Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2016) 24: 2638 – 2653

doi:10.3906/elk-1310-191

# Abnormal event detection in crowded scenes via bag-of-atomic-events-based topic model

**Xing HU[1], Shiqiang HU[1,*], Lingkun LUO[1], Guoxiang LI[2]**
[1]School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai, P.R. China
[2]Academic Affairs Division, Guangxi University of Finance and Economic, Guangxi, P.R. China

**Received:** 24.10.2013 • **Accepted/Published Online:** 26.10.2014 • **Final Version:** 15.04.2016

**Abstract:** In this paper, we propose a novel framework for abnormal event detection in crowded scenes. A new concept of atomic event is introduced into this framework, which is the basic component of video events. Different from previous bag-of-words (BoW) modeling-based methods that represent feature descriptors using only one code word, a feature descriptor is represented using a few more atomic events in bag-of-atomic-events (BoAE) modeling. Consequently, the approximation error is reduced by using the obtained BoAE representation. In the context of abnormal event detection, BoAE representation is more suitable to describe abnormal events than BoW representation, because the abnormal event may not correspond to any code word in BoW modeling. Fast latent Dirichlet allocation is adopted to learn a model of normal events, as well as classify the testing event with low likelihood under the learned model. Our proposed framework is robust, computationally efficient, and highly accurate. We validate these advantages by conducting extensive experiments on several challenging datasets. Qualitative and quantitative results show the promising performance compared with other state-of-the-art methods.

**Key words:** Bag-of-atomic-events, abnormal event detection, fast latent Dirichlet allocation

## 1. Introduction

Abnormal event detection in crowded scenes is a challenging problem in computer vision, which has attracted a large amount of research interest [1–26]. In daily life, crowded scenes can be found in many places, such as subway stations, amusement parks, shopping malls, and so on. There are numerous pedestrians or moving objects with various dynamics in crowded scenes. If an abnormal event happens suddenly in such scenes, such as people running or scattering, it might lead to stampedes as happened at the Loveparade 2010 in Duisburg, Germany [1]. Therefore, finding an abnormal event in time may help to avoid tragedy.

Most existing abnormal event detection methods fall into two categories. The first category is tracking-based methods, such as in [2–6]. In these methods, pedestrians or moving objects are first detected and tracked by tracking algorithm, and then the system learns a normalcy model using the obtained trajectories. The trajectories that are not represented by the learnt model are detected as abnormal events. However, tracking algorithms tend to fail in crowded scenes due to large numbers of individuals and frequent occlusions. The second category is motion feature-based methods, such as in [11–16, 21, 23–26]. These methods adopt motion features, such as social force, dynamic texture, and so on, to describe events. For example, in [11], crowd behavior was characterized by a social force model, which reflects the interaction force within a crowd; in [12],

*Correspondence: sqhu@sjtu.edu.cn

local optical flow patterns were modeled by a mixture of probabilistic principal component analyzers models to obtain the prototype of crowd events; in [13], appearance and dynamics of crowded scenes were jointly modeled using mixtures of dynamic textures (MDT); in [15, 16], the statistics of spatiotemporal interest points were exploited to characterize crowd behaviors. However, these methods are either computationally expensive or not robust in crowded scenes.

The problem of abnormal event detection refers to finding patterns in video that do not conform to expected patterns [27]. The major difficulties of abnormal event detection in crowded scenes are as follows:

- How to choose an appropriate feature descriptor significantly affects the performance of the detection system [17].
- It is impossible to capture either normal or abnormal events by a single distribution, because both normality and anomaly are diverse.
- It is difficult to collect enough abnormal samples for training, since abnormal events are rare and occur infrequently in the real world.
- Video data are typically high-dimensional and redundant, which requires an efficient modeling tool and a compact representation while yielding a stable and highly accurate classification rate [28].
- The computational cost and memory requirements of the detection system should be low because of the need for real-time detection in surveillance situations.

LDA is a generative probabilistic model introduced in [29], which is an excellent tool for finding underlying structures and capturing statistical properties from a collection of conditionally independent and identically distributed random variables [30]. LDA was initially applied in text modeling, where the number of documents is sometimes huge. Fortunately, LDA has discriminative power that can deal with large-scale databases. However, LDA only supports discrete input, i.e. no type of data is processed by LDA except for discrete data. Hence, most continuous feature descriptors are modeled as BoW representations. BoW-based topic models have been applied in video analysis and description for face recognition [30], action recognition [31], and so on. In BoW modeling, a code word is created from the feature set by clustering like K-means or vector quantization, and the feature descriptor is assigned to one and only one code word (using, e.g., the nearest neighbor in $R^d$). This leads to considerable amount of approximation error, and the number of code words has to be increased as the data exhibit more and more variations. In the context of abnormal event detection, crowd event is more appropriate to be represented by more code words than a single code word, because the feature descriptor corresponding to an abnormal event unseen in the past may not correspond well to any code word in the codebook.

Dictionary learning is an effective tool for modeling high-dimensional signals, such as audio, images, and videos [32]. Given a learned overcomplete dictionary, high-dimensional data can efficiently and accurately approximate by linear combination of dictionary atoms and corresponding sparse representations. Unlikely traditional models like GMM where fitting such data requires a large number of training samples, a dictionary can be efficiently learned from a limited number of training samples. Dictionary learning can be regarded as a generalization of the clustering or vector-quantization-based codebook learning process in BoW modeling [33], where dictionary atoms can be considered as code words. Sparse representation can reconstruct data with low approximation error, since it allow more atoms to participate. In contrast, BoW representation only uses one code word, so it will lead to a considerable amount of approximation error.

In this paper, we propose a BoAE-based topic model framework for abnormal event detection in crowded scenes. Inspired by the efficiency of sparse representation, we represent events using a "collection of atomic events" by novel BoAE modeling. We consider the dictionary atoms as the atomic events that are the basic components of an event. In BoAE modeling, events are well approximated by the linear combination of atomic events and their corresponding BoAE representations. Even if the event is unseen in the training set, it can be accurately represented by BoAE representation. In BoAE representation, each element indicates the frequency of the corresponding atomic event in the event. Given a sequence, we first extract feature descriptors, and then we obtain their corresponding BoAE representations by BoAE modeling. Figure 1 illustrates the process of BoAE modeling. The underlying structure in BoAE representations is found by FLDA [34]. Thus, a video sequence is modeled as a distribution over events, and each event is modeled as a distribution over atomic events. Finally, the testing BoAE representation with low likelihood under the trained LDA is labeled as abnormal.

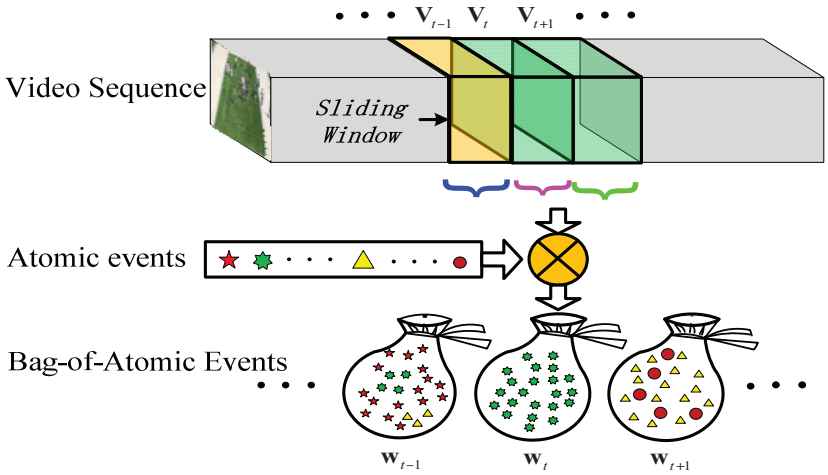


**Figure 1.** Illustration of the process of BoAE modeling.

The major contributions of this paper are summarized as follows:

- We propose novel BoAE modeling to represent the high-dimensional feature descriptor as BoAE representation. Compared with previous clustering or vector-quantization-based BoW modeling, the BoAE model can produce more accuracy and compact representation.
- Dictionary learning is adopted for BoAE modeling, which is an effective tool to model high-dimensional feature descriptors with low approximation error.
- We introduce a spatiotemporal gradient-based local motion pattern (STG-LMP) descriptor, which is distinctive, is fast to compute, and can account for both motion and texture.
- FLDA is introduced to discover the underlying structure in the BoAE representations, which has low computational complexity and memory requirements.

The rest of this paper is organized as follows: Section 2 describes the proposed framework in detail and the experimental results are compared and analyzed in Section 3. The conclusion is draw and the future work is sketched in Section 4.

## 2. The details of the proposed framework

Our framework consists of two phases: a training phase and a testing phase. A schematic representation of the framework is illustrated in Figure 2.

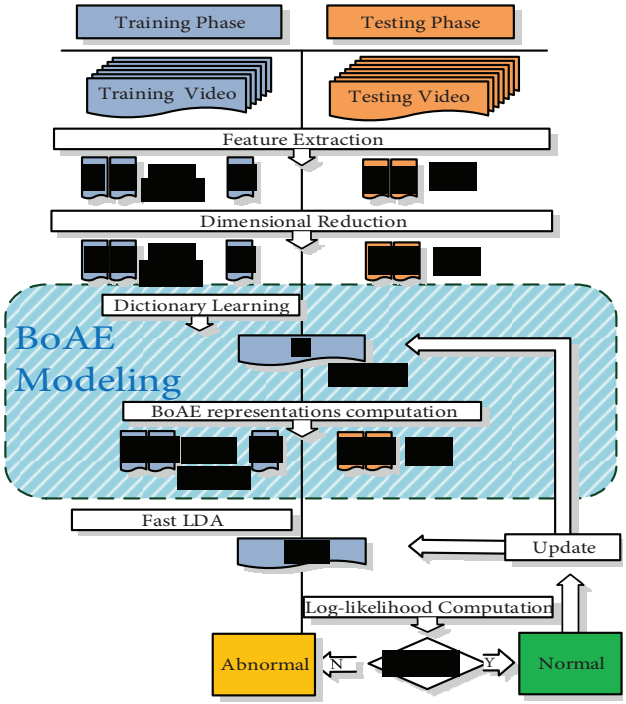


**Figure 2.** System overview of abnormal event detection framework.

### 2.1. Feature descriptor computation

Choosing an appropriate feature descriptor significantly influences the performance of abnormal detection systems, since in an appropriate feature space, an abnormal event is more salient than a normal one. A local motion pattern (LMP) descriptor was developed in [33], which is used for describing human action. Based on the LMP, we propose a STG-LMP descriptor for describing the crowd motion in the scene, which is computed for the spatiotemporal gradient magnitude of each pixel. This modification makes our descriptor able to account for both motion and texture in the scene. Consider a training sequence $V(x,y,t)$ with size of $H \times W \times L$, we first perform Gaussian smoothing of each frame to reduce the influence of noise. Then we compute the spatiotemporal gradient magnitude for each pixel as

$$V_G = \sqrt{G_x^2 + G_t^2 + G_t^2}, \tag{1}$$

where

$$G_x = V(x+1,y,t) - V(x-1,y,t), \tag{2}$$

$$G_y = V(x,y+1,t) - V(x,y-1,t), \tag{3}$$

$$G_t = V(x,y,t+1) - V(x,y,t-1). \tag{4}$$

Considering a spatiotemporal gradient magnitude sequence $V_G(x,y,t)$, we divide it into a set of clips $V_G = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_C]$ by sliding a window along the temporal axis, where each clip $\mathbf{V}_c$ is the size of $H \times W \times l$.

The window moves forward $\iota\,(\iota \le l)$ frames at each time step, so every two adjacent clips have $l-\iota$ frames overlapping. We further divide $\mathbf{V}_c$ into small cubes $v$ with size of $\rho \times \rho \times l$. We compute three central moments for each pixel along the temporal direction within cube $v$. The three center moments are the 2nd, 3rd, and 4th center moments denoted as $M_2$, $M_3$, and $M_4$, respectively. They reflect three important statistical properties, i.e. variance, skewness, and kurtosis, of the temporal change of the pixel gradient magnitude, respectively. We define the average moment value $\bar{M}_r$ for each cube as

$$\bar{M}_r = \frac{1}{\rho^2}\sum_{i=1}^{\rho}\sum_{j=1}^{\rho} m_{ij}, \quad r = \{2,3,4\}, \tag{5}$$

where

$$m_{ij} = \frac{1}{l}\sum_{t=1}^{l}(v_{ijt})^r, \quad 1 \le i \le \rho, 1 \le j \le \rho, \tag{6}$$

where $v_{ijt}$ is the spatiotemporal gradient magnitude of the pixel at location $\{i,j\}$ of the $t$th frame. The three average moments are formed into vector $\bar{\mathbf{M}} = \left[\bar{M}_2, \bar{M}_3, \bar{M}_4\right]$. The original feature descriptor of clip $\mathbf{V}_c$ is obtained by concatenating on the vector of all cubes in clip $\mathbf{V}_c$ as

$$u = \left[\bar{\mathbf{M}}_{hw}\right]^T, h = 1, \ldots, \left\lfloor H/\rho \right\rfloor, w = 1, \ldots, \left\lfloor W/\rho \right\rfloor, \tag{7}$$

where $\lfloor\cdot\rfloor$ is a rounded down function. The vector $u$ is the proposed STG-LMP. STG-LMP has two advantages: first, it is distinctive. Figure 3 shows the 3D surface of the STG-LMP descriptors extracted from UMN[1] scene 1. We can see from it that the abnormal events are more salient than the normal events in this feature domain. Second, the computational cost of STG-LMP extraction is very low. Table 1 lists the computation time of the proposed STG-LMP, multiscale histogram of optical flow (MHOF) [21], and MDT [13], respectively. From Table 1, we can observe that the proposed STG-LMP is much faster than MHOF and MDT.

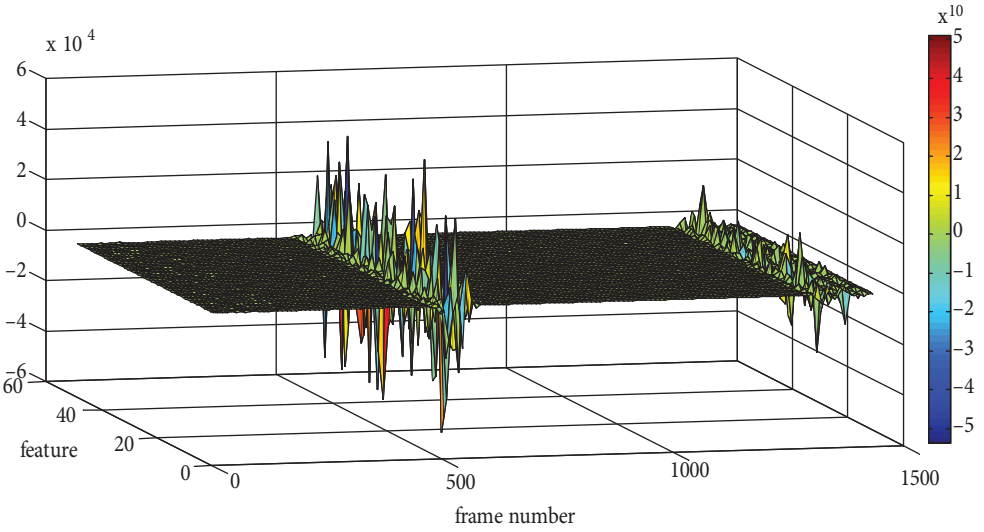


**Figure 3.** The surface of STG-LMP descriptors extracted from UMN scene 1.

[1] Unusual crowd activity dataset of the University of Minnesota. Available from http://mha.cs.umn.edu/movies/crowdactivity-all.avi.

**Table 1.** Quantitative comparisons between MHOF, MDT, and our proposed feature representation.

| Descriptor | Frame size | Run time |
|---|---|---|
| MHOF [21] | 0.957 | 0.57079 s |
| MDT [13] | 0.970 | 24s |
| STG-LMP | 0.993 | 0.14701 s |

The STG-LMP descriptor is high-dimensional and contains a large amount of redundant information. In order to reduce the dimension and capture the most salient information, we use RP [35] to reduce the high dimensionality. RP can preserve the distances between vectors quite reliably by projecting high-dimensional data onto a random lower-dimensional subspace. It is simple, fast, and data-independent and it can avoid the limitations of traditional methods like PCA. Considering a matrix $\mathbf{U} = [u_1, u_2, \ldots, u_C] \in R^{d \times C}$, where each column is a high-dimensional STG-LMP, we project the matrix $\mathbf{U}$ onto a low-dimensional random subspace $\mathbf{R} \in R^{q \times d}$ $(q \ll d)$, given by

$$\mathbf{Y} = \mathbf{R}\mathbf{U}. \tag{8}$$

The obtained low-dimensional matrix $\mathbf{Y} = [y_1, y_2, \ldots, y_C] \in R^{q \times C}$ contains projections of $\mathbf{U}$ on some random $q$ dimensional subspace. The random matrix $\mathbf{R}$ can be any zero-mean unit variance and normally distributed matrix. Figure 4 illustrates the process of extracting the spatial temporal features from video clips as well as the following dimensionality reduction.

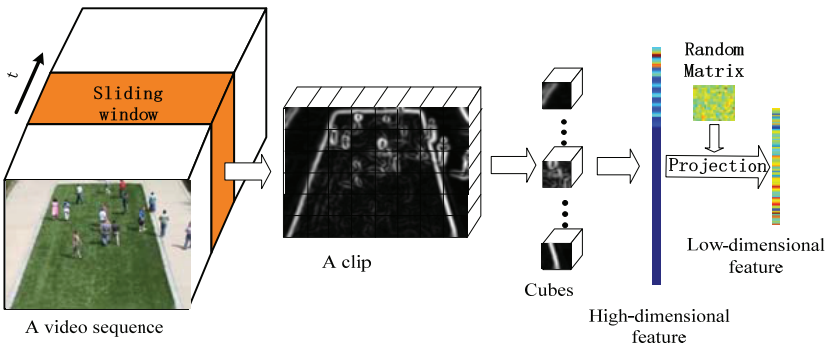


**Figure 4.** Illustration of the process of STG-LMP extraction and dimensionality reduction.

### 2.2. BoAE modeling

In this subsection, we model the STG-LMP descriptors as BoAE representations by BoAE modeling. First, we briefly describe the core of BoAE modeling: dictionary learning. Considering a set of feature descriptors in matrix $\mathbf{Y} \in R^{q \times C}$, the goal of dictionary learning is to learn a dictionary $\mathbf{D} \in R^{q \times b}$ that represents the input $\mathbf{Y}$ approximately as $\mathbf{Y} = \mathbf{D}\mathbf{S} + e$ using dictionary $\mathbf{D}$ and the corresponding sparse representations $\mathbf{S} = [s_1, s_2, \ldots, s_C] \in R^{b \times C}$, where $e$ is an additive component with bounded energy $(\|e\|_2^2)$ modeling both the noise and deviation from the model. The dictionary can be overcomplete $(b > q)$, and can thus capture a large number of high-level patterns in the input dataset. The optimization problem is formally written as

$$\arg\min_{\mathbf{D},\mathbf{S}} \frac{1}{2} \|\mathbf{D}\mathbf{S} - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{S}\|_1, \tag{9}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\|\cdot\|_1$ is the $\ell_1$-norm. In this paper, we solve the optimization problem using K-SVD [36], which solves Eq. (9) by performing sparse coding and dictionary updates at every iteration. We keep a fixed $\mathbf{D}$ and compute $\mathbf{S}$ in the sparse coding step as

$$\arg\min_{\mathbf{S}} \frac{1}{2}\|\mathbf{DS}-\mathbf{Y}\|_F^2 + \lambda\|\mathbf{S}\|_1 . \tag{10}$$

We then update the dictionary $\mathbf{D}$ in the dictionary update step as

$$\mathbf{D}^{(\tau+1)} = \mathbf{D}^{(\tau)} + \eta\mathbf{E}\mathbf{S}^{(\tau)T}, \tag{11}$$

where

$$\mathbf{E}_i = \mathbf{Y} - \tilde{\mathbf{D}}_i\tilde{\mathbf{S}}_i, \tag{12}$$

where $\tilde{\mathbf{D}}_i$ refers to $\mathbf{D}$, which is the $i$th column is removed, and $\tilde{\mathbf{S}}_i$ refers to $\mathbf{S}$, which is the $i$th row is removed. For more details about K-SVD, please refer to [36].

After dictionary learning, we obtain the dictionary $\mathbf{D}$ and the sparse representations of matrix $\mathbf{S}$ corresponding to $\mathbf{Y}$ by solving Eq. (10). Each row $S_i \in R^b$ of $\mathbf{S}$ contains $k$ ($k \ll b$) nonzero coefficients. The coefficients indicate the contribution of each dictionary atom in the data reconstruction. The positive value of the element indicates that the corresponding dictionary atom is additive in data reconstruction, and the negative value indicates that the corresponding dictionary atom is subtractive. In order to account for both additive and subtractive dictionary atoms, we define an atomic event dictionary that consists of positive and negative atomic events. The positive atomic event corresponds to an additive dictionary atom, and the negative atomic event corresponds to a subtractive dictionary atom. Consequently, the atomic event dictionary has $2b$ atomic events. On the other hand, LDA only supports discrete input. Given a feature descriptor, its corresponding BoAE representation $\mathbf{w}$ is computed as follows:

$$\mathbf{w}^{b\times\left|sgn(S^i-|S^i|)\right|+i} = \left\lfloor |S^i| \right\rfloor, \quad 1 \le i \le b, \tag{13}$$

where $sgn(\cdot)$ is the signum function. $\mathbf{w}^i$ corresponds to the frequency of the $i$th atomic event in the atomic event dictionary, and $S^i$ is the $i$th element in the feature descriptor corresponding to sparse representation. Through the BoAE modeling, the feature descriptor is represented as a BoAE representation, and we treat the BoAE representation as a document in which each word corresponds to one type of atomic event.

### 2.3. FLDA training

Given a corpus $W = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C\}$, each document $\mathbf{w}_c$ is modeled as a mixture of $K$ assumed known topics, and each topic is modeled as a multinomial distribution over a vocabulary by LDA. In our framework, documents, topics, and words correspond to feature descriptors corresponding to BoAE representations, events, and atomic events, respectively. The generative process of LDA for each document $\mathbf{w}$ in the collection $W$ is as follows (Figure 5) [29]:

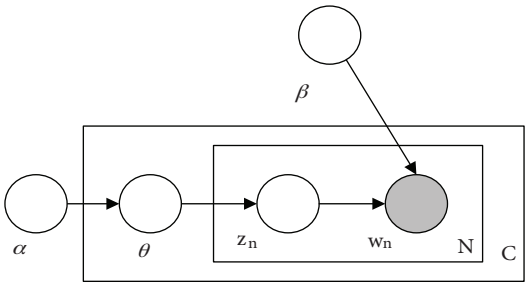


**Figure 5.** Graphic models for LDA.

1) Choose $\theta \sim Dir(\alpha)$.

2) For each of the $N$ words $w_n$:

   a) Choose a topic $z_n \sim Mult(\theta)$;

   b) Choose a word $w_n$ from $w_n \sim p(w_n | z_n, \beta)$.

In a particular document, the mixing proportion of different topics is indicated by parameter $\theta$. The parameter $\alpha$ is a $K$-dimensional Dirichlet prior shared by all documents. The parameter $\beta$ is a $K \times V$ matrix, where each row is a $V$-dimensional distribution of words within a particular topic $z_n$, where $V$ is the size of vocabulary. The probability of a document $\mathbf{w} = \{w_1, w_2, \ldots, w_N\}$ is given by

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left( \prod_{n=1}^{N} \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta. \tag{14}$$

The optimal parameters $\alpha*$ and $\beta*$ can be estimated by maximizing the log-likelihood of the corpus:

$$(\alpha*, \beta*) = \arg\max_{\alpha,\beta} \sum_{c=1}^{C} \log P(\mathbf{w}_c | \alpha, \beta) . \tag{15}$$

Traditional solutions of this problem are variational approximation and Gibbs sampling. However, these algorithms are too computationally expensive to be applied to large-scale datasets. In order to speed up the computation, we adopt FLDA [34] into our framework. FLDA uses a substantially smaller number of variational parameters, with no dependency on the dimensionality of the training set. This leads to the result that the FLDA is orders of magnitudes faster than the original LDA.

The variational distribution in FLDA introduces one Dirichlet distribution parameterized by $\theta$ and one Discrete distribution parameterized by $\phi$ for each document $\mathbf{w}$. The variational distribution is given by:

$$q_2(\theta, \mathbf{z} | \phi, \gamma) = q_2(\theta | \gamma) \prod_{n=1}^{N} q_2(z_n | \phi). \tag{16}$$

Compared to the original variational distribution,

$$q_1(\theta, \mathbf{z} | \phi, \gamma) = q_1(\theta | \gamma) \prod_{n=1}^{N} q_1(z_n | \phi_n), \tag{17}$$

it only uses one discrete distribution parameter $\phi$ for each document. The complexity is reduced to $O(N)$ from $O(N^2)$, so FLDA is much faster than LDA. Meanwhile, the number of $\phi$s is reduced to 1 from $N$ for each document, so the memory requirement is accordingly saved. For more details about FLDA training, please refer to [34].

### 2.4. Abnormal event detection and update

After the training phase, we will classify the testing sample as normal or abnormal in the testing phase. Given a testing sequence, it is first subjected to feature extraction, dimensionality reduction, and BoAE modeling. The log-likelihood of each BoAE representation $\mathbf{w}$ is calculated under the trained FLDA as

$$\ell(\mathbf{w}|\alpha,\beta) = \log p(\mathbf{w}|\alpha,\beta). \tag{18}$$

To further reduce the influence of noise, we smooth the log-likelihood distribution with a Gaussian smoothing filter. Figures 6a and 6b show the normalized log-likelihood distributions of two sequences from the UMN and PETS 2009[2] dataset, respectively. The distribution curves demonstrate that our proposed framework is not only robust to noise but also responds to abnormal events rapidly. Finally, we label testing samples as normal or abnormal based on the empirical threshold $\delta$.

$$Label = \begin{cases} Normal & ,\ell(\mathbf{w}|\alpha,\beta) > \delta \\ Abnormal & ,\ell(\mathbf{w}|\alpha,\beta) \leq \delta \end{cases}. \tag{19}$$

In order to adapt the environment gradually to change, such as illumination change, it is necessary to incrementally update the dictionary and LDA parameter. Given an identified normal testing feature descriptor, we use it to update the dictionary atom by an online dictionary update algorithm [22], and then use its BoAE representation to update the parameters in FLDA by the online-LDA algorithm [37].

## 3. Experiments

In this section, we conduct experiments on different datasets for evaluating the performance of our proposed framework. All experiments are conducted on a PC with Intel Core i3 CPU at 2.13 GHZ and 2G RAM. The UMN, PETS 2009, UCSD[3], and real highway traffic video are used as the experimental dataset. These datasets exhibit various events in different scenes, such as indoor and outdoor scenes, and both global and local abnormal events are included.

### 3.1. UMN dataset

The UMN dataset is captured from 3 different scenes, including indoor and outdoor scenes, and its resolution is $320 \times 240$. The abnormal event in the dataset is a crowd running suddenly. We portion each scene into two parts; the first part has 400 frames only containing normal events, and we use it as the training set, and the rest is used as a testing set, which contains both normal and abnormal events. In the training stage, video sequences are portioned into clips with size of $320 \times 240 \times 5$ by a sliding window. For each clip, we further divide it into cubes with size of$20 \times 20 \times 10$. The STG-LMP descriptor is computed for each cube and then concatenated into a high-dimensional feature descriptor. We reduce its dimension from 576 to 70 by RP. The overcomplete

[2] Pets 2009 dataset. http://ftp.cs.rdg.ac.uk/PETS2009/.
[3] UCSD Ped1 dataset. http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm.

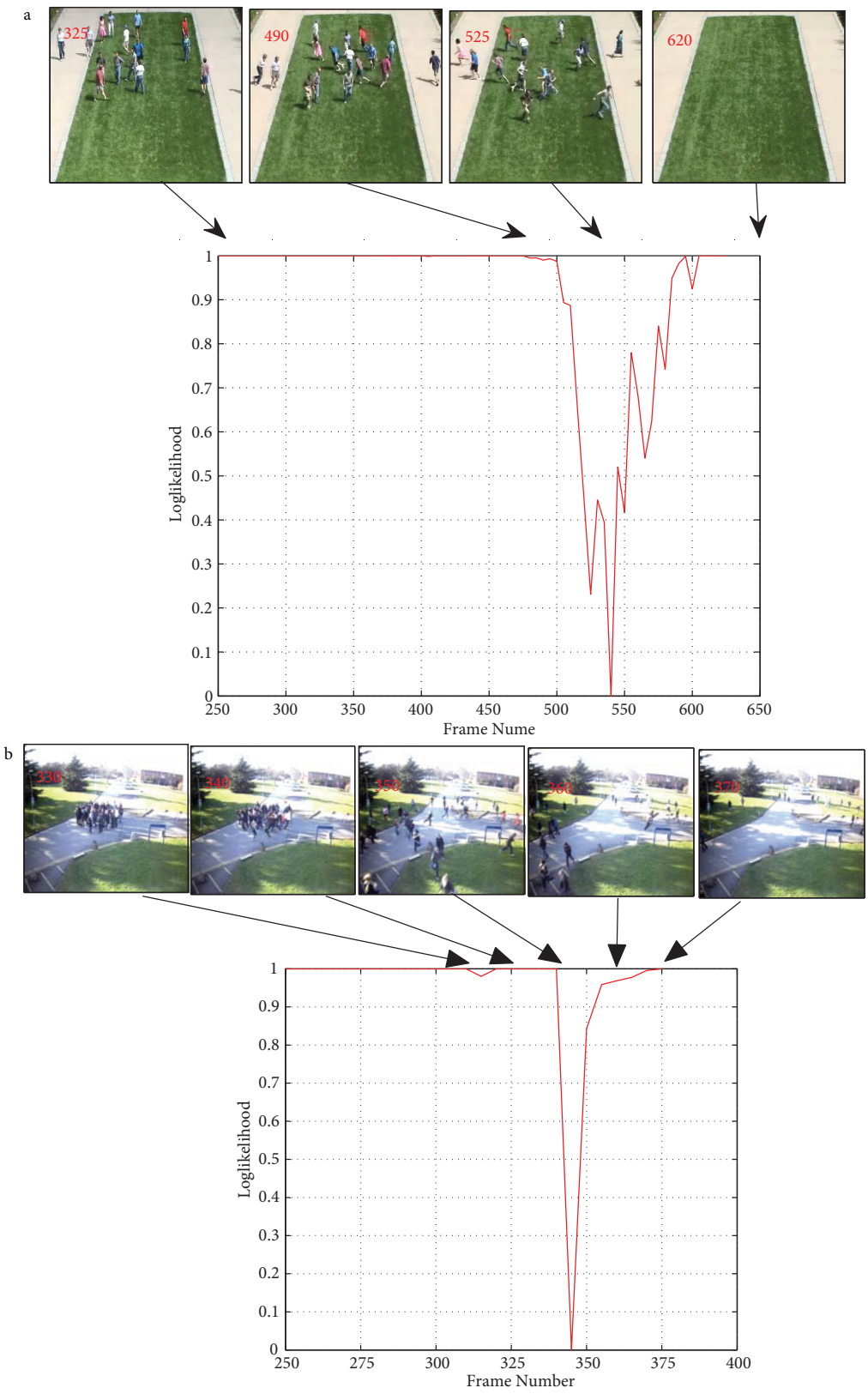


**Figure 6.** The log-likelihood distributions of two sequences from UMN and PETS 2009 datasets, respectively: a) UMN dataset, b) PETS 2009 dataset.

dictionary **D** is learned by a K-SVD algorithm with 10 iterations. The number of dictionary atoms is 120, and the number of latent topics of LDA is 10. Figure 7 shows some detection results of abnormal event detection from the UMN dataset. We compare our method with the K-means-based BoW method and other state-of-the-art methods [11, 15, 21, 23, 33]. Figure 8 plots the ROC curves of different methods. Table 2 provides their AUC values for comparison. We can see from Table 2 that our method outperforms the K-means-based BoW method and is comparable to other state-of-the art methods.

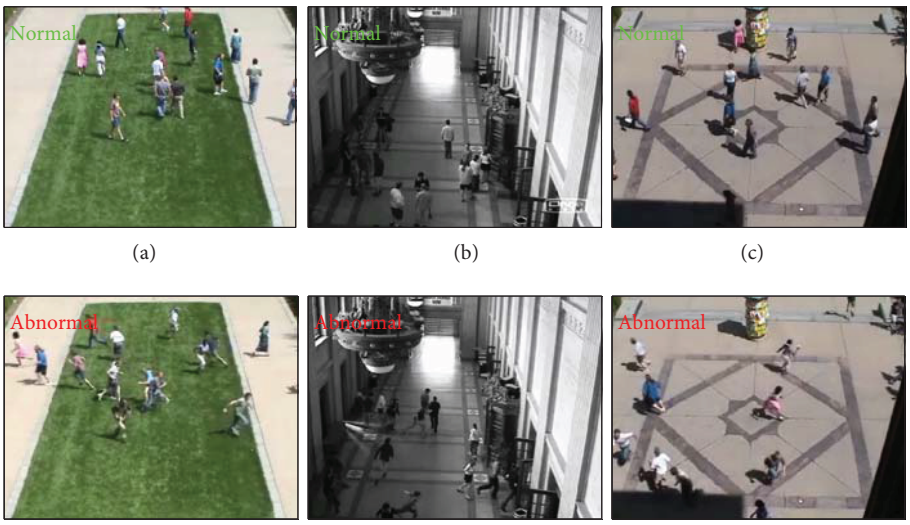


**Figure 7.** The detection results of our proposed method for three scenes for UMN dataset.

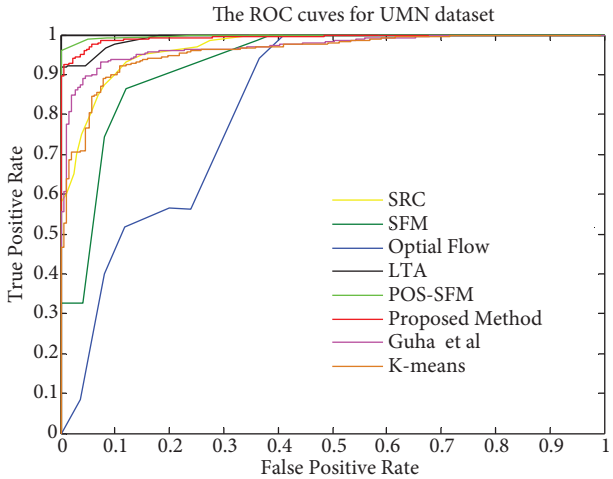


**Figure 8.** Comparison of abnormal event detection results for UMN dataset.

### 3.2. PETS 2009 dataset

In this subsection, we conduct experiments on View001 and View002 sequences in the PETS 2009 dataset. Each sequence has 378 frames and the abnormal event begins from frame 336. The normal event is people merging with normal speed, and the abnormal one is people dispersing suddenly. The first 250 frames of each sequence are used as training samples and the rest as testing samples. The frame size is resized to $320 \times 240$, and the

other parameters are the same as in the experiment on the UMN dataset. Figure 9 shows some detection results of our method. Figure 10 plots the ROC curves of our method and the other state-of-the art methods [23, 24]. Table 3 provides the AUC values for comparison. It is obvious that the performance of our proposed framework outperforms the other state-of-the-art methods.

**Table 2.** Summary of quantitative system performance and the comparison with the state-of-the-art methods for UMN dataset according to AUC value.

| Representation | AUC |
|---|---|
| Chaotic [21] | 0.99 |
| Social force [11] | 0.96 |
| Optical flow | 0.84 |
| MHOF [21] | 0.978 |
| PSO-social force[23] | 0.996 |
| Interaction potential [15] | 0.992 |
| BoW | 0.957 |
| LMP [33] | 0.970 |
| BoAE | 0.993 |

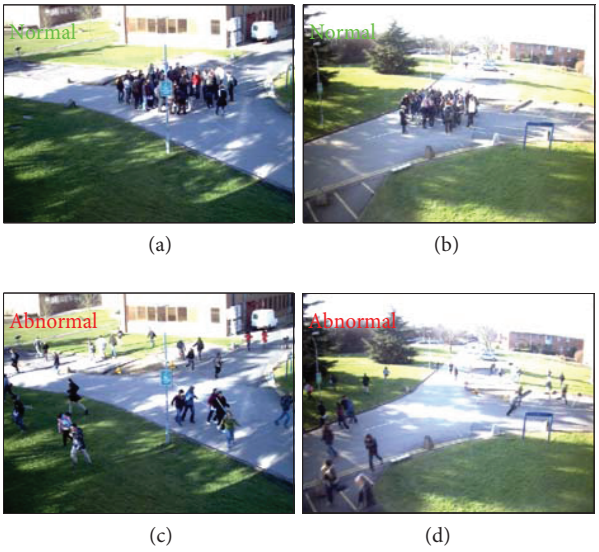


**Figure 9.** The detection results of our proposed method for PETS 2009 dataset.

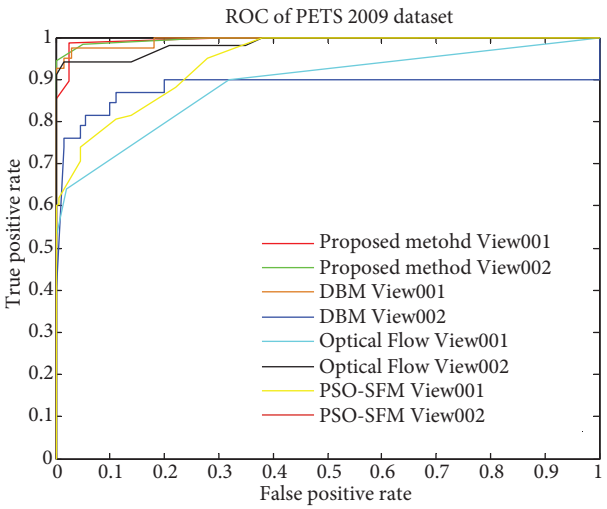


**Figure 10.** Comparison of abnormal event detection results for PETS 2009 dataset.

**Table 3.** Summary of quantitative system performance and the comparison with the state-of-the-art methods for PETS 2009 dataset according to AUC value.

| Representation | AUC (View001) | AUC (View002) |
|---|---|---|
| DBM [24] | 0.9939 | 0.8784 |
| PSO-SFM | 0.9414 | 0.9914 |
| Optical flow | 0.9801 | 0.8834 |
| BoAE | 0.9945 | 0.9979 |

### 3.3. UCSD Ped1 dataset

In this subsection, we validate the efficiency of our method on the UCSD Ped1 dataset. The normal event in this dataset is a crowd moving with normal speed, and abnormal events include skaters, cars speeding, and people cycling on the walkway. We resize its resolution from $158 \times 238$ to $160 \times 240$. The training set contains 34 clips of normal events, and the testing set contains 36 testing clip. Each frame is split into $16 \times 16$ local regions with 8 pixels overlapping. For each local region, we extract STG-LMP descriptors from the cuboids with size of $16 \times 16 \times 10$. The dimension of the original feature descriptor is reduced to 40. The number of dictionary atoms is 70 and the topic number is 10. A spatiotemporal smoothing is adopted to further eliminate the influence of noise, which can be seen as a simple version of a spatiotemporal Markov model. Figure 11 shows different types of detected abnormal events. Figure 12 plots the ROC curves of our method and other state-of-the-art methods [11–13, 25, 26]. The performances of different methods are evaluated by equal error rate (EER) values, where the lower the EER is the better the performance is. We can see from Table 4 that the performance of our method is better than other state-of-the-art methods.

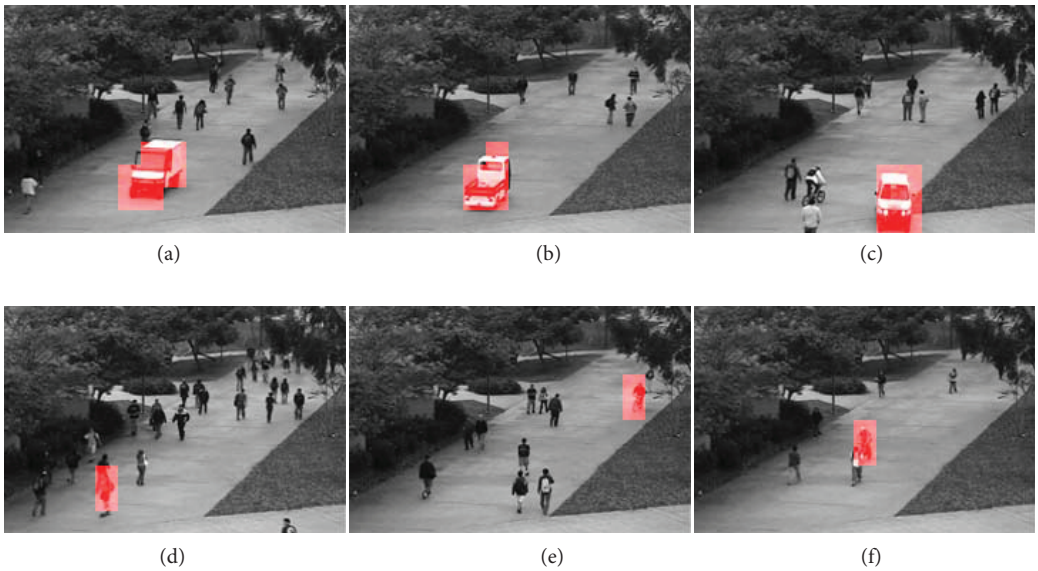


**Figure 11.** The detection results of our proposed method for UCSD Ped1 dataset. The abnormal events, such as bicyclers, skaters, and cars, are detected by red masks.

**Table 4.** Summary of quantitative system performance and the comparison with the state-of-the-art methods for UCSD Ped1 dataset according to EER value.

| Representation | EER |
|---|---|
| MPPCA [12] | 40% |
| Adam et al. [25] | 38% |
| Bertini et al. [26] | 31% |
| SFM | 31% |
| MDT | 25% |
| BoAE | 23% |

### 3.4. Highway traffic dataset

Beyond crowd abnormal event detection, our proposed framework can also be applied in traffic surveillance in daily life. The traffic video sequences were captured by a highway surveillance camera in the real world. The

abnormal event is a traffic accident that occurred suddenly. Figure 13 shows the detection results of abnormal events, such as illegal parking and traffic accidents. The AUC value of our method is 0.9892.

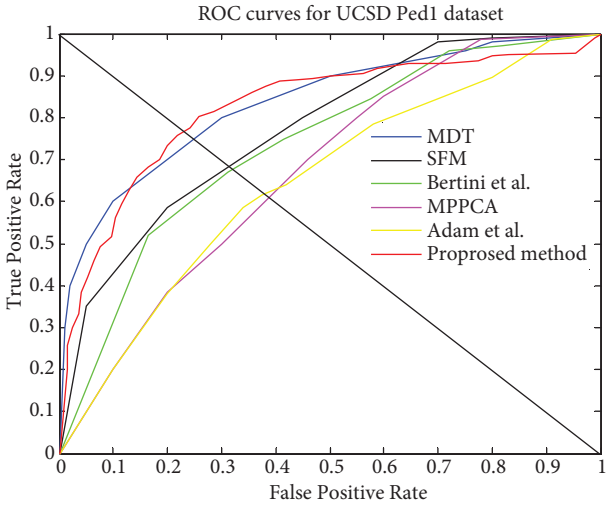


**Figure 12.** Comparison of abnormal event detection results for UCSD Ped1 dataset.


**Figure 13.** The detection results of abnormal event detection for highway traffic video.

## 4. Conclusion and future work

In this paper, we address the problem of abnormal event detection in crowded scenes by developing a novel BoAE modeling. In BoAE modeling, a high-dimensional feature descriptor is modeled as a collection of atomic events. Compared with clustering or vector quantization-based BoW modeling, which represent descriptors by only one code word, BoAE representation has lower approximation error and is more suitable to represent the abnormal event. The BoAE reorientations are used to train FLDA, and then the testing BoAE representation with low likelihood under the trained FLDA is classified as abnormal. By taking advantage of LMP descriptors, RP, BoAE representations, and FLDA, we show that the proposed framework is robust and efficient and it works well in crowded scenes. The experimental results analysis and comparisons confirm our claims. In the future, we will investigate the possibilities of incorporating multifeature and multiscale methods for abnormal event detection.

## Acknowledgment

This paper was jointly supported by the National Natural Science Foundation of China, "61374161" and "61074106".

## References

[1] Krausz B, Bauckhage C. Loveparade 2010: Automatic video analysis of a crowd disaster. Comput Vis Image Und 2012; 116: 307-319.

[2] Stauffer C, Grimson WEL. Learning patterns of activity using real-time tracking. IEEE T Pattern Anal 2000; 22: 747-757.

[3] Piciarelli C, Micheloni C, Foresti GL. Trajectory-based anomalous event detection. IEEE T Circ Syst Vid 2008; 18: 1544-1554.

[4] Chen TP, Haussecker H, Bovyrin A, Belenov R, Rodyushkin K, Kuranov A, Eruhimov V. Computer vision workload analysis: case study of video surveillance systems. Intel Techn J 2005; 9: 109-118.

[5] Johnson N, Hogg D. Learning the distribution of object trajectories for event recognition. Image Vision Comput 1996; 14: 609-615.

[6] Jiang F, Wu Y, Katsaggelos AK. A dynamic hierarchical clustering method for trajectory-based unusual video event detection. IEEE T Image Process 2009; 18: 907-913.

[7] Zhang Y, Qin L, Yao H, Huang Q. Abnormal crowd behavior detection based on social attribute-aware force model. In: 19th IEEE International Conference on Image Processing; 30 September–3 October 2012; Orlando, FL, USA. pp. 2689-2692.

[8] Zhang Y, Huang Q, Qin L, Zhao S, Yao H, Xu P. Dense crowd event recognition using bag of trajectory graphs. SIGNAL Image Video P 2014; 8 (Suppl. 1): S173-S181.

[9] Zhang Y, Zhang S, Huang Q, Thomas S. Learning sparse prototypes for crowd perception via ensemble coding mechanisms. In: 5th International Workshop on Human Behavior Understanding; 12 September 2014. pp 86-100.

[10] Zhang Y, Qin L, Ji R, Yao, H, Huang Q. Social attribute-aware force model: exploiting richness of interaction for abnormal crowd detection. IEEE T Circ Syst Vid 2014; 25: 1231-1245.

[11] Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model. In: IEEE 2009 Computer Vision and Pattern Recognition; 19–26 June 2009; Miami Beach, FL, USA. pp. 935-942.

[12] Kim J, Grauman K. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: IEEE 2009 Computer Vision and Pattern Recognition; 19–26 June 2009; Miami Beach, FL, USA. pp. 2921-2928.

[13] Mahadevan V, Li W, Bhalodia V, Vasconcelos N. Anomaly detection in crowded scenes. In: IEEE 2010 Computer Vision and Pattern Recognition; 13–18 June 2010; San Francisco, CA, USA. pp. 1975-1981.

[14] Kratz L, Nishino K. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: IEEE 2009 Computer Vision and Pattern Recognition; 19–26 June 2009; Miami Beach, FL, USA: IEEE. pp. 1446-1453.

[15] Cui X, Liu Q, Gao M, Metaxas DN. Abnormal detection using interaction energy potentials. In: IEEE 2011 Computer Vision and Pattern Recognition; 20–25 June 2011; Colorado Springs, CO, USA. pp. 3161-3167.

[16] Haidar Sharif M, Djeraba C. An entropy method for abnormal activities detection in video streams. Pattern Recogn 2012; 45: 2543-2561.

[17] Saligrama V, Konrad J, Jodoin P. Video anomaly identification. IEEE Signal Proc Mag 2010; 27: 18-33.

[18] Wang X, Ma X, Grimson WEL. Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. IEEE T Pattern Anal 2009; 31: 539-555.

[19] Li J, Gong S, Xiang T. Learning behavioural context. Int J Comput Vision 2012; 97: 276-304.

[20] Zhao B, Fei-Fei L, Xing EP. Online detection of unusual events in videos via dynamic sparse coding. In: IEEE 2011 Computer Vision and Pattern Recognition; 20–25 June 2011; Colorado Springs, CO, USA. pp. 3313-3320.

[21] Cong Y, Yuan J, Liu J. Abnormal event detection in crowded scenes using sparse representation. Pattern Recogn 2012; 46: 1851-1864.

[22] Mairal J, Bach F, Ponce J. Online dictionary learning for sparse coding. In: ACM 2009 International Conference on Machine Learning; 14–18 June 2009; Montreal, Canada. pp. 689-696.

[23] Raghavendra R, Del Bue A, Cristani M. Optimizing interaction force for global anomaly detection in crowded scenes. In: IEEE 2011 International Conference on Computer Vision Workshop; 20–25 June 2011; Colorado Springs, CO, USA. pp. 136-143.

[24] Xu J, Denman S, Fookes C, Sridharan S. Unusual scene detection using distributed behaviour model and sparse representation. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance; 18–21 September 2012; Beijing, China. pp. 48-53.

[25] Adam A, Rivlin E, Shimshoni I, Reinitz D. Robust real-time unusual event detection using multiple fixed-location monitors. IEEE T Pattern Anal 2008; 30: 555-560.

[26] Bertini M, Del Bimbo A, Seidenari L. Multi-scale and real-time non-parametric method for anomaly detection and localization. Comput Vis Image Und 2012; 116: 320-329.

[27] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Comput Surv 2009; 41: 15.

[28] Castrodad A, Sapiro G. Sparse modeling of human actions from motion imagery. Int J Comput Vision 2012; 100: 1-15.

[29] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003; 3: 993-1022.

[30] Vretos N, Nikolaidis N, Pitas I. Video fingerprinting using latent Dirichlet allocation and facial images. Pattern Recogn 2012; 45: 2489-2498.

[31] Wang Y, Mori G. Human action recognition by semi-latent topic models. IEEE T Pattern Anal 2009; 31: 1762-1774.

[32] Wright J, Ma Y, Mairal J, Sapiro G, Huang TS, Yan S. Sparse representation for computer vision and pattern recognition. P IEEE 2010; 98: 1031-1044.

[33] Guha T, Ward RK. Learning sparse representations for human action recognition. IEEE T Pattern Anal 2012; 34: 1576-1588.

[34] Shan H, Banerjee A. Mixed-membership naive Bayes models. Data Min Knowl Disc 2011; 23: 1-62.

[35] Baraniuk RG, Wakin MB. Random projections of smooth manifolds. Found Comput Math 2009; 9: 51-77.

[36] Aharon M, Elad M, Bruckstein A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE T Signal Proces 2006; 54: 4311-4322.

[37] Hoffman M, Bach FR, Blei DM. Online learning for latent Dirichlet allocation. In: Advances in Neural Information Processing Systems; 2010. pp. 856-864.