International Journal of Neural Systems

MULTI-INSTANCE DICTIONARY LEARNING FOR DETECTING ABNORMAL EVENTS IN SURVEILLANCE VIDEOS

Jing Huo, Yang Gao*, Wanqi Yang
State Key Laboratory for Novel Software Technology, Nanjing University, China
Hujun Yin
School of Electrical and Electronic Engineering, The University of Manchester, UK

In this paper, a novel method termed Multi-Instance Dictionary Learning (MIDL) is presented for detecting abnormal events in crowded video scenes. With respect to multi-instance learning, each event (video clip) in videos is modeled as a bag containing several sub-events (local observations); while each sub-event is regarded as an instance. The MIDL jointly learns a dictionary for sparse representations of sub-events (instances) and multi-instance classifiers for classifying events into normal or abnormal. We further adopt three different multi-instance models, yielding the Max-Pooling based MIDL (MP-MIDL), Instance based MIDL (Inst-MIDL) and Bag based MIDL (Bag-MIDL), for detecting both global and local abnormalities. The MP-MIDL classifies observed events by using bag features extracted via max-pooling over sparse representations. The Inst-MIDL and Bag-MIDL classify observed events by the predicted values of corresponding instances. The proposed MIDL is evaluated and compared with the state-of-the-art methods for abnormal event detection on the UMN (for global abnormalities) and the UCSD (for local abnormalities) datasets and results show that the proposed MP-MIDL and Bag-MIDL achieve either comparable or improved detection performances. The proposed MIDL method is also compared with other multi-instance learning methods on the task and superior results are obtained by the MP-MIDL scheme.

Keywords: Multi-instance learning, Sparse coding, Dictionary learning, Abnormal event detection

1. Introduction

Detecting abnormal events in crowded video scenes is an important and challenging task in computer vision. Automatically detecting anomalies in surveillance videos, which are accumulating rapidly in the digital era, can facilitate efficient search and screening. The task otherwise may prove too costly or even impossible by manual operations.

In the literature, the definition of abnormal events is often qualitative and subjective under different application scenarios. However, following the definitions in Refs. 1 and 3, abnormal events generally possess the following characters. One character is that the events seldom occur or have not been observed before. The other is that the events are unpredictable. Detection of abnormal events is challenging due to the fact that anomalies in videos often occur with a very low probability and also have dramatic appearance variations. Thus, the problem of abnormal event detection is to identify anomalies, given a large number of normal events and possibly a small portion of available abnormal events. In this case, it becomes an unbalanced learning problem. Much previous work has used one-class unsupervised methods to tackle the problem.[1,4] Recently, Yang[2] et al. proposed a framework based on trajectory segmentation and multi-instance learning to detect local anomalies. However, trajectories of objects are often hard to obtain in crowded video scenes. We herein propose to use motion-based abnormal event detection under the framework of multi-instance learning.

Motion based abnormal events in video can be classified into two categories: local and global.[4] Local abnormal events are local behaviors with different motion patterns compared with its spatial-temporal neighbors on the scene (e.g. vehicles on crowded sidewalks). Global abnormal events are scenes where all

* Corresponding author, e-mail: gaoy@nju.edu.cn

the local behaviors are abnormal (e.g. crowded escape events). The task thus is to identify frames containing either local or global abnormal events. In order to better depict local anomalies that appear in local regions of video frames, a given short video clip of several frames is first divided into small spatial-temporal cuboids. Motion features are then extracted in these cuboids. We define a video clip as an event and local motion patterns as sub-events. An event is abnormal if at least one of its sub-events is abnormal, that is, one local region contains an abnormal sub-event. An event is normal only if all its sub-events are normal. This can be naturally framed under the multi-instant learning methodology. In multi-instance learning, a bag is defined as a collection of several instances. A bag is labeled positive if at least one of the instances is positive; or it is labeled negative if all of its instances are negative. Therefore if we define an event as a bag and abnormal sub-events as positive instances, then a positive bag corresponds to an abnormal event, while a negative bag corresponds to a normal event. Then abnormal event detection is to perform multi-instance classification to find positive bags, which correspond to abnormal events. The frames in a clip corresponding to a positive bag are identified as abnormal.

In order to effectively detect abnormal events in videos, learning a good representation of events plays an essential role. Sparse coding has been used as an effective feature representation method in the literature.[5,6,7] This is because when compared with other methods such as principal component analysis, sparse coding does not impose orthogonality constraints on basis vectors, thus leading to more flexible representations. Several previous abnormal event detection methods[3,4,8] adopt the sparse coding technique as feature representation for individual events and have shown superior performance. Sparse coding is also employed in the proposed method due to its effectiveness in representing events. Previous sparse coding based methods, however, learn the feature representation for each individual event separately in an unsupervised manner, leading to the learned sparse representation being good for reconstruction but inefficient for multi-instance classification. For multi-instance classification, it is better to learn a dictionary that is able to produce sparse codes more effectively. To achieve this, we have developed a novel dictionary learning method called multi-instance dictionary learning (MIDL). The MIDL jointly learns the dictionary as well as solves the multi-instance learning problem by minimizing a classification loss function. The dictionary is learned for sparse coding of instances and the classification model for classifying bags. By using different classification models, three different schemes of the MIDL are naturally produced:

(1) Max-Pooling based MIDL (MP-MIDL): It classifies bags by using bag features extracted via max-pooling over the sparse codes of instances.

(2) Instance based MIDL (Inst-MIDL): The label of a bag is determined by the maximal classification value of all instances in the bag. The learning of a dictionary uses all the instances.

(3) Bag based MIDL (Bag-MIDL): The label of a bag is also determined by the maximal classification value of all instances in the bag. The learning of a dictionary uses selected instances in bags. In each bag, an instance with the maximal classification value is selected.

Experimental results show that MP-MIDL is suited for global abnormal event detection and Bag-MIDL for local abnormal event detection. Inst-MIDL also shows a comparable result compared with MP-MIDL and Bag-MIDL while it outperforms some existing abnormal event detection methods.

This work is a significant extension of our earlier work,[9] in which only a prototype Bag-MIDL was experimented with. Here, the MIDL is considered as a general formulation and a further three different schemes of the MIDL are derived. Moreover, the relationships and differences among these three schemes are analyzed in detail.

The contributions of the work can be summarized as follows:

(i) Abnormal event detection is modeled as a multi-instance learning task for effective detection of abnormal events in crowded scenes.

(ii) A novel dictionary learning method, i.e. MIDL, is proposed for learning dictionaries and sparse representations of sub-events that are adapted to the problem of multi-instance learning.

(iii) Three schemes of the MIDL are developed depending on the classification model used, together with an analysis of the three schemes and comparisons with other methods.

The rest of the paper is organized as follows. Section 2 gives a brief overview of related work.

Section 3 provides a detailed explanation of the proposed method, followed by experimental results, comparisons and analysis in Section 4. Section 5 concludes the paper.

2. Related Work

Abnormal event detection is an active topic in the area of video processing.[10,11,12,41,43,44] The related methods can be categorized into two categories: trajectory based and motion feature based.

The trajectory based methods rely mainly on tracking of an object.[2,13,14,42] However, reliable tracking[15,16] is still a challenging task. Besides, in many crowded scenes, tracking of an object can be unrealistic due to occlusions. This results in the trajectory based methods being unsuited for crowded scenes.

There have been a lot of efforts devoted to motion feature based methods. In these methods, features such as optical flow, motion history, gradients are extracted at pixel level. Then different models are built to learn the spatial-temporal relations between different feature patterns. These models include the Markov Random Field,[17] Gaussian Mixture Model[18,19] and Social Force Model.[20] Such methods avoid explicitly tracking moving objects and therefore are more suited for detecting abnormal events in crowded scenes.

Recent studies show that sparse coding is effective for the task of abnormal event detection.[3,4,8] Previous work follows almost the same idea: using the sparse reconstruction cost to classify events. For example, in the method proposed by Zhao et al,[3] motion features of events are calculated first. Then with a learned dictionary, sparse codes of these motion features are computed. The final step is to evaluate the value of the sparse reconstruction cost. If the reconstruction cost is greater than a user defined threshold, the event is considered as an abnormal event. The threshold is used to control the sensitivity of the detection. The methods proposed by Cong et al[4] and Xu et al[8] use similar ideas. Generally, a reconstruction cost is adopted in these methods. Unlike these previous sparse-coding based methods, the proposed method is able to learn discriminative sparse codes of instances for effective multi-instance classification.

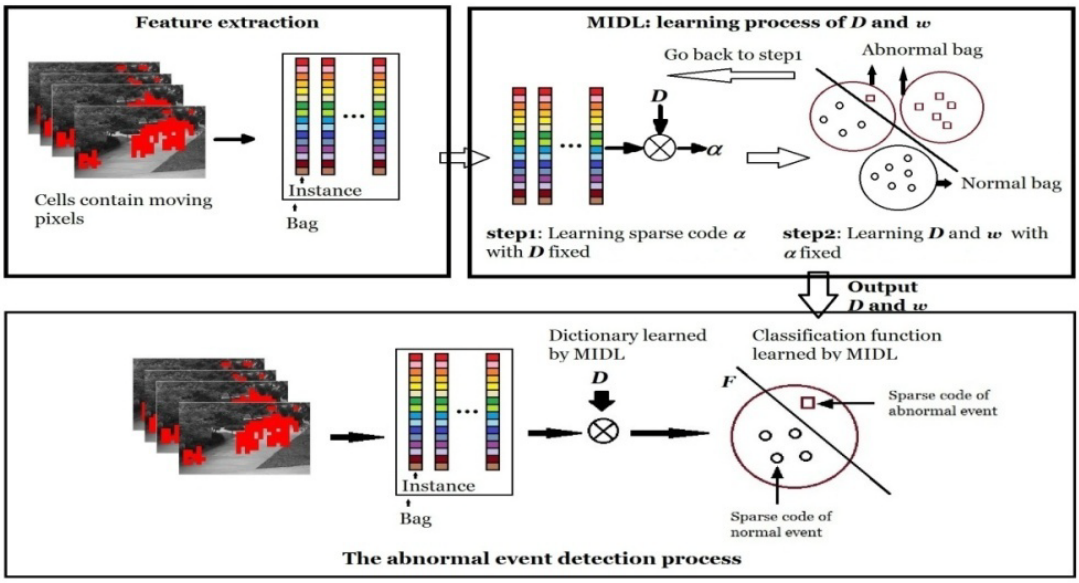

Fig. 1. Flowchart of the proposed method. The first step is to extract features of events. Corresponding details are described in Section 3.2. In the MIDL learning process, dictionary $\boldsymbol{D}$ and a classification parameter $\boldsymbol{w}$ are learned together, as detailed in Sections 3.3 and 3.4. In the last process (abnormal event detection), features are extracted, and then with the learned dictionary, sparse codes of instance are computed, finally a new event is classified as normal or abnormal by using the obtained classification function.

3. Multi-instance Dictionary Learning for Abnormal Event Detection

3.1. *The framework*

Fig. 1. shows the flowchart of the proposed method. There are three steps in the proposed abnormal event detection method. First, a given video is divided into a set of clips, each containing a fixed number of frames. The number of frames in a clip is decided via cross validation. A video clip is regarded as a bag (an event). Then video clips are partitioned into spatial-temporal cuboids.[†] Frame differencing is used to calculate pixels of moving objects (or moving pixels). For the cuboids that contain moving pixels, motion based features are extracted. The motion feature extraction procedure will be described in Section 3.2. The motion features within a bag are considered as instances (sub-events).

The second step is a dictionary learning procedure. Here we use multi-instance dictionary learning to jointly learn a dictionary with a multi-instance classification function. Three different multi-instance dictionary learning schemes are proposed. Details of the proposed multi-instance dictionary learning method are given in Sections 3.3 and 3.4.

In the last step, given a video clip to classify, it is first represented as a bag and instances (local motion features) using the feature extraction method. With the learned dictionary, sparse codes of instances are first computed. Using the sparse codes of instances as feature vectors and together with the learned multi-instance classification function, the bag is classified as positive or negative, corresponding to an abnormal or a normal event. Further details are presented in Section 3.5.

3.2. *Event representation*

For normal and abnormal events in surveillance videos, the major differences between them are their motion direction and motion magnitude. Since Multi-scale Histogram of Optical Flow (MHOF)[4] can capture well both motion direction and motion magnitude, it is adopted in our framework to depict events. Other feature representation methods designed for surveillance videos are also applicable. The MHOF feature has 16 bins. The first 8 bins denote the 8 directions of the optical flow with its magnitude smaller than a threshold τ. If the magnitude of the optical flow is greater than the threshold, it is quantized into 8 directions in the next 8 bins. The threshold τ is selected based on a cross validation test in the experiments.

The entire feature extraction procedure used in this paper has the following four steps: (1) Given a video, we first partition it into small clips. Each clip contains a fixed number of frames. In our case, every four frames create a clip, as shown in Fig. 2. This frame number was selected based on a cross validation test. By varying the number of frames in a clip and calculating the corresponding final prediction precision via cross validation, the number of frames in a clip with the highest prediction precision was then selected. (2) Moving pixels are then detected using the frame differencing method. Partition the video view into small overlapping cells, the cells between several successive frames form small spatial-temporal cuboids. And the moving pixels will fall into different cuboids. In Fig. 2, cells containing moving pixels are marked in red. The selection of the cell size is fairly flexible. Generally, smaller cells can capture smaller moving abnormal objects and thus give higher detection precision, but will lead to more instances in a bag and possible computational inefficiency. For a reasonable trade-off between the two aspects, in this work we have used the fixed settings with cells of size 24×24 and cuboids of size $4\times24\times24$. Other choices are also possible. (3) Optical flow[21] is then computed between these frames. (4) For those cuboids that contain moving pixels, optical flow within the cuboids is used to extract the MHOF features.

For the frames in a succession, the MHOF features are considered as instances (i.e. sub-events). All the features jointly form the concept of a bag (i.e. event) in the multi-instance learning framework. The feature extraction process is summarized in Fig. 2.

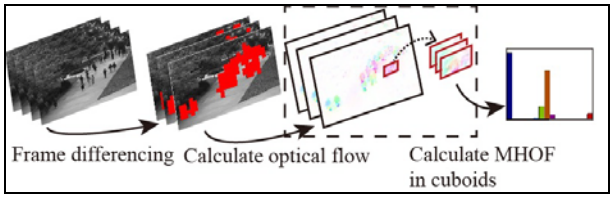

Fig. 2. Flowchart of the feature extraction process.

[†] The frame number in a clip and the size of a cuboid are discussed in Section 3.2.

3.3. *General formulation of Multi-instance Dictionary Learning*

3.3.1 Sparse coding

The basic formulation of sparse coding is two-fold: an input sample is modeled as a linear combination of the basis in a dictionary, and the coefficients are sparse. This is the so called sparse representation. Here, a learned dictionary given is $\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_k) \in \mathbf{R}^{m \times k}$. The dictionary can be overcomplete with the number of its basis vectors greater than the dimension of the sample, $k > m$. An input sample is represented by $\boldsymbol{x} \in \mathbf{R}^m$, and the sparse representation of $\boldsymbol{x}$ can be represented as $\hat{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{D}) \in \mathbf{R}^k$ ($\hat{\boldsymbol{\alpha}}$ is used for short in the rest of the paper).

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{D}) = \text{argmin}_{\boldsymbol{\alpha} \in \mathbf{R}^k} (\tfrac{1}{2} || \boldsymbol{x} - \boldsymbol{D}\boldsymbol{\alpha} ||_2^2 + \lambda_1 || \boldsymbol{\alpha} ||_1) \quad (1)$$

where the first term is the sparse reconstruction error; the second term is the sparsity regularization term and λ_1 is a regularization parameter. The l_1 norm used in the second term guarantees that there are only a few non-zero entries in $\hat{\boldsymbol{\alpha}}$.

With $\boldsymbol{D}$ fixed, the above optimization task is an l_1-regularized least-square problem. Solutions of this problem include: the Interior Point,[22] a modification of the Least Angle Regression (LARS),[7,23]Feature Sign Search,[5] etc.

3.3.2 Multi-instance Dictionary Learning

In the above formulation, $\boldsymbol{D}$ is assumed as given or fixed. However, in practice, $\boldsymbol{D}$ is learned from a set of training samples. A classical approach to obtain $\boldsymbol{D}$ is by minimizing the reconstruction error.

$$\text{argmin}_{\boldsymbol{\alpha}, \boldsymbol{D}} \tfrac{1}{n} \sum_{i=1}^{n} (\tfrac{1}{2} || \boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{\alpha}_i ||_2^2 + \lambda_1 || \boldsymbol{\alpha}_i ||_1) \quad (2)$$

where n is the total number of training samples.

In Eq. (2), usually there is a constraint on the column of $\boldsymbol{D}$, such that, $|| \boldsymbol{d}_j ||_2^2 \le c, \quad \forall j \in \{1,2,...,k\}$. This is to avoid the elements of $\boldsymbol{D}$ being arbitrarily large.

Eq. (2) aims to learn a dictionary that is best suited for signal reconstruction tasks.[6,7] As has been pointed out in Refs. 24, 25, 26 and 27, for classification tasks, it is not optimal to learn a dictionary in this way since the label information of samples is not used. It would be better to learn dictionaries while considering the labels of samples, so generating sparse codes that are discriminating with respect to the labels. Like in supervised dictionary learning, there is some prior information of the relationship between labels and bags that should be considered and used for multi-instance dictionary learning. Thus, we present here a multi-instance dictionary learning method specifically designed for multi-instance classification.

Suppose a set of training bags is given with their labels as: $B = \{B_1, B_2, ..., B_n\}$ and $Y = \{Y_1, Y_2, ..., Y_n\}$, where n is the number of training bags, $B_i = \{\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, ..., \boldsymbol{x}_{n_i}^{(i)}\}$ is the ith bag containing n_i instances, $\boldsymbol{x}_j^{(i)} \in \mathbf{R}^m$ is the jth instance in bag B_i, and $Y_i \in \{+1, -1\}$ is the label of the ith bag. We use $A = \{A_1, A_2, ..., A_n\}$ to represent a set of sparse codes, where $A_i = \{\hat{\boldsymbol{\alpha}}_1^{(i)}, \boldsymbol{\alpha}_2^{(i)}, ..., \boldsymbol{\alpha}_{n_i}^{(i)}\}$ is the set of sparse codes of instances in bag B_i, with $\hat{\boldsymbol{\alpha}}_j^{(i)} \in \mathbf{R}^k$ being the sparse code of instance $\boldsymbol{x}_j^{(i)}$.

For the task of multi-instance classification, the primary goal is to consider a multi-instance classification function to classify bags as positive or negative. The sparse codes of instances are treated as feature vectors and then the classification is carried out with respect to these sparse codes. By minimizing the total classification loss on the training set, the formulation becomes:

$$\min_{\boldsymbol{D}, \boldsymbol{w}} (f(\boldsymbol{D}, \boldsymbol{w}) + \tfrac{\nu}{2} || \boldsymbol{w} ||_2^2) \quad (3)$$

where $\boldsymbol{D} \in \mathbf{R}^{m \times k}$ is the dictionary and $\boldsymbol{w} \in \mathbf{R}^k$ is the classification parameter. It aims to jointly learn both $\boldsymbol{D}$ and $\boldsymbol{w}$. ν is a regularization parameter for avoiding over fitting. The function f is defined as the total classification loss of bags and is represented as,

$$f(\boldsymbol{D}, \boldsymbol{w}) = \tfrac{1}{n} \sum_{i=1}^{n} C(Y_i, B_i, \boldsymbol{w}) \quad (4)$$

where C is a function measuring the classification loss of a bag. Thus, f represents the total classification loss of classifying bags.

In the next subsection, we discuss the multi-instance classification functions. For the multi-instance classification problem, finding good representations of instances and bags is imperative. The multi-instance dictionary learning method plays an important role in the proposed abnormal event detection framework and the resulting three schemes.

3.4. *Three schemes of Multi-instance Dictionary Learning*

3.4.1 Multi-instance Dictionary Learning using bag features extracted by max-pooling (MP-MIDL)

Yang et al. proposed a supervised dictionary learning method[26] for image classification. In their work, an image is represented using features extracted by max-pooling over the sparse codes of local descriptors within a spatial pyramid. The dictionary is trained for local descriptors through the back-propagation. Similar to this method, we also apply max-pooling over the sparse codes of instances to extract features of bags.

The max-pooling procedure is shown as,

$$\boldsymbol{\beta}_i = \phi(B_i) = \begin{bmatrix} \max\{\hat{\boldsymbol{\alpha}}_{11}^{(i)}, \boldsymbol{\alpha}_{21}^{(i)}, ..., \boldsymbol{\alpha}_{n_i 1}^{(i)}\} \\ \max\{\hat{\boldsymbol{\alpha}}_{12}^{(i)}, \boldsymbol{\alpha}_{22}^{(i)}, ..., \boldsymbol{\alpha}_{n_i 2}^{(i)}\} \\ ... \\ \max\{\hat{\boldsymbol{\alpha}}_{1k}^{(i)}, \boldsymbol{\alpha}_{2k}^{(i)}, ..., \boldsymbol{\alpha}_{n_i k}^{(i)}\} \end{bmatrix} \quad (5)$$

where ϕ represents the max-pooling operator.

The resulting feature vector for bag B_i is $\boldsymbol{\beta}_i \in \mathbf{R}^k$, where its jth element is $\boldsymbol{\beta}_{ij} = \max\{\hat{\boldsymbol{\alpha}}_{1j}^{(i)}, \boldsymbol{\alpha}_{2j}^{(i)}, ..., \boldsymbol{\alpha}_{n_i j}^{(i)}\}$, and the operator $\max\{\hat{\boldsymbol{\alpha}}_{1j}^{(i)}, \boldsymbol{\alpha}_{2j}^{(i)}, ..., \boldsymbol{\alpha}_{n_i j}^{(i)}\}$ means taking the maximum with respect to the elements at the jth dimension of every sparse codes of instances in bag B_i.

For multi-instance classification tasks, we are interested in predicting bags. Linear classification function is employed to make predictions, described as

$$F(B_i, A_i, \boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\beta}_i \quad (6)$$

where $\boldsymbol{\beta}_i$ is used as feature vector and $\boldsymbol{w} \in \mathbf{R}^k$ is the parameter of the linear classification model.

Logistic loss,[38,39] $L(x) = log(1+e^{-x})$, is chosen to measure the classification loss of a bag, as it is both convex and differentiable. So the complete formulation of the classification loss of a bag, C, is shown as,

$$C(Y_i, B_i, \boldsymbol{w}) = L(Y_i F(B_i, A_i, \boldsymbol{w})) = log(1+e^{-Y_i F(B_i, A_i, \boldsymbol{w})}) \quad (7)$$

And the final objective function of the Max-Pooling based MIDL (MP-MIDL) is given by,

$$\min_{\boldsymbol{D}, \boldsymbol{w}} \left(\frac{1}{n} \sum_{i=1}^{n} log(1+e^{-Y_i F(B_i, A_i, \boldsymbol{w})}) + \frac{\nu}{2} ||\boldsymbol{w}||_2^2\right) \quad (8)$$

As can be seen from Eq. (8), dictionary $\boldsymbol{D}$ and the classification parameter $\boldsymbol{w}$ are jointly optimized to minimize the sum of the classification loss of bags on the training set.

Once $\boldsymbol{D}$ and $\boldsymbol{w}$ are learned, the label of bag B_i is given by the sign of the value $(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\beta}_i)$. Let $\hat{Y}_i$ denote the predicted label of B_i, and then $\hat{Y}_i$ is represented by the following formulation,

$$\hat{Y}_i = \mathrm{sign}(F(B_i, A_i, \boldsymbol{w})) = \begin{cases} -1, \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\beta}_i \le 0 \\ 1, \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\beta}_i > 0 \end{cases} \quad (9)$$

In the above equation, $\hat{Y}_i$=1 means B_i is predicted as a positive bag and $\hat{Y}_i$=-1 means B_i is predicted as negative.

3.4.2 Multi-instance Dictionary Learning at instance level (Inst-MIDL)

In this scheme, the classification model is learned with respect to instances. It is clear that every instance (local motion pattern) has its own associated label though not directly accessible. If the instance corresponds to an abnormal sub-event, the instance is positive, or negative otherwise. A bag is labeled positive if there is at least one positive instance in the bag, or is labeled negative otherwise. So once we are able to classify instances, the label of a bag is also determined.

The classification of an instance is performed with respect to its sparse code using a linear classification model represented as,

$$l(\hat{\boldsymbol{\alpha}}_j^{(i)}, \boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\alpha}_j^{(i)} \quad (10)$$

where $\boldsymbol{w} \in \mathbf{R}^k$ is the classification parameter and $\hat{\boldsymbol{\alpha}}_j^{(i)}$ is the sparse code of instance $\boldsymbol{x}_j^{(i)}$.

Sparse codes of instances are used as feature vectors. The predicted label $\hat{y}_j^{(i)}$ of instance $\boldsymbol{x}_j^{(i)}$ can be represented using,

$$\hat{y}_j^{(i)} = \mathrm{sign}(l(\hat{\boldsymbol{\alpha}}_j^{(i)}, \boldsymbol{w})) = \mathrm{sign}(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\alpha}_j^{(i)}) \quad (11)$$

If $\hat{y}_j^{(i)}$=1, $\boldsymbol{x}_j^{(i)}$ is predicted as a positive instance and $\hat{y}_j^{(i)}$=-1 means $\boldsymbol{x}_j^{(i)}$ is labeled as negative.

Suppose the true label of an instance is known, then the classification loss of a bag can be defined as,

$$C(Y_i, B_i, \boldsymbol{w}) = \sum_{j=1}^{n_i} log(1+e^{-y_j^{(i)} l(\hat{\boldsymbol{\alpha}}_j^{(i)}, \boldsymbol{w})}) \quad (12)$$

where $y_j^{(i)}$ is the true label of instance $\boldsymbol{x}_j^{(i)}$ and n_i is the number of instances in bag B_i.

As can be seen, the classification loss of bag B_i is the sum of all the classification loss of instances within this bag.

However, in Eq. (12), $y_j^{(i)}$ is unknown in positive bags in practice. Motivated by the scheme of the mi-SVM,[28] we use a heuristic scheme to infer the labels of instances. To solve for the final objective function, two steps are used at each iteration:

1) For negative bags, all the instance labels are set as -1.

2) For positive bags, firstly, the label of instances are set using Eq. (11). Then, if there is no positive instance in the positive bag, the label of the instance with the maximal ($\boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_j^{(i)}$) value is set as +1, where $\boldsymbol{w}$ is the current learned classification parameter.

The final objective function of this scheme is given as,

$$\min_{\boldsymbol{D},\boldsymbol{w}} (\frac{1}{s}\sum_{i=1}^{n}\sum_{j=1}^{n_i} log(1+e^{-y_j^{(i)} l(\hat{\boldsymbol{\alpha}}_j^{(i)},\boldsymbol{w})}) + \frac{\nu}{2}||\boldsymbol{w}||_2^2) \tag{13}$$

where $s=\sum_{i=1}^{n} n_i$ is the total number of instances in all the training bags.

Since the classification loss of a bag is defined as the sum of all the classification loss of instances within this bag, Eq. (13) can be interpreted as jointly optimizing $\boldsymbol{D}$ and $\boldsymbol{w}$ in order to minimize the total classification loss of all instances in all the training bags.

Given the learned parameters $\boldsymbol{D}$ and $\boldsymbol{w}$, the classification function for instances is determined. Then the classification for a bag is carried out by predicting the labels of all the instances in the bag, if there is at least one positive instances in the bag, the bag is classified as positive. It means, for positive bags, the maximum classification value of instances is positive, while the maximum classification value of instances in a negative bag is negative. That is, the classification value of a bag is represented as,

$$F(B_i, A_i, \boldsymbol{w}) = \max_{j=1,...,n_i} (\boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_j^{(i)}) \tag{14}$$

And the predicted label of a bag is given by,

$$\hat{Y}_i = \mathrm{sign}(F(B_i, A_i, \boldsymbol{w})) = \mathrm{sign}(\max_{j=1,...,n_i} (\boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_j^{(i)})) \tag{15}$$

3.4.3 Multi-instance Dictionary Learning at bag level (Bag-MIDL)

The above instance based learning relies on a heuristic scheme to get the labels of instances. In many cases, however, what we know is the labels of bags. If $Y_i = -1$, then $y_{ij} = -1$, for all $j = 1,...,n_i$. On the other hand, if $Y_i = 1$, then there is at least one instance in the bag that is positive. Therefore, as stated in the previous subsection, the classification value of a bag is represented by Eq. (14). The Inst-MIDL tries to learn a linear classifier for all the instances. Now using Eq. (14), the target is changed to learning a classifier for a set of selected instances. In each bag, an instance is selected to learn the classifier, the instance with the maximum classification value in a bag. For negative bags, the selected instance is the one nearest to the separation plane, which is the most uncertain or most discriminative instance in the bag. While for positive bags, the selected instance is the one farthest from the separation plane, which is the most certain or least discriminative instance in the bag. The difference between this scheme (Bag-MIDL) and the Inst-MIDL scheme is illustrated in Fig. 3.

The classification loss of a bag and the final objective function of Bag-MIDL are presented in the following two formulae, respectively.

$$C(Y_i, B_i, \boldsymbol{w}) = log(1+e^{-Y_i \max_{j=1,...,n_i} (\boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_j^{(i)})}) \tag{16}$$

$$\min_{\boldsymbol{D},\boldsymbol{w}} (\frac{1}{n}\sum_{i=1}^{n} log(1+e^{-Y_i \max_{j=1,...,n_i} (\boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_j^{(i)})}) + \frac{\nu}{2}||\boldsymbol{w}||_2^2) \tag{17}$$

Once $\boldsymbol{D}$ and $\boldsymbol{w}$ are learned, the label of a bag is given by Eq. (15).

3.4.4 Optimization method

In the previous subsections, three different schemes of MIDL are presented. The three objective functions can be solved using the same optimization method. We optimize alternatively between sparse code $\hat{\boldsymbol{\alpha}}$, dictionary $\boldsymbol{D}$ and classification parameter $\boldsymbol{w}$. Sparse code $\hat{\boldsymbol{\alpha}}$ are optimized with dictionary $\boldsymbol{D}$ fixed. Then with $\hat{\boldsymbol{\alpha}}$ fixed, $\boldsymbol{D}$ and $\boldsymbol{w}$ are optimized.

Learning sparse code $\hat{\boldsymbol{\alpha}}$ with $\boldsymbol{D}$ fixed. This is the coding phase. With $\boldsymbol{D}$ fixed, the optimization of $\hat{\boldsymbol{\alpha}}$ in Eq. (1) is solved using the Euclidean projection based method by projecting the coefficient vector onto the l_1 ball.[40,30,29] Other sparse coding methods can also be used, such as a modified LARS[7] and the Feature Sign Search.[5]

Learning dictionary $\boldsymbol{D}$ and $\boldsymbol{w}$ with $\hat{\boldsymbol{\alpha}}$ fixed. This is the dictionary learning phase, together learning a classifier. Gradient and subgradient methods are used for optimization. We adopt the method proposed in Ref. 26. A brief description of the optimization method of $\boldsymbol{D}$ and $\boldsymbol{w}$ is given below, taking the objective function of Eq. (17) as an example (the other two schemes can also be solved using the same scheme).

The optimization of $\boldsymbol{w}$ in Eq. (17) is straightforward. Since $\boldsymbol{w}$ is explicit in Eq. (17). Gradient of $\boldsymbol{w}$ can be calculated directly. If the gradient is not available, we use subgradient instead. The optimization of $\boldsymbol{D}$ in Eq.

(17) is not as easy as the optimization of $\boldsymbol{w}$, since $\boldsymbol{D}$ is not explicit in Eq. (17). Therefore implicit differentiation and chain rule are used to compute the gradient of $\boldsymbol{D}$ in Eq. (17). Compute the gradient of $\boldsymbol{D}$ using chain rule:

$$\frac{\partial f}{\partial \boldsymbol{D}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial C}{\partial \boldsymbol{D}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial C}{\partial F}\frac{\partial F}{\partial \boldsymbol{D}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial C}{\partial F}\frac{\partial F}{\partial \hat{\boldsymbol{\alpha}}}\frac{\partial \hat{\boldsymbol{\alpha}}}{\partial \boldsymbol{D}} \qquad (18)$$

The problem turns into computing $\frac{\partial \hat{\boldsymbol{\alpha}}}{\partial \boldsymbol{D}}$, since $\boldsymbol{D}$ and $\hat{\boldsymbol{\alpha}}$ are not explicitly linked. However, this can be solved by taking the derivation of $\boldsymbol{D}$ with respect to a fixed point equation. The way of calculating a fixed point equation is described in Ref. 26. Finally, we obtain the partial derivative of $\boldsymbol{D}$ in the following form,

$$\frac{\partial \tilde{\boldsymbol{\alpha}}}{\partial \boldsymbol{D}_{mn}} = (\tilde{\boldsymbol{D}}^T\tilde{\boldsymbol{D}})^{-1}\left(\frac{\partial \tilde{\boldsymbol{D}}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{D}_{mn}} - \frac{\partial \tilde{\boldsymbol{D}}^T\tilde{\boldsymbol{D}}}{\partial \boldsymbol{D}_{mn}}\tilde{\boldsymbol{\alpha}}\right) \qquad (19)$$

where $\tilde{\boldsymbol{\alpha}}$ means the nonzero elements in $\hat{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{D}}$ the corresponding bases, $\boldsymbol{D}_{mn}$ is the element of dictionary $\boldsymbol{D}$ at the mth row and the nth column.

Implementation details: 1) Stochastic gradient descent is used in the implementation for efficient training. A sample is drawn from the training set at each iteration. 2) The learning rate of $\boldsymbol{w}$ is updated in the form of $\lambda_{\boldsymbol{wt}} = \min(\lambda_{\boldsymbol{w}}, \lambda_{\boldsymbol{w}} t_0 / t)$, where $\lambda_{\boldsymbol{w}}$ is a constant. And if the current iteration number t satisfies $t \le t_0$, $\lambda_{\boldsymbol{wt}} = \lambda_{\boldsymbol{w}}$, or $\lambda_{\boldsymbol{wt}}$ decreases if $t > t_0$. t_0 is therefore a threshold, set to $t_0 = T/10$, where T is the total number of iterations. $\lambda_{\boldsymbol{Dt}}$ is set as a constant in the experiments, that is, $\lambda_{\boldsymbol{Dt}}$ is always equal to $\lambda_{\boldsymbol{D}}$. Based on a cross validation test, the best parameters $\lambda_{\boldsymbol{w}}$ and $\lambda_{\boldsymbol{D}}$ are selected. For ν, it is set as a constant. 3) For $\boldsymbol{D}$ and $\boldsymbol{w}$, $\boldsymbol{D}$ is initialized using unsupervised dictionary learning, and $\boldsymbol{w}$ is initialized as a vector with its all elements set as one. The algorithm is summarized in Algorithm 1.

In Algorithm 1, $f(\boldsymbol{D},\boldsymbol{w})$ in Eq. (20) is either the first term in Eq. (8), Eq. (13) or Eq. (17) depending on which scheme is adopted. $f(\boldsymbol{D},\boldsymbol{w})$ in Eq. (21) has the same meaning. Now we discuss the training speed of the three schemes using the stochastic gradient descent. If one bag is chosen at each iteration, then $f(\boldsymbol{D},\boldsymbol{w})$ reduces to the classification loss of one bag equal to C as defined in Eqs. (7), (12) and (16). From Eqs. (7), (12) and (16), if the time complexity of the MP-MIDL and Bag-MIDL is $O(1)$ in each iteration, then it is $O(n_i)$ for the Inst-MIDL, where n_i is the number of instances in the selected bag. This is because the Inst-MIDL defines the classification loss of a bag as the sum of all the classification loss of instances in Eq. (12). We have observed the same result from experiments that the optimization speed of the Inst-MIDL is slower than the other two schemes. The convergence of the algorithm can be improved by randomly selecting a set of bags at each iteration instead of one bag at a time.

Algorithm 1: MIDL

Input: Initialize $\boldsymbol{D}$ and $\boldsymbol{w}$. Training set B, A, Y
Output: $\boldsymbol{D}$ and $\boldsymbol{w}$
Step 1:
For $t = 1:T$
Randomly select one or several bags in the training set.
Update $\lambda_{\boldsymbol{Dt}}$ and $\lambda_{\boldsymbol{wt}}$.

1. Compute $\hat{\boldsymbol{\alpha}}$ in Eq. (1).
2. Optimize $\boldsymbol{D}$ and $\boldsymbol{w}$ using gradient descent.
   2.1 Optimize $\boldsymbol{D}$
   $$\boldsymbol{D}_t = \boldsymbol{D}_{t-1} - \lambda_{\boldsymbol{Dt}}(\nabla_{\boldsymbol{D}} f(\boldsymbol{D},\boldsymbol{w})) \qquad (20)$$
   2.2 Optimize $\boldsymbol{w}$
   $$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \lambda_{\boldsymbol{wt}}(\nabla_{\boldsymbol{w}} f(\boldsymbol{D},\boldsymbol{w}) + \nu\boldsymbol{w}_{t-1}) \qquad (21)$$

end
Step2:
Output $\boldsymbol{D}$ and $\boldsymbol{w}$

3.5. *Abnormal event detection*

Section 3.4 describes and discusses the three schemes for learning multi-instance dictionary and the corresponding classification parameters. As stated before, the detection of abnormal events in the proposed framework is to perform multi-instance classification in order to find positive bags corresponding to abnormal events. So once the dictionary $\boldsymbol{D}$ and the classification parameter $\boldsymbol{w}$ are learned, the classification of an unlabeled bag B_i (or the detection of an abnormal event) follows the following procedure.

(1) With the learned dictionary $\boldsymbol{D}$, solve Eq. (1) to learn the sparse representations of the instances in B_i.

(2) With the learned sparse codes of instances and also the classification parameter: if MP-MIDL is adopted, use Eq. (9) to classify bag B_i; if Inst-MIDL or Bag-MIDL is adopted, Eq. (15) is used for classification of B_i. After classification, we obtained the predicted label $\hat{Y}_i$ of bag B_i.

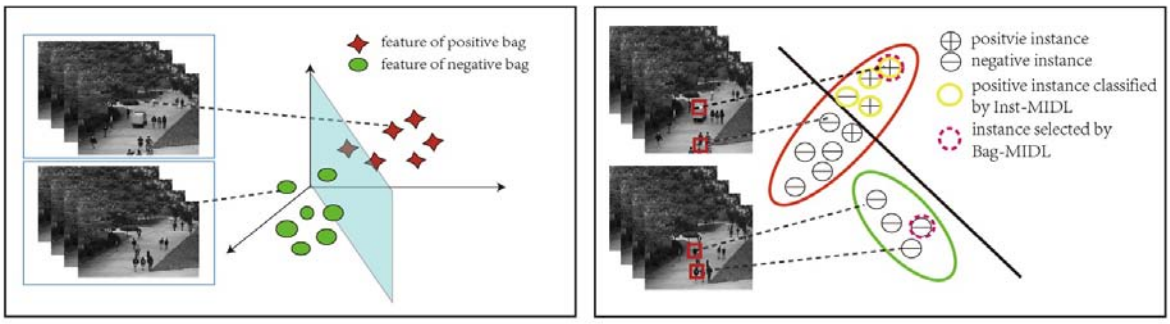

Fig. 3. Illustration of the differences between three schemes of MIDL. Left figure shows in MP-MIDL, a bag is projected into a new bag feature space; the classification function is learned directly for classification of bags. Right figure shows the differences between Inst-MIDL and Bag-MIDL, the main one being the training instances selected by the two methods. The classifier is learned in the feature space of instances.

If $\hat{Y}_i$ is equal to -1, bag B_i is negative and corresponds to a normal event. While $\hat{Y}_i$ is +1, then bag B_i is positive and it relates to an abnormal event. When a positive bag is detected, the frames in the corresponding video clip of this positive bag are all labeled as abnormal. It deals with the frame level abnormal event detection problem and is a problem we are mostly interested in.

We now describe how to detect pixel level abnormal events using the Inst-MIDL and Bag-MIDL. Eq. (11) can be used to predict the label of sub-events. If an instance is classified as positive, it corresponds to an abnormal sub-event, so the cuboid corresponding to this instance contains abnormal sub-event. The region of this cuboid in the frame is marked as an abnormal region, as shown in Section 4.2.

3.6. *Comparison of the three schemes*

A comparison of the three methods is illustrated in Fig. 3. The interpretation of the max-pooling based method is learning a bag based feature space. Projecting bags into the new feature space, a linear classifier is able to separate those bags. The dictionary learning procedure is to find a best feature space that helps classification. While the Inst-MIDL and Bag-MIDL learn to classify bags in the instance feature space, the Inst-MIDL uses all the instances to learn a classifier. Though the labels of instances are unknown in positive bags, we can use a heuristic scheme to infer the labels of these instances. Contrary to the Inst-MIDL, Bag-MIDL selects one instance in every bag for training, the one with the maximal classification value.

The three schemes are suited where multi-instance learning can be adopted. For the abnormal event detection task, the MP-MIDL mainly deals with frame level abnormal event detection. Whilst the Inst-MIDL and Bag-MIDL learn a classifier in the instances feature space. These two schemes can predict the labels of instances (sub-events). Therefore, not only frames of abnormal events are detected, but also the cuboids containing abnormal sub-event, thus the locations of these abnormal sub-events.

The performances and differences of these three methods will be further discussed in Section 4. A conclusion is that the MP-MIDL is relatively suited for global abnormal event detection, while the Bag-MIDL for local abnormal event detection. The Bag-MIDL is able to select discriminated normal instances, with such discriminative information added into dictionary learning, it achieves better performance.

4. Experiments and Results

To evaluate and verify the effectiveness of the proposed method, two public available datasets: the UMN dataset[31] and the UCSD dataset[32] were used. The UMN dataset is commonly used for global abnormal event detection, while the UCSD for local abnormal event detection. Fig. 4 shows examples of normal and abnormal events in the two datasets.

Fig. 4 Examples of events in UMN scene 1 (top) and UCSD Peds 1 (bottom). Top left image is a frame of normal walking events and top right image is a frame of abnormal crowd escape event (global abnormal event). Bottom left shows a frame of normal pedestrian walking event and bottom right a frame of abnormal vehicles on crowded sidewalk event (local abnormal event).

4.1. *Detection of global abnormal events*

The UMN dataset consists of three different scenes of crowded escape events with people walking as normal event and people running to escape as abnormal event. And the total number of frames of the video is 7740 (1450, 4415 and 2145 for scenes 1, 2, and 3, respectively). The original resolution of the UMN dataset is 320×240. For the MHOF feature extraction procedure, there are two parameters to adjust. One is the motion magnitude threshold and the other is the number of frames in a clip. The two parameters are set by two-fold cross validation. The motion magnitude threshold is first adjusted by fixing the number of frames in a clip and then in turn the number of frames is adjusted by fixing the threshold. The MIDL models are trained separately on three scenes. The parameters of the MIDL models are adjusted using grid search and two-fold cross validation.

4.1.1 Comparison of frame level abnormal event detection of the three schemes of MIDL

We first compare the results of frame level abnormal event detection of the three schemes of MIDL. Area under the ROC curve (AUC) is calculated as the criterion. The results are summarized in Table 1. The MP-MIDL performs the best on all three scenes compared to other two schemes, though the differences are small. The AUC values of the Inst-MIDL and Bag-MIDL are slightly lower than that of the MP-MIDL on Scene 2. On Scene 3, the Bag-MIDL has a marginally lower AUC value compared with the MP-MIDL and Inst-MIDL. But the differences are very small. This shows that the proposed MIDL method is capable for the task of frame level abnormal events and all the three schemes perform well. From the result, the MP-MIDL performs best and this demonstrates the proposed bag feature extraction method works well in practice.

Table 1. AUC of frame level abnormal event detection of the three schemes on three scenes of UMN dataset.

	Scene 1	Scene 2	Scene 3
MP-MIDL	0.99	0.98	0.99
Inst-MIDL	0.99	0.96	0.99
Bag-MIDL	0.99	0.97	0.98

4.1.2 Comparison of MIDL with other abnormal event detection methods

Table 2 lists the AUC results of other abnormal event detection methods[4,20,33] on the UMN dataset. From Tables 1 and 2, it can be seen that the proposed method outperforms the methods of Social Force[20] and Optical Flow[20] and is comparable with the Chaotic Invariants[33] and SRC[4]. This demonstrates strength of the proposed method for global abnormal event detection.

Table 2. AUC results of other abnormal event detection methods on UMN dataset

Method	Area under ROC
Social Force[20]	0.96
Optical Flow[20]	0.84
Chaotic Invariants[33]	0.99
SRC (Scene 1)[4]	0.99
SRC (Scene 2)[4]	0.97
SRC (Scene 3)[4]	0.96

4.2. *Detection of local abnormal events*

The UCSD datasets contains two scenes of pedestrian walking on a sidewalk. The UCSD Peds1 contains 34 clips of videos for training, and 36 clips for testing with resolution of 158×238, and the UCSD Peds2 contains 16 clips for training and 12 clips for testing with resolution of 360×240. The training clips only contain normal events. As this is a scene of sidewalk, normal events of this dataset are pedestrian walking. And examples of abnormal events include buses, wheelchairs,

bicycles, and skaters which seldom appear in the scene. These abnormal events only exist in the testing data.

For both sets, we only used the testing clips. The testing clips were partitioned for training and testing because the proposed method requires both normal and abnormal events in the training stage. The parameters were adjusted and two-fold cross validation was used.

4.2.1 Comparison of MIDL with other abnormal event detection methods

The equal error rates (EER) is computed as the criterion. The EER is where the false accept rate equates the false reject rate. For a good classification algorithm, the EER should be as low as possible. A comparison of EER for frame level detection is shown in Table 3. On both Peds1 and Peds2, the lowest EER values are highlighted in boldface.

Table 3. Equal Error Rates (EER) for frame level abnormal event detection on UCSD Peds1 and UCSD Peds2

	Peds1	Peds2
MP-MIDL	31%	25%
Inst-MIDL	32 %	24%
Bag-MIDL	**27%**	**8%**
SF[34]	31%	42%
MPPCA[17]	40%	30%
SF-MPPCA[34]	32%	36%
MDT[34]	**25%**	25%
Adam[35]	38%	42%
SRC[4]	**19%**	/
Sparse[8]	33 %	**9%**
Sparse-CS[8]	31%	**6%**

As can be seen, the Bag-MIDL performs the best among the three schemes of the proposed method. On Peds1, though SRC[4] is better than Bag-MIDL, Bag-MIDL achieves comparable result compared with MDT[34] and outperforms the rest of the methods. The result of MP-MIDL and Inst-MIDL is also comparable with SF[34], SF-MPPCA[34], Sparse[8] and Sparse-CS[8] and is better than MPPCA[17] and Adam[35]. On Peds2, the EER value of Bag-MDIL, Sparse[8] and Sparse-CS[8] are close. In particular, these three methods are significantly superior to the rest of the methods. While the results of MP-MIDL and Inst-MIDL are close to the results of MDT, they are much better than that of SF[34], MPPCA[17], SF-MPPCA[34] and Adam[35]. This shows that for the detection of local abnormal events, the proposed method also performs fairly well. The reason for the good result is that local abnormal events are modeled as instances and thus are not neglected in the framework of multi-instance learning.

4.2.2 Results on pixel level abnormal event detection

We now discuss the results of pixel level abnormal event detection. One thing needs to be mentioned is that only the Bag-MIDL and Inst-MIDL are able to detect abnormal events at pixel level. The MP-MIDL is only suited for frame level abnormal event detection. The pixel level detection has been mentioned in Section 3.5. Some pixel level detection results of the Bag-MIDL are shown in Fig. 5, in which cells containing sub-events that are classified as abnormal are marked. As can be seen, though no information about the label of sub-events was used, the method was able to learn the concept of normal and abnormal automatically with the help of the labels of training bags. Detection of abnormal events at pixel level can give more detailed information about abnormal events, i.e. not only in which frames but also the locations of the abnormal events. The good results achieved at pixel level may also illustrate why the proposed method achieves good result at the frame level detection.

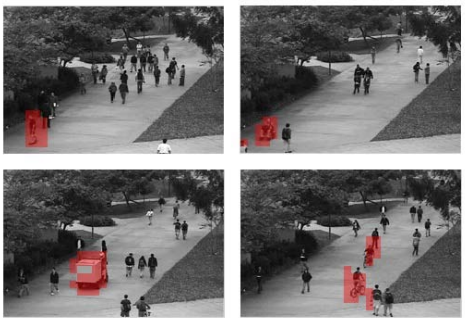

Fig. 5 Examples of pixel level abnormal event detection on UCSD Peds1 using Bag-MIDL. Abnormal events are well detected (skater, wheelchair, vehicle and bicycle). The local regions are classified using the function defined by Eq. (11).

4.3. *Comparison with other multi-instance learning methods*

For a fair comparison, several other multi-instance methods have been applied to the same datasets for abnormal event detection. They include the mi-SVM,[28] MI-SVM,[28] EM-DD,[36] and Citation-KNN.[37] The

methods were tested on the UMN datasets. On all the three scenes, 200 bags were used for both training and testing after feature extraction. Only a part of the datasets was used because methods such as Citation-KNN are very time consuming when running on the entire dataset. For all the methods, we ran ten times and the average precision was taken. Each time, two-fold cross validation was used. Results are shown in Table 4. On Scene 1, MP-MIDL, Inst-MIDL, EM-DD and Citation-KNN achieved the best results. On Scene2, the best result was achieved by MP-MIDL. On Scene 3, all the methods performed fairly well with Inst-MIDL, mi-SVM and MI-SVM having a slightly lower prediction precision. From the table, it is apparent that the proposed method is either better or comparable with the current state-of-the-art multi-instance learning method for this abnormal event detection task.

Table 4. Precision of various multi-instance learning methods on UMN dataset.

Method	Scene1	Scene2	Scene3
MP-MIDL	0.99	0.95	0.99
Inst-MIDL	0.99	0.85	0.95
Bag-MIDL	0.96	0.87	0.99
mi-SVM[28]	0.87	0.79	0.93
MI-SVM[28]	0.88	0.79	0.94
EM-DD[36]	0.99	0.84	0.99
Citation-KNN[37]	0.98	0.87	0.98

5. Conclusions

A method termed Multi-instance Dictionary Learning (MIDL) is proposed for automatic detecting abnormal events in videos. A dictionary learning procedure is carried out together with the learning of a multi-instance classifier. By adopting different multi-instance learning models, the proposed method yields three schemes. MP-MIDL is suited for frame level abnormal event detection; while Inst-MIDL and Bag-MIDL for both frame level and pixel level abnormal event detection. Various experiments have been conducted to verify effectiveness of the method. The results show that, compared with the state-of-the-art abnormal event detection techniques, the proposed method demonstrates its strength and compared with the current multi-instance learning methods, the proposed method is either superior or comparable. Specifically, among the three schemes, MP-MIDL is the most suited for global abnormal event detection and Bag-MIDL performs best for local abnormal event detection.

The future work will include exploring how to automatically set the parameters of the MHOF feature extraction procedure, so that optimal parameters can be learned automatically for different usage scenarios. Faster optimization methods will also be explored such as the Alternative Least Squares for the multi-instance dictionary learning task and further parameter selection tests to verify the robustness of the method. We also plan to apply the method to other multi-instance learning tasks, such as image classification and to abnormal event detection scenarios other than surveillance videos.

Acknowledgement

The authors would like to acknowledge the support for this work from the National Science Foundation of China (Grant Nos. 61035003, 61175042, 61321491), the Graduate Research Innovation Program of Jiangsu, China (CXZZ13_0055), the 973 Program of Jiangsu, China (Grant No. BK2011005) and Program for New Century Excellent Talents in University (Grant No. NCET-10-0476). The authors would also like to thank anonymous reviewers for their insightful comments and many helpful suggestions.

References

1. C. C. Loy, *Activity Understanding and Unusual Event Detection in Surveillance Videos*, PhD Thesis, Queen Mary University of London (2010).
2. W. Yang, Y. Gao and L. Cao, TRASMIL: A local anomaly detection framework based on trajectory segmentation and multi-instance learning, in *Computer Vision and Image Understanding* (2012) doi: 10.1016/j.cviu. 2012.08.010.
3. B. Zhao, F.-F. Li and E. P. Xing, Online detection of unusual events in videos via dynamic sparse coding, in *IEEE Conference on Computer Vision and Pattern Recognition* (2011), pp. 3313-3320.
4. Y. Cong, J. Yuan and J. Liu, Sparse reconstruction cost for abnormal event detection, in *IEEE Conference on Computer Vision and Pattern Recognition* (2011), pp. 3449-3456.
5. H. Lee, A. Battle, R. Raina and A. Y. Ng, Efficient sparse coding algorithms, *Advances in Neural Information Processing Systems* , eds. B. Schölkopf, J. Platt and T. Hoffman **19** (2007), pp. 801-808.
6. M. Elad and M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, *IEEE Transactions on Image Processing* **15**(12) (2006) 3736-3745.

7. J. Mairal, F. Bach, J. Ponce and G. Sapiro, Online dictionary learning for sparse coding, in *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), pp. 689-696.
8. J. Xu, S. Denman, S. Sridharan, C. Fookes and R. Rana, Dynamic texture reconstruction from sparse codes for unusual event detection in crowded scenes, in *Proceedings of the 2011 Joint ACM Workshop on Modeling and Representing Events* (2011), pp. 25-30.
9. J. Huo, Y. Gao, W. Yang and H. Yin, Abnormal event detection via multi-instance dictionary learning, in *Proceedings of Intelligent Data Engineering and Automated Learning* (2012), pp. 76-83.
10. K. Subramanian and S. Suresh, Human action recognition using meta-cognitive neuro-fuzzy inference system, *International Journal of Neural Systems* **22**(06) (2012) doi: 10.1142/S0129065712500281.
11. J. A. Iglesias, P. Angelov, A. Ledezma and A. Sanchis, Human activity recognition based on evolving fuzzy systems, *International Journal of Neural Systems* **20**(05) (2010) 355-364.
12. E. López-Rubio, R. M. Luque-Baena and E. Domínguez, Foreground detection in video sequences with probabilistic self-organizing maps, *International Journal of Neural Systems* **21**(3) (2011) 225-246.
13. I. Saleemi, K. Shafique and M. Shah, Probabilistic modeling of scene dynamics for applications in visual surveillance, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(8) (2009) 1472-1485.
14. C. Piciarelli and G. L. Foresti, On-line trajectory clustering for anomalous events detection, *Pattern Recognition Letters* **27**(15) (2006) 1835-1842.
15. L. Quesada and A. J. León, Unsupervised markerless 3-DOF motion tracking in real time using a single low-budget camera, *International Journal of Neural Systems* **22**(5) (2012) doi: 10.1142/S0129065712500190.
16. K. A. Kramer and S. C. Stubberud, Analysis and implementation of a neural extended Kalman filter for target tracking, *International Journal of Neural Systems*, **16**(1) (2006) 1-13.
17. J. Kim and K. Grauman, Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates, in *IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 2921-2928.
18. Y. Shi, Y. Gao and R. Wang, Real-time abnormal event detection in complicated scenes, in *20th IEEE International Conference on Pattern Recognition* (2010), pp. 3653-3656.
19. L. Kratz and K. Nishino, Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models, in *IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 1446-1453.
20. R. Mehran, A. Oyama and M. Shah, Abnormal crowd behavior detection using social force model, in *IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 935-942.
21. C. Liu, W. T. Freeman, E. H. Adelson and Y. Weiss, Human-assisted motion annotation, in *IEEE Conference on Computer Vision and Pattern Recognition* (2008), pp. 1-8.
22. S. S. Chen, D. L. Donoho and M. A. Saunders, Atomic decomposition by basis pursuit, *SIAM Journal on Scientific Computing* **20**(1) (1999) 33-61.
23. B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least angle regression, *The Annals of Statistics* **32**(2) (2004) 407-499.
24. J. Mairal, F. Bach and J. Ponce, Task-driven dictionary learning, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(4) (2012) 791-804.
25. J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman, Supervised dictionary learning, *Advances in Neural Information Processing Systems*, eds. D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, **21** (2008), pp. 1033-1040.
26. J. Yang, K. Yu and T. Huang, Supervised translation-invariant sparse coding, in *IEEE Conference on Computer Vision and Pattern Recognition* (2010), pp. 3517-3524.
27. J. A. Bagnell and D. M. Bradley, Differential sparse coding, *Advances in Neural Information Processing Systems*, eds. D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, **21** (2008), pp. 113-120.
28. S. Andrews, I. Tsochantaridis and T. Hofmann, Support vector machines for multiple-instance learning, *Advances in Neural Information Processing Systems* **15** (2002), pp. 561-568.
29. J. Liu, J. Chen and J. Ye, Large-scale sparse logistic regression, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2009), pp. 547-556.
30. J. Liu, S. Ji and J. Ye, SLEP: Sparse learning with efficient projections, (Arizona State University, 2009), url: http://www.public.asu.edu/~jye02/Software/SLEP.
31. UMN: Unusual crowd activity dataset of University of Minnesota, available from: http://mha.cs.umn.edu/movies/crowdactivity-all.avi
32. V. Mahadevan, W. Li, V. Bhalodia and N. Vasconcelos, Anomaly detection in crowded scenes, in *IEEE Conference on Computer Vision and Pattern Recognition* (2010), pp. 1975-1981.
33. S. Wu, B. E. Moore and M. Shah, Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes, in *IEEE Conference on Computer Vision and Pattern Recognition* (2010), pp. 2054-2060.
34. V. Mahadevan, W. Li, V. Bhalodia and N. Vasconcelos, Anomaly detection in crowded scenes, in *IEEE Conference on Computer Vision and Pattern Recognition* (2010), pp. 1975-1981.
35. A. Adam, E. Rivlin, I. Shimshoni and D. Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(3) (2008) 555-560.
36. Q. Zhang and S. A. Goldman, EM-DD: an improved multi-instance learning technique, *Advances in Neural Information Processing Systems* **14** (2001), pp. 1073-1080.
37. J. Wang and J.-D. Zucker, Solving multiple-instance

problem: A lazy learning approach, in *Proceedings of the Seventeenth International Conference on Machine Learning* (2000), pp. 1119-1126.
38. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, (Springer New York, 2001).
39. T. Zhang and F. J. Oles, Text categorization based on regularized linear classiffication methods, *Information Retrieval* **4**(1) (2001) 5-31.
40. J. Liu and J. Ye, Efficient Euclidean projections in linear time, in *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), pp. 657-664.
41. F. Verbist, N. Deligiannis, M. Jacobs, J. Barbarien, P. Schelkens and A. Munteanu, Maximum Likelihood Motion Compensation for Distributed Video Coding, *Integrated Computer-Aided Engineering* **19**(3) (2012) 215-227.
42. P. M. Ciarelli, E. O. T. Salles, and E. Oliveira, Human Automatic detection and tracking for outdoor video, *Integrated Computer-Aided Engineering*, **18**(4) (2011) 379-390.
43. Y. Tsai and Y. Huang, A Generalized Framework for Parallelizing Traffic Sign Inventory of Video Log Images Using Multi-Core Processors, *Computer-Aided Civil and Infrastructure Engineering*, **27**(7) (2012) 476-493.
44. Z. Hu, Intelligent Road Sign Inventory (IRSI) with Image Recognition and Attribute Computation from Video Log, *Computer-Aided Civil and Infrastructure Engineering*, **28**(2) (2013) 130-145.