is trained using normal events through unsupervised one-class learning and events that deviate statistically from the resulting normal profile are deemed unusual. This strategy offers a practical way of bypassing the problems of imbalanced class distribution and inadequate unusual event training samples. However, the unsupervised nature of this outlier detection methods is subject to a few inextricable limitations:

1. Difficulty in detecting unusual events whose distributions are partially over-lapped with normal events. Specifically, in a busy public scene, unusual events are visually similar to a large number of normally behaving objects co-existing in a scene (see Fig. 1 for example). Without human supervision, it is hard to spot these subtle unusual events.
2. No subsequent exploitation of flagged unusual events. Outlier detection approach is therefore less effective in distinguishing the true unusual events from noise.
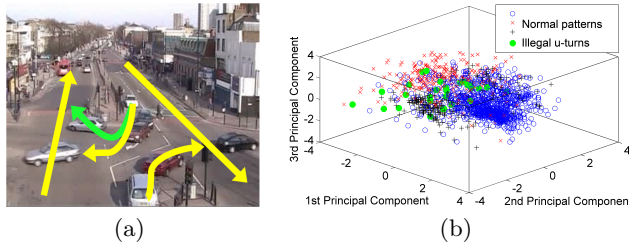
Fig. 1. An example of illegal u-turn event (Fig. 1(a)). It is subtle due to its visual similarity with large number of co-occurring normal patterns in a scene. This can observed from a plot in a principal component analysis space (Fig .1(b)), where similar u-turn cases (plotted as green dots) are partially overlapped with other normal patterns.

3. Large amount of uninteresting outliers causing false alarms. Normal behaviour patterns in a public scene are complicated and highly diverse. Hence, preparation of well-defined and complete normal data for off-line learning becomes unfeasible. Training a model using incomplete normal patterns is likely to result in a large amount of uninteresting outliers, since some outlying regions of the normal class may be consistently and wrongly flagged as unusual event.

In most video surveillance tasks, human knowledge is readily available in practise to remedy the aforementioned issues. Although it is unfeasible to label every single instances at hand, it is still desirable to make use of occasional human inputs for guiding the creation of an activity model. In particular, human inputs may exist in the form of feedback, i.e. indicating the exact event classes or whether a particular detection is right/wrong, when the activity model encounters difficulty in distinguishing an equivocal event or subtle unusual activity. The feedback would be extremely useful to resolve ambiguities of interest and to strengthen the decision boundary of activity classes on what is normal/abnormal, leading to more robust detection on inconspicuous and unknown unusual events.

Active learning strategy emerges as an compelling alternative to conventional supervised and unsupervised unusual event detection methods, since it is capable of seeking human feedback automatically on critical instances to improve event classification performance according to some predefined query criteria [5]. Note that it differs from supervised or semi-supervised strategies, which perform random labelling that treats all samples equally, whilst in essence, not all samples are critical for learning the correct decision boundary. In this study, we formulate a novel *stream-based active learning* strategy with several key features outlined as follows: (1) The method is formulated as a stream-based approach to ensure real-time response, i.e. the model makes immediate decision on whether to query for labels as new video data are streamed in. (2) Multiple criteria are employed for joint exploration and exploitation. In particular, some classes, especially unusual event classes have to be discovered (exploration) since they are not available in the early stage of training. At the same time, it is necessary to gradually improve the model by refining the decision boundary (exploitation). Thus, different criteria are needed to achieve these goals. (3) Query criterion is

adaptively selected from multiple criteria. This is important because good active learning criteria are *dataset dependent* [5]. Importantly, we typically do not know the best suited criterion for a specific dataset at different phases of learning.

In our approach, the first query criterion is a *likelihood criterion*, which favours samples that have low likelihood w.r.t. the current model. Consequently, unknown classes or unexplored regions of existing classes can be discovered. Note that our method does not assume availability of predefined classes, i.e. once a new class is discovered, the model will expand itself automatically. The second criterion is an *uncertainty criterion* based on a modified Query-by-Committee (QBC) algorithm [6–8]. It is used to refine the decision boundary by selecting controversial samples in uncertain regions that give rise to the most disagreement among classes, with more emphasis given to the regions surrounding unusual event classes to address the problem of imbalance class distribution. The two query criteria are dynamically re-weighted based on the Kullback-Leibler (KL) divergence [9] measured on the model before and after it is trained using a queried sample. The premise behind this adaptive weighting scheme is to favour the criterion that is more likely to return a queried sample that brings most influence to the current model.

Comparative experiments are carried out on busy public space surveillance videos. We show that by exploiting a small cost of human supervision through active learning, more robust and accurate detection of subtle unusual events is achieved compared to conventional supervised and unsupervised learning methods. In addition, the results also suggest that our adaptive multi-criteria approach outperforms single criterion and multi-criteria methods we evaluated.

2   Related Work

Most existing unusual event detection methods follow *unsupervised one-class learning strategy* by employing different models such as topic models [1–3] and Markov random field [4]. On the other hand, there are several studies that perform event classification [10,11] based on *supervised strategy*. Our method differs significantly from these methods in that our approach is capable of discovering unknown classes and resolve inter-class ambiguities by exploiting human feedback. It is thus more suitable for on-line mining of unusual events. It is worth pointing out that Sillito and Fisher [12] attempt to incorporate human feedback using a one-class semi-supervised model. However, their method is limited to learning the normal event class. Therefore, it is still facing the same problems encountered by unsupervised approaches (see Sec. 1).

In the active learning perspective, most studies to date assume pool-based setting [13–16], which requires access to a fixed pool of unlabelled data for searching the most informative instance for querying. For surveillance task since activity patterns are dynamic and unusual events are often unpredictable, preparing a pool of unlabelled data that encompasses complete event classes is impractical. Moreover, performing exhaustive search in the pool is expensive therefore unsuitable for surveillance task that demands real-time performance. Stream-based

setting is preferred in this context as it is capable of making immediate query decision without the need of accessing a data pool.

Most existing stream-based approaches are based on single query criterion [6, 8, 17], which are obviously not sufficient for exploration and exploitation that pursue different goals in nature. Even though there are attempts in combining multi-criteria for active learning, they are either not adaptive [13, 14] or limited to pool-based setting [16, 18]. Non-adaptive methods (e.g. iterate over different criteria with constant weights) cannot apply the right criteria at different phases of learning, e.g. the active learner may waste effort refining the boundary before discovering the right classes, or vice versa. Methods proposed by Baram et al. [16] and Cebron and Berthold [18], though adjusting weights of different criteria online, they require access to a pool of unlabelled data, which are often unavailable to stream-based environments.

Our uncertainty criterion is based the QBC algorithm [6, 8], in which an ensemble of committee members are maintained. Query will be triggered if class label of a sample is controversial among the members. Various measures of disagreement have been proposed [6, 7, 19]. These measures, however, only return the disagreement score among members without identifying conflicting classes, i.e. the classes closest to the uncertain point. We formulate a new QBC scoring method to identify conflicting classes, thereby incorporate a prior constraint to favour uncertain samples surrounding unusual event classes, leading to a more balanced sample selection for class imbalanced data.

In summary, the main novelties and contributions of this study are:

1. We propose a new active learning approach to incorporate crucial human supervision to resolve ambiguities for more robust and accurate unusual event detection over conventional unsupervised and supervised approaches. To the best of our knowledge, this problem has not been addressed before.
2. We introduce a new adaptive weighting scheme suitable for combining multiple query criteria in a stream-based setting. This method does not need to access a fixed pool of unlabelled data.
3. To have a more balanced sample selection, we introduce prior to constrain uncertainty criterion to favour unusual classes during decision boundary formation. For this purpose, a new QBC scoring method is formulated to identify conflicting classes.

3   Active Unusual Event Detection

We consider active learning in a stream-based setting in which an unlabelled sample $\mathbf{x}_t$ is observed at each time step t from an input data stream $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots)$. Consequently, a classifier $\mathcal{C}_t$ is required to determine on-the-fly whether or not to query for label y_t or discard $\mathbf{x}_t$. Our goal is to select critical samples from $\mathcal{X}$ for annotation to achieve two tasks simultaneously: (1) to discover unusual event classes or unknown region of existing classes in the input feature space and (2) to refine the classification boundary with higher priority being given to regions surrounding the unusual classes so as to improve the detection accuracy of unusual events.

3.1   Activity Representation

We wish to represent activity patterns using location-specific motion information over a temporal window without relying on object segmentation and tracking. This is achieved through the following steps: (1) Given an input video, we extract optical flow in each pair of consecutive frames using [20]. (2) A method similar to that in [21] is employed to automatically decompose a complex scene into D regions, $\mathbf{r} = \{r_i | i = 1, \ldots, D\}$ according to the spatial-temporal distribution of motion patterns observed (Fig. 2(b,d)). (3) Motion direction of each moving pixel in each region are quantised into four directions and put into bins. (4) A histogram hist_{f,r_i} with a size of four bins is constructed for each region r_i in each frame f. We uniformly divide the whole video sequence into non-overlapping clips, each having 50 frames in length. We then sum up individual bins of a regional histogram within each clip t as $\mathrm{hist}_{t,r_i} = \sum_{f \in \mathrm{clip}\ t} \mathrm{hist}_{f,r_i}$. (5) Non-dominant motion directions are then removed as they are more likely to be caused by error in optical flow computation[1]. (6) Finally, the histogram is discretised to construct a codebook with 16 words ω_j, $j \in \{1, 2, ..., 16\}$, representing the dominant motion directions of each region. For example, word ω_1 represents motionless region, word ω_2 means only direction bin 1 is observed, and word ω_4 indicates both occurrence of direction bins 1 and 2, etc. Consequently, the ith region of the tth clip is represented as a variable $x_{i,t}$ of 16 possible discrete values x_{ij} according to its word label and the clip is denoted as $\mathbf{x}_t = (x_{1,t}, \ldots, x_{D,t})$.

3.2   Bayesian Classification

We wish to classify the D-dimensional observed vector $\mathbf{x} = (x_1, \ldots, x_D)$ into one of the K classes, where a class variable is represented by $y = k \in \{1, \ldots, K\}$. We approach the classification task as Bayesian classification. To facilitate efficient incremental learning, we employ a naïve Bayesian classifier with Bayesian learning by assuming conditional independence among the distributions of input attributes $x_1, \ldots, x_D$ given the class label. The classifier is quantified by a parameter set $\boldsymbol{\theta}$ specifying the conditional probability distributions (CPDs). We assume separate multinomial distribution $p(x_i|y)$ on each x_i for each class label. Consequently, we use $\boldsymbol{\theta}_{x_i|y}$ to represent a vector of parameters $\theta_{x_{ij}|y}$ for multinomial $p(x_i|y)$. Given the multinomial CPDs, the conditional probability $p(\mathbf{x}|y = k)$ for an observed vector given class $y = k$ is given as $p(\mathbf{x}|y = k) = \prod_{i=1}^{D} p(x_i|y = k)$. Given $p(\mathbf{x}|y)$ and $p(y)$, posterior conditional distribution $p(y|\mathbf{x})$ can be computed via Bayes rule. A class y^* that best explains $\mathbf{x}$ can then be obtained as follows:

$$y^* = \operatorname*{argmax}_{k \in \{1, \ldots, K\}} p(y = k|\mathbf{x}) = \operatorname*{argmax}_{k \in \{1, \ldots, K\}} p(y = k)p(\mathbf{x}|y = k). \qquad (1)$$

Incremental learning - Efficient incremental learning is required for stream-based active learning. Since we have fully observed data, we use conjugate prior

---

[1] The four direction bins are ranked in a descending order based on their values. The dominant motion directions is identified from the first few bins in the rank that account for a given fraction $P \in [0, 1]$ of total bin values ($P = 0.8$ in this study). Motion directions in the remainder of the bins are considered as non-dominant.

to facilitate efficient Bayesian learning. The conjugate prior of a multinomial distribution with parameters $\boldsymbol{\theta}_{x_i|y}$ is the Dirichlet distribution, which is given as:

$$\text{Dir}(\boldsymbol{\theta}_{x_i|y} \mid \boldsymbol{\alpha}_{x_i|y}) \propto \prod_j [\theta_{x_{ij}|y}]^{\alpha_{x_{ij}|y}-1}, \tag{2}$$

where $\alpha_{x_{ij}|y} \in \mathbb{R}^+$ are hyper-parameters of the distribution.

3.3   Query Criteria

In stream-based setting, the query decision is typically determined by a query score p^{query} derived from a query criterion Q. The query score will be compared against a threshold Th. Specifically, if $p^{\text{query}} \geq \text{Th}$, query is made; otherwise $\mathbf{x}_t$ is discarded. In this study, we propose to employ two widely used criteria with clear complementary nature, namely likelihood criterion and uncertainty criterion for joint unknown event discovery and classification boundary refinement. Next we formulate methods to compute the respective query scores based on these criteria.

Likelihood criterion - Using this criterion a point is selected by comparing its likelihood against current distribution modelled by the classifier. In particular, given a sample $\mathbf{x}$, we first find a class y^* that best explains the sample according to Eqn. (1). Secondly, for each feature node, we compute the normalised probability score of x_i given y^*:

$$\hat{p}(x_i|y^*) = \frac{p(x_i|y^*) - \mathrm{E}\left[p(x_{ij}|y^*)\right]}{\sqrt{\mathrm{E}\left[p(x_{ij}|y^*) - \mathrm{E}\left[p(x_{ij}|y^*)\right]\right]}}.$$

The normalised probability score $\hat{p}(x_i|y^*)$ is bounded to ensure $-0.5 \leq \hat{p}(x_i|y^*) \leq 0.5$. Finally, the likelihood score at time step t is calculated as:

$$p_t^l = 1 - \left(\frac{1}{2} + \frac{1}{D}\sum_{i=1}^{D} \hat{p}(x_i|y^*)\right). \tag{3}$$

The likelihood score lies within [0,1]. If p_t^l of a sample is closer to 1, it is more likely to be queried.

Uncertainty criterion - Our uncertainty criterion is re-formulated from the existing QBC algorithm [6,7], with additional consideration on conflicting classes for yielding a more balanced sample selection.

(1) *Generating a committee* - Given a classifier $\mathcal{C}_t$ and training data $\mathcal{S}_t$, we generate M committee members corresponding to hypotheses $\mathbf{h} = \{h_i\}$ of the hypotheses space $\mathcal{H}_t$, where each hypothesis is consistent with the training data seen so far [8], i.e., $h_i \in \mathcal{H}_t | \forall(\mathbf{x}, y) \in \mathcal{S}_t, h_i(\mathbf{x}) = y$.

In a naïve Bayes classifier with multinomial CPDs, this can be done by sampling new parameters from the posterior Dirichlet distribution of classifier [6,7]. It has been proven that parameters of a Dirichlet distribution can be generated

from a Gamma distribution (Chapter XI, Theorem 4.1 in [22]). Consequently, we sample $\hat{\boldsymbol{\theta}}_{x_i|y}$ from its posterior Dirichlet distribution $\mathrm{Dir}(\hat{\boldsymbol{\theta}}_{x_i|y}|\boldsymbol{\alpha}_{x_i|y})$, by drawing new weights $\hat{\alpha}_{x_{ij}|y}$ from the Gamma distribution, i.e. $\hat{\alpha}_{x_{ij}|y} \sim \mathrm{Gam}\left(\alpha_{x_{ij}|y}\right)$. The parameter of a committee member is then estimated as:

$$\hat{\theta}_{x_{ij}|y} = \frac{\hat{\alpha}_{x_{ij}|y} + \lambda}{\sum_j \left(\hat{\alpha}_{x_{ij}|y} + \lambda\right)}. \tag{4}$$

where λ is a weight added to compensate data sparseness, i.e. to prevent zero probabilities for infrequently occurring values x_{ij}.

(2) *Measure of disagreement* - As discussed in Sec. 2, existing approaches of measuring member disagreement are not able to return the corresponding classes that cause the most disagreement. In this study, we formulate a new uncertainty score as follows: first, a class disagreement score is computed over all possible class labels:

$$s_{y=k,t} = \left\{ \max_{h_i \in \mathcal{H}_t, h_j \in \mathcal{H}_t} [p_i(y = k|\mathbf{x}_t) - p_j(y = k|\mathbf{x}_t)] \right\}, \tag{5}$$

where $i \neq j$. Consequently, the top two classes that return the highest $s_{y=k,t}$ are identified as c_1 and c_2. The final uncertainty score is computed as:

$$p_t^u = \frac{1}{2}.\gamma_u. \left[s_{y=c_1,t} + s_{y=c_2,t}\right], \tag{6}$$

where γ_u is the prior introduced to favour the learning of classification boundary for unusual classes. Specifically, γ_u is set to a low value if c_1 and c_2 are both normal event class and a high value if any one of c_1 and c_2 is unusual event class. If p_t^u of a sample is closer to 1, it is more likely to be queried.

3.4   Adaptive Selection of Multiple Query Criteria

As explained in Sec. 1, adaptive selection of multiple criteria is necessary for joint unknown event discovery and classification boundary refinement. In particular, different criteria can be more suitable for different datasets as well as different learning stages. Since we usually do not know the right choice *a priori*, selecting different criteria adaptively has the potential to provide a more reliable and even more optimal solution than using any single criterion alone.

To this end, we formulate an adaptive approach in selecting different query criteria for stream-based active learning. Specifically, given multiple query criteria $\mathcal{Q} \in \{Q_1, \ldots, Q_a, \ldots, Q_A\}$, a weight $w_{a,t}$ is assigned to each query criterion Q_a at time step t. A criterion is then chosen by sampling from a multinomial distribution, $a \sim \mathrm{Mult}(\mathbf{w}_t)$, where $\mathbf{w}_t \in \{w_{1,t}, \ldots, w_{a,t}, \ldots, w_{A,t}\}$.

The weights $\mathbf{w}$ are guided by the change in distribution modelled by our naïve Bayes classifier before and after it is updated using a newly queried sample. Intuitively, a criterion is preferred, therefore being assigned higher weight if it asks for samples that give greater impact to the existing distribution modelled by the

classifier. To measure the distance between two distributions $p_{\boldsymbol{\theta}}(\mathbf{x})$ and $p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})$, we employ the KL-divergence, which is given as $\mathcal{KL}(\boldsymbol{\theta} \parallel \tilde{\boldsymbol{\theta}}) = \sum_{\mathbf{x}} p_{\boldsymbol{\theta}}(\mathbf{x}) \ln \frac{p_{\boldsymbol{\theta}}(\mathbf{x})}{p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}$.

Algorithm 1: Stream-based active unusual event detection.

Input: Data stream $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots)$, an initial classifier $\mathcal{C}_0$ trained with a small set of labelled samples from known classes

Output: A set of labelled samples $\mathcal{S}$ and a classifier $\mathcal{C}$ trained with $\mathcal{S}$

1 Set $\mathcal{S}_0 = $ a small set of labelled samples from known classes ;

2 for t *from* $1, 2, \ldots$ *until the data stream runs out* **do**

3   Receive $\mathbf{x}_t$;

4   Compute p_t^l (Eqn. (3)) ;

5   Compute p_t^u (Eqn. (6)) ;

6   Select query criterion by sampling $a \sim \text{Mult}(\mathbf{w})$, assign p_t^{query} based on the selected criterion ;

7   **if** $p_t^{\text{query}} \geq \text{Th}$ **then**

8   Request y_t and set $\mathcal{S}_t = \mathcal{S}_{t-1} \bigcup \{(\mathbf{x}_t, y_t)\}$;

9   Obtain classifier $\mathcal{C}_{t+1}$ by updating classifier $\mathcal{C}_t$ with $\{(\mathbf{x}_t, y_t)\}$;

10   Update query criteria weights $\mathbf{w}$ (Eqn. (9)) ;

11   **else**

12 $\mathcal{S}_t = \mathcal{S}_{t-1}$;

13   **end**

14 end

15 Unusual event is detected if $p(y = \text{unusual}|\mathbf{x})$ is higher than $\text{Th}_{\text{unusual}}$;

In particular, given a naïve Bayes classifier $\mathcal{C}_t$ and an updated classifier $\mathcal{C}_{t+1}$ trained using $\mathcal{S}_t \bigcup \{(\mathbf{x}_t, y_t)\}$, the KL-divergence between their distributions can be decomposed as follows [23]:

$$\mathcal{KL}(\boldsymbol{\theta} \parallel \tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{D} \mathcal{KL}\left(p_{\boldsymbol{\theta}}(x_i|y) \parallel p_{\tilde{\boldsymbol{\theta}}}(x_i|y)\right)$$

$$= \sum_{i=1}^{D} \sum_{k=1}^{K} p(y = k) \mathcal{KL}\left(p_{\boldsymbol{\theta}}(x_i|y = k) \parallel p_{\tilde{\boldsymbol{\theta}}}(x_i|y = k)\right). \qquad (7)$$

where $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ represent sets of parameters of classifiers $\mathcal{C}_t$ and $\mathcal{C}_{t+1}$ respectively. A symmetric KL-divergence $\overline{\mathcal{KL}}(\boldsymbol{\theta} \parallel \tilde{\boldsymbol{\theta}})$ is computed as follows:

$$\overline{\mathcal{KL}}(\boldsymbol{\theta} \parallel \tilde{\boldsymbol{\theta}}) = \frac{1}{2} \cdot \left[\mathcal{KL}(\boldsymbol{\theta} \parallel \tilde{\boldsymbol{\theta}}) + \mathcal{KL}(\tilde{\boldsymbol{\theta}} \parallel \boldsymbol{\theta})\right]. \qquad (8)$$

A weight $w_{a,t}$ at time step t associated to query Q_a is defined as:

$$w_{a,t} = \beta w_{a,t-1} + (1 - \beta) \frac{\overline{\mathcal{KL}}_a(\boldsymbol{\theta} \parallel \tilde{\boldsymbol{\theta}})}{\sum_{a=1}^{A} \overline{\mathcal{KL}}_a(\boldsymbol{\theta} \parallel \tilde{\boldsymbol{\theta}})}, \qquad (9)$$

where $\overline{\mathcal{KL}}_a(\boldsymbol{\theta} \parallel \tilde{\boldsymbol{\theta}})$ (see Eqn. (8)) represents the symmetric KL-divergence yielded by a query criterion Q_a when it last triggered a query. Parameter β is an update coefficient that controls the updating rate of weights. Algorithm 1 summaries the proposed approach.
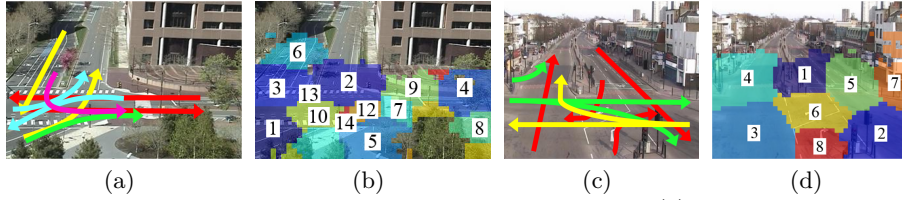
Fig. 2. Dominant traffic flows observed in MIT traffic dataset (a) and Junction dataset (c) are treated as normal event classes. The scene decomposition results of both datasets according to the spatial distribution of activity patterns are shown in (b) and (d), respectively.

4   Experiments

4.1   Datasets and Settings

Two public video datasets[2] captured at busy public scenes are employed in our experiments.

MIT traffic dataset [1] - This dataset with an approximate length of 1.5 hours (168822 frames), is recorded at 30 fps and scaled to a frame size of 360×240. The traffic is controlled with traffic lights and dominated by five different traffic flows (Fig. 2(a)). The scene decomposition result is given in Fig. 2(b), showing the fourteen regions discovered.

Junction dataset - The length of the video is approximately 60 minutes (89999 frames) captured with 360×288 frame size at 25 fps. The traffic is regulated by traffic lights and dominated with three traffic flows as shown in Fig. 2(c). The scene decomposition result is depicted in Fig. 2(d), showing the eight regions discovered.

Both datasets feature complex activities exhibited by multiple objects. In particular, behaviours and the correlations among vehicles are determined by not only the traffic light cycles, but also traffic volume and driving habits of drivers. For instance, vehicles waiting in region 6 of Junction dataset can perform horizontal turning whenever there is a gap in vertical flow. This type of activity is more frequent in MIT dataset, in which vehicles are allowed to do turning between gaps of traffic flows. As a consequence, the traffic phases of MIT traffic dataset are less distinctive visually and become harder to model compared to Junction dataset.

Ground truth - The videos were segmented into non-overlapping clips of 50 frames long each, resulting 1800 clips and 3376 clips for Junction dataset and MIT traffic dataset respectively. Each clip was manually labelled into different event classes as listed in Table 1. The ground truth is used as feedback returned to a classifier when it requests for labels during active learning process[3]. It is also employed for comparison during testing phase.

---

[2] Processed data with ground truth are available for download at: `http://www.eecs.qmul.ac.uk/~ccloy/files/accv_2010_dataset.zip`.

[3] In reality, these labels are assumed to be provided by human operators.

Class	No. of clips (% from total)	Description
		MIT Traffic Dataset
1	874 (25.89)	Horizontal traffic flow (red arrows in Fig. 2(a))
2	1249 (37.00)	Vertical traffic flow (yellow arrows in Fig. 2(a))
3	376 (11.14)	Right-turn from zone 1 toward zone 4 (green arrow in Fig. 2(a))
4	185 (5.48)	Left-turn from zone 3 toward zone 4 (magenta arrow in Fig. 2(a))
5	517 (15.31)	Turning from left-exit toward zone 2, turning from zone 9 to zone 1 (cyan arrows in Fig. 2(a))
6	75 (2.22)	[Unusual] Left-turn from zone 1 to left-exit
7	79 (2.34)	[Unusual] Turning right from zone 7 to zone 2
8	21 (0.62)	[Unusual] U-turn at zone 7
		Junction Dataset
1	1078 (59.89)	Vertical traffic flow (red arrows in Fig. 2(c))
2	323 (17.94)	Rightward traffic flow (yellow arrows in Fig. 2(c))
3	355 (19.72)	Leftward traffic flow (green arrows in Fig. 2(c))
4	29 (1.61)	[Unusual] Illegal u-turns from zone 1 to zone 4 via zone 6
5	3 (0.17)	[Unusual] Emergency vehicles using an improper lane of traffic
6	12 (0.67)	[Unusual] Traffic interruptions by fire engines

Table 1. Ground truth.

Settings - The clips (see Table 1) were randomly partitioned into training/test sets with equal size. Different partitions were used in different runs in the experiments. In this study, all experimental results were averaged over 30 runs.

We followed similar experimental setting reported in [17]. In particular, if a model did not request for any labels after observing a sufficiently large number of samples (100 in this study), the query threshold Th (preset to 0.5) was reduced to Th$'$ where Th$'$ was the largest p^{query} computed since the last query. A budget constraint, i.e. the number of samples a classifier can request on the data stream was specified as 250. There are three free parameters in our active learning approach, namely λ (Eqn. (4)), uncertainty weights γ_u (Eqn. (6)), and update coefficient β (Eqn. (9)). We used coarse values in parameters setting without optimisation: $\lambda = 0.1$ for a weak prior, $\gamma_u = 0.9$ among normal classes, $\gamma_u = 10$ among normal-unusual classes, and $\beta = 0.9$ for slow adaptation rate. The number of committee members for all QBC approaches was set to three. It is reported in [7] that a committee size of three is sufficient and varying the size has little effect. Initially, the classifier was given a sample from a random normal event class to start the learning process. In the QBC approaches, two random samples from different classes were needed.

4.2   Active Learning vs. Unsupervised Learning

We first compared the proposed method with unsupervised learning approach. To build an unsupervised model, a random set of 250 normal samples was selected and normal classes were automatically determined through Gaussian clustering with automatic model selection based on Bayesian Information Criterion score [24]. The samples together with the predicted cluster labels were then employed to train a model described in Sec. 3.2. Note that the unsupervised learning strategy employed here is similar in spirit to that in [1–4]. For fair comparison, we used the same feature representation and model for both active learning and unsupervised learning strategies. As can be seen from Fig. 3(a), our method outperformed the fully unsupervised method given as little as 90 samples annotated

through active learning. Figure 3(b) suggests that the performance of an unsupervised model was still inferior than the proposed approach even if the number of unlabelled samples used for unsupervised learning was increased to 800. It is evident from the results that without exploiting human feedback, unsupervised learning method was unable to learn the correct decision boundary even if a large amount of unlabelled samples were employed.
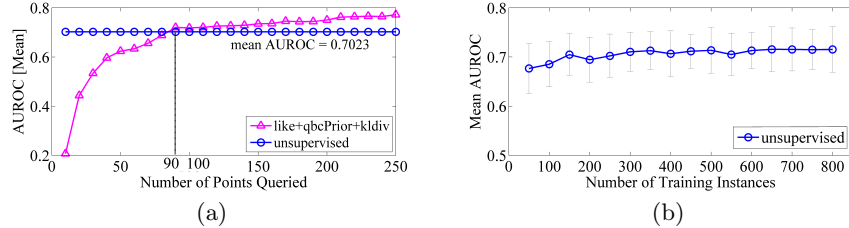

Fig. 3. Comparison with unsupervised approach. Results were averaged over 30 runs.

4.3 Active Learning vs. Passive Supervised Learning and other Active Learning Strategies
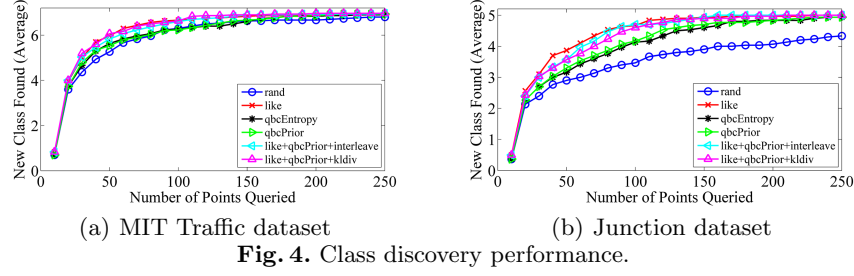
Passive supervised learning (random sampling strategy) and different active learning strategies evaluated in this experiment are summarised as follow:

1. *rand* - passive supervised learning (random sampling strategy), i.e. samples are randomly chosen from the data stream
2. *like* - likelihood criterion as explained in Sec. 3.3
3. *qbcEntropy* - QBC approach with vote entropy measure [6]
4. *qbcPrior*[4] - QBC approach with the proposed measure as described in Sec. 3.3
5. *like+qbcPrior+interleave* - combine both *like* and *qbcPrior* using interleave strategy, i.e. iterating different criteria during learning. This method is similar to the multi-criteria strategy proposed in [14]
6. *like+qbcPrior+KLdiv* - combine both *like* and *qbcPrior* using the KL-divergence-based strategy as described in Sec. 3.4
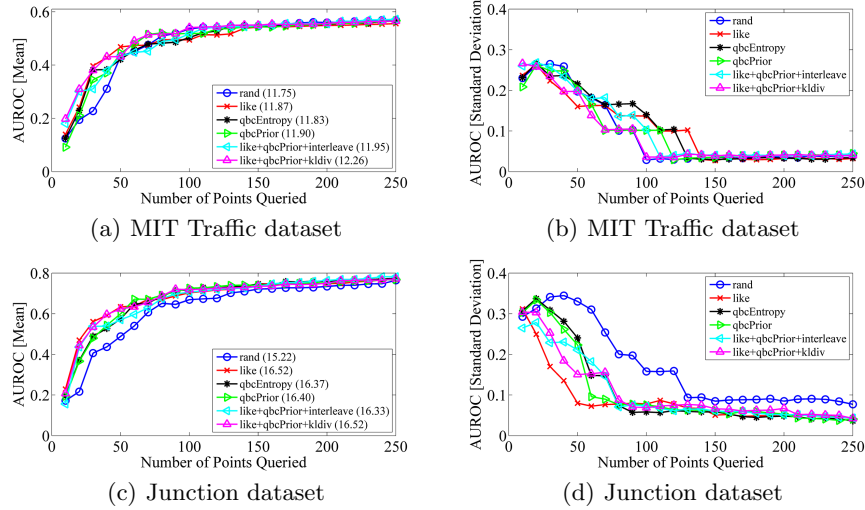
We evaluated different active learning strategies according to: (1) how fast they can discover unknown classes (including normal and unusual event classes) and (2) how accurately the learned classifier can detect unusual events. The former case was measured based on the number of classes discovered vs. number of samples queried. The latter case was evaluated using Area Under Receiver Operating Characteristic curve (AUROC) computed in each active learning iteration against the number of queried samples. ROC was obtained by varying $Th_{unusual}$.

Discover unknown event classes - As can be seen from Figure 4, *like* showed the best performance in discovering unknown event classes in both datasets. The QBC approaches (*qbcEntropy* and *qbcPrior*) yielded slightly inferior result compared to *like* but still performed better than random sampling strategy.

---

[4] Matlab codes are available for download at: `http://www.eecs.qmul.ac.uk/~ccloy/files/qbcPrior.zip`.

(a) MIT Traffic dataset   (b) Junction dataset

Fig. 4. Class discovery performance.

Specifically, with the introduction of prior (see Eqn. (6)) for dealing with the imbalanced data problem, the performance of our *qbcPrior* is better to that of *qbcEntropy* (see Fig. 4(b)). Both the proposed *like+qbcPrior+KLdiv* method and the alternative multi-criteria method *like+qbcPrior+interleave* showed comparable results and performed better than *qbcPrior* and *qbcEntropy* after combining with *like*.

(a) MIT Traffic dataset   (b) MIT Traffic dataset

(c) Junction dataset   (d) Junction dataset

Fig. 5. Unusual events detection performance. Numbers shown in the brackets within graph legend are area under the mean AUROC of different approaches.

Unusual event detection - Figure 5 shows the performance of different active learning strategies in detecting unusual events, measured as averaged AUROC over 30 runs. Overall, it can be seen that the detection performance of all active learning methods monotonically increase as more data is queried and importantly, all methods significantly outperformed random sampling (passive supervised learning). In particular, we observed that by incorporating prior constraint into uncertainty criterion (*qbcPrior*) yielded slightly better result compared to method without prior constraint (*qbcEntropy*). This is because that, without the prior constraint, *qbcEntropy* wasted some effort in refining boundary between normal classes, whilst *qbcEntropy* focussed on the uncertainty regions surrounding the unusual event classes.

In both datasets, as can be seen from Fig. 5(a) and 5(c), the proposed method *like+qbcPrior+KLdiv* yields the best performance. Adaptive selection of multiple criteria leads to reliable and good performance with our *like+qbcPrior+KLdiv* outperforming the alternative *like+qbcPrior+interleave*. The reliability of the multi-criteria methods is also reflected by the smaller variance across multiple trials shown in Fig. 5(b) and 5(d).
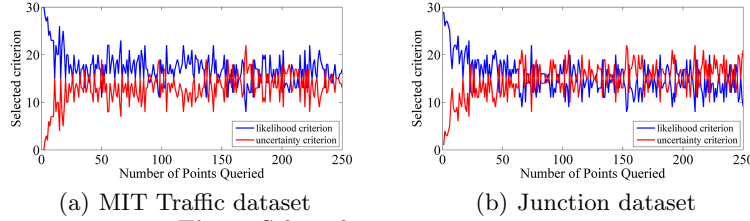

(a) MIT Traffic dataset   (b) Junction dataset

Fig. 6. Selected criterion over 30 runs.

Adaptive selection of criteria - In contrast to iterative strategy reported in [14], the proposed strategy assigns weights adaptively to different criterion at different stage of active learning based on the KL-divergence of a model (see Sec.3.4). For the MIT traffic dataset (Fig. 6(a)), likelihood criterion was more frequently selected than uncertainty criterion before the number of queried samples reached 150. This observation suggests that when the visual distinctiveness between event classes were less obvious (see Sec. 4.1), our method was capable of selecting the right criterion and avoid uncertainty criterion that may keep querying uncertain points located at highly overlapped area of class boundary, which are less useful for improving the detection performance. On the other hand, in the Junction dataset (Fig. 6(b)), likelihood criterion dominated at the beginning, since it discovered unknown events that caused greater change in parameter values to the classifier compared to uncertainty criterion. The model eventually switched from likelihood criterion to uncertainty criterion (after 80 samples were queried) to refine the classification boundary when the exploratory learning was no longer fruitful (see Fig. 4(b), approximately 90% of total event classes were discovered after 80 samples).

5   Conclusion

To our best knowledge, this study is the first investigation into the use of active learning to exploit human feedback for on-line unusual event detection. Importantly, the proposed approach yielded more robust and accurate detection on subtle unusual events in public space as compared to conventional supervised and unsupervised learning strategies, by exploiting a small cost of human supervision through active learning. Experimental results also showed that the proposed stream-based multi-criteria approach is capable of balancing different query criteria for joint unknown event discovery and decision boundary refinement. It therefore results in a more reliable detection performance than using single criterion alone, and it outperforms an existing multi-criteria strategy [14] applied in a stream-based manner. In addition, by introducing a prior to deal with imbalanced data, our re-formulated QBC criterion improves the performance.

References

1. Wang, X., Ma, X., Grimson, W.E.L.: Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. TPAMI **31** (2009) 539–555
2. Hospedales, T., Gong, S., Xiang, T.: A Markov clustering topic model for mining behaviour in video. In: ICCV. (2009)
3. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behaviour detection using social force model. In: CVPR. (2009) 935–942
4. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: CVPR. (2009) 2921–2928
5. Settles, B.: Active learning literature survey. Technical report, University of Wisconsin Madison (2010)
6. Argamon-Engelson, S., Dagan, I.: Committee-based sample selection for probabilistic classifiers. JAIR **11** (1999) 335–360
7. McCallum, A.K., Nigam, K.: Employing EM in pool-based active learning for text classification. In: ICML. (1998) 350–358
8. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: COLT. (1992) 287–294
9. Cover, T., Thomas, J.: Information Theory. Wiley (1991)
10. Remagnino, P., Jones, G.: Classifying surveillance events from attributes and behaviour. In: BMVC. (2001)
11. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: ICCV. (2009)
12. Sillito, R., Fisher, R.: Semi-supervised learning for anomalous trajectory detection. In: BMVC. (2008)
13. Pelleg, D., Moore, A.: Active learning for anomaly and rare-category detection. In: NIPS. (2004)
14. Stokes, J.W., Platt, J.C., Kravis, J., Shilman, M.: ALADIN: Active learning of anomalies to detect intrusions. Technical report, Microsoft Research (2008)
15. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. IJCV (2009)
16. Baram, Y., El-Yaniv, R., Luz, K.: Online choice of active learning algorithms. JMLR **5** (2004) 255–291.
17. Ho, S.S., Wechsler, H.: Query by transduction. TPAMI **30** (2008) 1557–1571
18. Cebron, N., Berthold, M.R.: Active learning for object classification: from exploration to exploitation. Data Min. Knowl. Disc. **18** (2008) 283–299
19. Dagan, I., Engelson, S.: Committee-based sampling for training probabilistic classifiers. In: ICML. (1995) 150–157
20. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Imaging Understanding Workshop. (1981) 121–130
21. Loy, C.C., Xiang, T., Gong, S.: Multi-camera activity correlation analysis. In: CVPR. (2009) 1988–1995
22. Devroye, L.: Non-Uniform Random Variate Generation. Springer-Verlag (1986)
23. Tong, S., Koller, D.: Active learning for parameter estimation in Bayesian networks. In: NIPS. (2000) 647–653
24. Xiang, T., Gong, S.: Video behaviour profiling for anomaly detection. TPAMI **30** (2008) 893–908