Ye R, Li X. Collective representation for abnormal event detection. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 32(3): 470–479 May 2017. DOI 10.1007/s11390-017-1737-8

Collective Representation for Abnormal Event Detection

Renzhen Ye [1,2] and Xuelong Li [1], *Fellow, IEEE*

[1] *Center for Optical Imagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China*
[2] *School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710119, China*

E-mail: yerenzhen@mail.hzau.edu.cn; xuelong_li@opt.ac.cn

Received December 26, 2016; revised March 21, 2017.

Abstract Abnormal event detection in crowded scenes is a hot topic in computer vision and information retrieval community. In this paper, we study the problems of detecting anomalous behaviors within the video, and propose a robust collective representation with multi-feature descriptors for abnormal event detection. The proposed method represents different features in an identical representation, in which different features of the same topic will show more common properties. Then, we build the intrinsic relation between different feature descriptors and capture concept drift in the video sequence, which can robustly discriminate between abnormal events and normal events. Experimental results on two benchmark datasets and the comparison with the state-of-the-art methods validate the effectiveness of our method.

Keywords abnormal detection, collective representation, dictionary learning

1 Introduction

In surveillance video, the automated detection and localization of anomalous behavior have become an active research area of computer vision and pattern recognition due to the increasing demand for security and safety[1-3]. The goal of abnormal event detection is to detect unusual behavior of individuals or a group in a crowded scene[4]. While much work on anomaly event detection has been reported in recent years, it is still a challenging problem to develop a robust method due to the general difficulties of the anomaly detection problem in crowded scenes. For crowds, for example, it is infeasible to list the set of anomalies that are possible in numerous vision applications[5-8].

One common solution to these problems is to find new patterns in data that do not conform to the expected case[9]. This is compounded by fitting a statistical model of anomaly detection, which tries to detect events with low probability as abnormality. However, it introduces a number of challenges. First, it needs a high-dimensional feature to better represent the event and train the statistical model. In this case, the number of training samples will increase exponentially with the feature dimension. In practice, it is difficult to collect enough data to train a statistical model. Second, crowded scenes require a statistical detection model robust to complex and dynamic scenes, containing a large number of moving persons that occlude each other in complex ways, and can have low resolution. Hence, it is difficult to effectively identify all abnormal behaviors. Third, different tasks may require different models of normalcy. However, it is unrealistic to collect sufficient samples of abnormal video events, which brings challenges to build a robust video anomaly detector[10-11].

Recently, researches have shown that high-dimensional natural signals of the same class usually lie in a low-dimensional subspace[12-14]. Hence, for a given sample, it can be represented by a linear combination of a few training samples from an overcomplete dictionary[15]. Inspired by recent advantages in sparse coding, sparse representation based methods have been exploited in video anomaly detection[16-17]. The main underlying assumption of these methods is that normal events can be well constructed by the normal basis with a small reconstruction cost, while abnormal events can-

Regular Paper

Special Section of CVM 2017

This work is supported by the Defense Basic Research Program under Grant No. BA1320110042, and the Open Research Fund of the Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences under Grant No. LSIT201408.

not with a large reconstruction cost. In sparse representation (SR) based methods, event representation is to extract distinguishable feature descriptors for different events. To enhance detection performance, several kinds of low-level feature descriptors, such as multiscale histogram of optical flow (MHOF), Gabor filter and object trajectory[18-19], saliency features[20-24], 3-D video patches[25], spatial-temporal gradient[26], and chaotic invariant features[27], are exploited to generate event representation. Each of these features can only describe certain aspects of object feature. In particular, MHOF descriptor considers amplitudes and directions of movements, while Gabor filter descriptors provide texture representation and discrimination. However, sparse representation based methods combine only these low-level feature descriptors to generate sparse event representation. In this case, these sparse representation (SR) based methods fail to build the intrinsic relation between different feature descriptors and capture high-level latent semantic information which has been proved that it can obtain better performance in computer vision.

In this paper, we propose a robust collective representation with multi-feature descriptors for abnormal event detection. The proposed method represents different features in an identical representation, in which different features of the same topic will show more common properties. In fact, the proposed method exploits sparse representation to capture the salient structures of different descriptors, and generate the same representation with respect to different dictionaries. Then the proposed method builds the intrinsic relation between different feature descriptors and captures concept drift in the video sequence. Finally, using sparse coding enables the algorithm to robustly discriminate between abnormal events and normal events.

The contributions of the proposed method can be summarized as follows.

1) We propose a robust collective representation for abnormal event detection, in which different features can be represented in an identical representation. In this case, the salient structures of different descriptors can be captured and multiple information will be merged effectively, which will increase the accuracy of anomaly detection.

2) We build the intrinsic relation between different feature descriptors and capture concept drift in the video sequence. In this case, the latent structure relation of different features generated from video events can be presented to detect the abnormal event, which outperforms the existing methods.

3) To evaluate the performance, a set of experiments are conducted on two publicly available video datasets to verify the effectiveness of the proposed method.

2 Related Work

In recent years, many anomaly detection algorithms[27-30] have been proposed. Reviews about abnormal event detection and human activity understanding can be referred to [31-33]. According to the specific applications, anomaly detection algorithms can be divided into two classes.

1) For uncrowded scenes, there are few moving objects in the scene, and they seldom interact with each other.

2) For crowded scenes, many objects exist in the scene, and they move complicatedly.

In uncrowded scenes, video events are usually represented with binary features based on background subtraction methods such as normalization cut clustering[34], or based on 3-D spatio-temporal foreground mask[35]. There are also many trajectory-based features based on frame-difference or tracking[36-39]. Trajectories occurring at much lower probabilities are treated as anomalies. For example, Wu *et al.*[27] extracted chaotic invariant features from trajectories, and modeled motion patterns with a probabilistic framework. Cheng and Hwang[28] conducted reliable tracking with an adaptive particle sampling and the Kalman filtering. Cui *et al.*[29] represented the crowd dynamic by tracking interest points and calculating interaction energy potentials. Trajectory-based features are generally of high-level semantics[40], but they may fail in density crowded scenes due to inevitable overlaps and occlusions.

In a crowded scene, video anomaly detection methods are generally developed with features based on local 2-D image patches or 3-D video blocks, such as spatio-temporal gradients and histograms of optical flow (HOF). Abnormal events are detected as ones rarely happening and divergent from normalcies. Adam *et al.*[41] represented the probability of optical flow at a group of spatial locations with histograms. Kim and Grauman[42] modeled local optical flows using a mixture of probability principle component analysis (MPPCA), and adopted a Markov random field to deal with space-time interactions. In [7], Mehran *et al.* proposed a social force (SF) model to analyze crowd behaviors. Kratz and Nishino[26] fit spatio-temporal gradient features with a statistical framework. In [43], Mahadevan *et al.* jointly modeled appearances and dynamics of the crowd using a mixture of dynamic textures (MDT). Based on the MDT, Li *et al.*[44] proposed a hierarchical MDT (H-MDT) algorithm, which takes advantages of background subtraction and discriminant saliency to detect temporal and spatial abnormal events, respectively. Cong *et al.*[1,45] described video events with a multi-scale HOF. By estimating sparse representation coefficients corresponding to a normal event dictionary, abnormal events are identified as samples owing to large reconstruction costs. Lu *et al.*[16] learned video event representations with a combination of several small dictionaries, which greatly increases the testing speed. In [46], Thida *et al.* extracted crowd features with a spatio-temporal Laplacian eigenmap, and learned their variations in an embedded space. Kaltsa *et al.*[47] described the characteristics of scenes with histograms of oriented gradients, which are newly introduced based on swarm theory.

3 Proposed Method

In this section, a novel sparse representation technique is formulated by considering the correction of different feature descriptors. We restrict the discussion to different feature descriptors consisting of multi-scale histogram of optical flow (MHOF) and histogram of oriented gradient (HOG) as they are appropriate for representation and widely used in computer vision.

3.1 Event Representation

Event representation aims to extract distinguishable and effective features to represent video events. Since appearance and motion patterns are the main differences between normal and abnormal events, in this paper, appearance and motion features are extracted to represent each spatio-temporal patch together. First, input image sequences are divided into overlapped spatio-temporal patches. Second, to filter out distractions (e.g., waving trees, illumination changes), foreground segmentation is exploited to estimate the background in crowded scenes[48]. Finally, object appearance is described by the spatial derivatives on horizontal and vertical directions, while the motion information of the scene is represented with the optical flow at each location. In this paper, histogram of oriented gradient (HOG) and multi-scale histogram of optical flow (MHOF) are exploited as appearance and motion features, respectively.

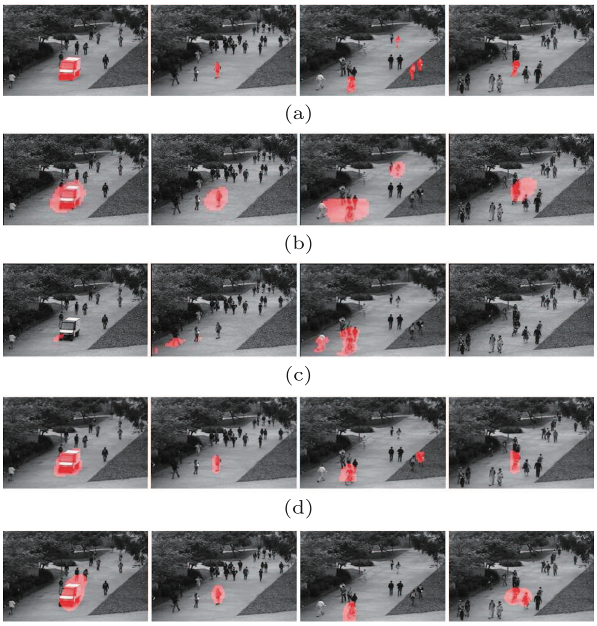
3.2 Latent Relation Cross Different Features

In this subsection, we explore the latent relation cross HOG feature and MHOF feature and assume that each feature of a video event generates identical representation with a given dictionary. For each video event, we learn the representation by collective representation learning with latent semantic model from different features.

Suppose that a training dataset $\boldsymbol{O} = \{\boldsymbol{o}_i\}_{i=1}^{n}$ consists of MHOF feature and Gabor feature of the same sample, i.e., $\boldsymbol{o}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^m$ is the m-dimensional MHOF feature, and $\boldsymbol{z}_i \in \mathbb{R}^d$ is the d-dimensional Gabor feature. In this paper, the purpose of the proposed method is to adaptively learn a correlation matrix which can bridge the semantic gap between different feature descriptors. As illustrated in Fig.1, we project different feature descriptors to the representation space respectively:

$$\boldsymbol{T}_1 : \mathbb{R}^m \to \boldsymbol{H}_1^M, \quad \boldsymbol{T}_2 : \mathbb{R}^d \to \boldsymbol{H}_2^D,$$

where $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ denote the projections, $\boldsymbol{R}$ stands for the feature descriptor, and $\boldsymbol{H}$ is the representation space. $\boldsymbol{M}$ and $\boldsymbol{D}$ are the dimension of $\boldsymbol{H}_1^M$ and $\boldsymbol{H}_2^D$, respectively. Then different representations are mapped into a common high-level abstraction space by linear projection as follows.


(a)

(b)

(c)

(d)

(e)

Fig.1. Four examples of abnormal event detections in UCSD Ped1 dataset. (a) Ground-truth. (b) Results of the MDT algorithm[43]. (c) Results of SF-MPPCA algorithm[43]. (d) Results of the Sparse algorithm[1]. (e) Results of the proposed algorithm.

$$\boldsymbol{P}_1:\ \boldsymbol{H}_1^M \rightarrow \boldsymbol{A}_k,\quad \boldsymbol{P}_2:\ \boldsymbol{H}_2^D \rightarrow \boldsymbol{A}_k,$$

where $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are linear projections. To build the correlation between different feature descriptors, we require different feature descriptors of each sample to be equal in $\boldsymbol{A}_k$:

$$\boldsymbol{P}_1\boldsymbol{T}_1(\boldsymbol{x}_i) = \boldsymbol{P}_2\boldsymbol{T}_2(\boldsymbol{y}_i),\quad \forall i.$$

3.3 Collective Sparse Representation for Anomaly Detection

Nowadays, sparse representation based methods are proposed to deal with the abnormal event detection. For sparse representation based methods, sparse reconstruction analysis of a given sample is exploited as a novel and promising idea in abnormal behavior detection. The fundamental underlying assumption of these methods is that any new sample can be represented by a linear combination of a few training samples from an overcomplete dictionary. Given an input test sample $\boldsymbol{y} \in \mathbb{R}^m$, it can be reconstructed by a sparse linear combination of an overcomplete normal base $\boldsymbol{D} \in \mathbb{R}^{m\times n}$ as follows:

$$\boldsymbol{s}^* = \arg\min_{\boldsymbol{s}} \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{s}\|_2^2 + \lambda\|\boldsymbol{s}\|_1, \tag{1}$$

where $\boldsymbol{s}$ is the representation coefficients. In anomaly event detection, a normal event $\boldsymbol{y}$ is more likely to generate sparse representation coefficients with respect to an overcomplete dictionary $\boldsymbol{D}$, but an abnormal event is not, thus generating a dense representation. To quantify the normalness, the standard sparse reconstruction cost (SRC) with l_1 regularization is described as follows:

$$SRC = \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{s}^*\|_2^2 + \lambda\|\boldsymbol{s}^*\|_1.$$

Generally, the SRC can be exploited as a measurement to identify anomalies. That is, the reconstruction costs of the normal frames are significantly higher than those of the abnormal frames. The model in (1) is generally utilized to improve the accuracy of abnormal event detection by combining different features, and fails to consider latent relation cross different features. In this case, it may lead to undesirable results by exploiting the combination or concatenation of different features. To address the problem, we exploit sparse representation to capture the salient structures of different descriptors, and generate the same representation with respect to different dictionaries. Then the proposed method builds the intrinsic relation between different feature descriptors and captures high level semantic information. We define the following objective function:

$$\|\boldsymbol{y}_1 - \boldsymbol{D}_1\boldsymbol{S}\|_F^2 + \|\boldsymbol{y}_2 - \boldsymbol{D}_2\boldsymbol{S}\|_F^2 + \lambda\|\boldsymbol{S}\|_1, \tag{2}$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ denote HOG feature and MHOF feature, respectively, $\boldsymbol{S}$ is the shared representation matrix for $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ represent the corresponding dictionaries, and λ is regularization parameter. The first term and the second term in (2) are the reconstruction errors. For a feature pair $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$, these terms should be small. This is because features from a usual event are more likely to be reconstructible from dictionaries $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, which agrees with our definition of usual events. The third term is the sparsity constraint. It can be seen from (2) that each column vector of $\boldsymbol{D}_1$ or $\boldsymbol{D}_1$ captures the high-level information and each column vector of $\boldsymbol{S}$ is the corresponding representation. By considering (2), there exists the linear map between the dictionary $\boldsymbol{D}_1$ of HOG feature $\boldsymbol{y}_1$ and the dictionary $\boldsymbol{D}_2$ of MHOF feature $\boldsymbol{y}_2$ by left multiplication inverse of $\boldsymbol{P}_1\boldsymbol{T}_1$:

$$\boldsymbol{D}_1 = (\boldsymbol{P}_1\boldsymbol{T}_1)^{-1}\boldsymbol{P}_2\boldsymbol{T}_2\boldsymbol{D}_2 = \boldsymbol{P}\boldsymbol{D}_2, \tag{3}$$

where $\boldsymbol{P} = (\boldsymbol{P}_1\boldsymbol{T}_1)^{-1}\boldsymbol{P}_2\boldsymbol{T}_2$ is the linear projection. We can approximate (3) by optimizing the cross-correlation between different features:

$$\|\boldsymbol{D}_1 - \boldsymbol{P}\boldsymbol{D}_2\|_F^2. \tag{4}$$

The overall objective function, combining the sparse coding on different features given in (3) and the cross-correlation between different features given in (4), is obtained as follows.

$$\begin{aligned}\min_{\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{S},\boldsymbol{P}} & \|\boldsymbol{y}_1 - \boldsymbol{D}_1\boldsymbol{S}\|_F^2 + \|\boldsymbol{y}_2 - \boldsymbol{D}_2\boldsymbol{S}\|_F^2 + \\ & \lambda\|\boldsymbol{S}\|_1 + \mu\|\boldsymbol{D}_1 - \boldsymbol{P}\boldsymbol{D}_2\|_F^2 + \gamma R(\boldsymbol{P}) \\ \text{s.t.}\ & \left\|\boldsymbol{d}_1^i\right\|_2^2 \leqslant 1,\ \left\|\boldsymbol{d}_2^i\right\|_2^2 \leqslant 1,\ i = 1,2,\cdots,T,\end{aligned} \tag{5}$$

where the regularization term is defined as $R(\cdot) = \|\cdot\|_F^2$ to avoid overfitting, and λ, μ and γ are regularization parameters. μ is to enforce the latent relation between different features by building the relationship between their corresponding dictionaries.

4 Optimization Algorithm

(5) is non-convex with four variables $\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{P}, \boldsymbol{S}$. Fortunately, it is convex with respect to any one of random three variables while fixing the other three ones. Thus, the optimization problem in (5) can be solved by the following steps until convergency.

Step 1. Learning sparse representation $\boldsymbol{S}$ by fixing other variables, (5) can be formulated as follows.

$$\begin{aligned}&\min_{\boldsymbol{S}} \|\boldsymbol{y}_1-\boldsymbol{D}_1\boldsymbol{S}\|_F^2+\|\boldsymbol{y}_2-\boldsymbol{D}_2\boldsymbol{S}\|_F^2+\lambda\|\boldsymbol{S}\|_1\\ \Leftrightarrow\ &\min_{\boldsymbol{S}}\left\|\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix}-\begin{pmatrix}\boldsymbol{D}_1\\ \boldsymbol{D}_2\end{pmatrix}\boldsymbol{S}\right\|_F^2+\lambda\|\boldsymbol{S}\|_1.\end{aligned}\tag{6}$$

(6) can be solved by using SLEP (sparse learning with efficient projections) package[1].

Step 2. Learning dictionary $\boldsymbol{D}_1$ by fixing other variables, (5) can be obtained as:

$$\min_{\boldsymbol{D}_1}\|\boldsymbol{y}_1-\boldsymbol{D}_1\boldsymbol{S}\|_F^2+\mu\left\|\boldsymbol{D}_1-\boldsymbol{P}\boldsymbol{D}_2\right\|_F^2.\tag{7}$$

(7) can be transformed as:

$$\begin{aligned}&\min_{\boldsymbol{D}_1}\left\|\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{P}\boldsymbol{D}_2\end{pmatrix}-\boldsymbol{D}_1\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}\right\|_F^2\\ &\text{s.t.}\quad \left\|\boldsymbol{d}_1^i\right\|_2^2\leqslant 1,\quad i=1,2,\cdots,T.\end{aligned}\tag{8}$$

(8) can be obtained when considering the Lagrangian:

$$\begin{aligned}L\left(\boldsymbol{D}_1,\boldsymbol{\lambda}\right)=&\min_{\boldsymbol{D}_1}\left\|\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{P}\boldsymbol{D}_2\end{pmatrix}-\boldsymbol{D}_1\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}\right\|_F^2\\ &\text{s.t. }\left\|\boldsymbol{d}_1^i\right\|_2^2\leqslant 1,\quad i=1,2,\cdots,T\\ =&\ \mathit{trace}\left(\left(\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{P}\boldsymbol{D}_2\end{pmatrix}-\boldsymbol{D}_1\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}\right)^{\mathrm{T}}\right.\\ &\left.\left(\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{P}\boldsymbol{D}_2\end{pmatrix}-\boldsymbol{D}_1\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}\right)\right)+\\ &\sum_{j=1}^{T}\omega_j\left(\sum_{i=1}^{k}\boldsymbol{d}_1^{ij}-1\right),\end{aligned}$$

where *trace* means to calculate the sum of the elements on the main diagonal of a matrix. The gradient and Hessian of $L\left(\boldsymbol{D}_1,\boldsymbol{\lambda}\right)$ are computed as follows.

$$\begin{aligned}&\frac{\partial L\left(\boldsymbol{D}_1,\boldsymbol{\lambda}\right)}{\partial\boldsymbol{\lambda}_i}\\ =&\left\|\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{P}\boldsymbol{D}_2\end{pmatrix}\boldsymbol{U}^{\mathrm{T}}\left(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}+\mathrm{diag}(\boldsymbol{\lambda})^{-1}\right)^{-1}\boldsymbol{e}_i\right\|^2-1,\\ &\frac{\partial L\left(\boldsymbol{D}_1,\boldsymbol{\lambda}\right)}{\partial\boldsymbol{\lambda}_i\partial\boldsymbol{\lambda}_j}\\ =&-2\left(\left(\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}^{\mathrm{T}}+\mathrm{diag}(\boldsymbol{\lambda})^{-1}\right)^{-1}\right.\\ &\left(\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{P}\boldsymbol{D}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{P}\boldsymbol{D}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}^{\mathrm{T}}\\ &\left.\left(\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}^{\mathrm{T}}+\mathrm{diag}(\boldsymbol{\lambda})^{-1}\right)^{-1}\right)_{ij}\times\\ &\left(\left(\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}^{\mathrm{T}}+\mathrm{diag}(\boldsymbol{\lambda})^{-1}\right)^{-1}\right)_{i,j},\end{aligned}\tag{9}$$

where $\boldsymbol{U}=\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}^{\mathrm{T}}$. After maximizing $L\left(\boldsymbol{D}_1,\boldsymbol{\lambda}\right)$, we can obtain $\boldsymbol{D}_1$ as follows.

$$\begin{aligned}\boldsymbol{D}_1=&\left(\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}^{\mathrm{T}}+\mathrm{diag}(\boldsymbol{\lambda})^{-1}\right)^{-1}\\ &\left(\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{P}\boldsymbol{D}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{S}\\ \sqrt{\mu}\boldsymbol{I}\end{pmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}.\end{aligned}$$

Step 3. Learning dictionary $\boldsymbol{D}_2$ by fixing other variables, (5) can be obtained as:

$$\begin{aligned}&\min_{\boldsymbol{D}_2}\|\boldsymbol{y}_2-\boldsymbol{D}_2\boldsymbol{S}\|_F^2+\mu\left\|\boldsymbol{D}_1-\boldsymbol{P}\boldsymbol{D}_2\right\|_F^2\\ &\text{s.t.}\quad\left\|\boldsymbol{d}_2^i\right\|_2^2\leqslant 1,\quad i=1,2,\cdots,T.\end{aligned}\tag{10}$$

The optimal solution $\boldsymbol{D}_2$ in (10) can be obtained by using gradient descent with iterative projection[49].

Step 4. Obtaining $\boldsymbol{P}$ by fixing other variables, (5) can be obtained as:

$$\mu\left\|\boldsymbol{D}_1-\boldsymbol{P}\boldsymbol{D}_2\right\|_F^2+\gamma\left\|\boldsymbol{P}\right\|_F^2.\tag{11}$$

By taking the derivative of (11) with respect to $\boldsymbol{P}$ and setting it to 0, we can obtain the close form solution:

$$\boldsymbol{P}=\frac{\mu\boldsymbol{D}_1\boldsymbol{D}_2^{\mathrm{T}}}{\left(\mu\boldsymbol{D}_2\boldsymbol{D}_2^{\mathrm{T}}+\gamma\boldsymbol{I}\right)}.$$

The overall optimization algorithm is summarized in Algorithm 1.

Algorithm 1. Optimization Algorithm of Objective Function

```
Input: HOG feature $y_1$ and HOF feature $y_2$ of training sample
Initialize step: Initialize $D_1$, $D_2$ and $P$ by random matrices respectively, iterate number $L$
Repeat
    1. Sparse representation $S$: optimizing for $S$:
       Fixing the variables $D_1$, $D_2$ and $P$, sparse representation $S$ is optimized by exploiting step 1
    2. Dictionary $D_1$: optimizing for $D_1$:
       Fixing the variables $S$, $D_2$ and $P$, update the dictionary $D_1$ by (9)
    3. Variable $D_2$: optimizing for $D_2$:
       Fixing variables $S$, $D_1$ and $P$, update the variable $D_2$ as illustrated in step 3
    4. Variable $P$: optimizing for $P$:
       Fixing the variables $S$, $D_2$ and $D_1$, update the variable $P$ as illustrated in step 4
until convergency
Output: the estimated variables $D_1$, $D_2$ and $S$
```

[1] http://parnec.nuaa.edu.cn/jliu/largeScaleSparseLearning.htm, Mar. 2017.

5 Abnormal Event Detection

In this subsection, details about determining whether a testing sample $\boldsymbol{Y}$ is normal or not are discussed. Similar to [2], $\boldsymbol{Y}$ is detected as an abnormal event if the cost function is beyond a given threshold. Mathematically, the cost function of $\boldsymbol{Y}$ is

$$J(\boldsymbol{Y}) = \|\boldsymbol{Y}_1 - \boldsymbol{D}_1\boldsymbol{S}\|_F^2 + \|\boldsymbol{Y}_2 - \boldsymbol{D}_2\boldsymbol{S}\|_F^2 + \lambda\|\boldsymbol{S}\|_1 + \\ \mu\|\boldsymbol{D}_1 - \boldsymbol{P}\boldsymbol{D}_2\|_F^2 + \gamma R(\boldsymbol{P}),$$

where $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ represent the HOG feature and the HOF feature of test sample $\boldsymbol{Y}$ respectively. The abnormal event detection framework is listed in Algorithm 2.

Algorithm 2. Abnormal Event Detection Framework

```
Input: HOG feature $\boldsymbol{y'}_1$ and HOF feature $\boldsymbol{y'}_2$ of testing sample $\boldsymbol{y'}$
Pursuit the sparse representation $\boldsymbol{S}$ by minimizing
    $\boldsymbol{S'} = \arg\min_{\boldsymbol{S}} J(\boldsymbol{S}) = \arg\min_{\boldsymbol{S}} \frac{1}{2}\|\boldsymbol{y}_1 - \boldsymbol{D}_1\boldsymbol{S}\|_F^2 +$
    $\|\boldsymbol{y}_2 - \boldsymbol{D}_2\boldsymbol{S}\|_F^2 + \lambda\|\boldsymbol{S}\|_1 + \mu\|\boldsymbol{D}_1 - \boldsymbol{P}\boldsymbol{D}_2\|_F^2 +$
    $\gamma\|\boldsymbol{P}\|_F^2$
if $J(\boldsymbol{S'}) > \xi$ then $\boldsymbol{y'}$ is an abnormal event
  else $\boldsymbol{y'}$ is a normal event
end if
```

6 Experiments

In this section, two real video datasets are used to verify the validation of the proposed algorithm. Results of state-of-the-art algorithms are adopted for comparison.

6.1 Evaluation Methodology

In this paper, three commonly used criteria are employed to evaluate the anomaly detection accuracy: frame-level, pixel-level, and object-level. All criteria consider the correct detections compared with the ground-truth. The ground-truth is defined as: the existing of an anomaly is a "positive", while the absence is a "negative".

- *Frame-Level.* If only a frame contains one abnormal pixel, it is marked as an abnormal frame. If the frame-level ground-truth of the corresponding frame is abnormal, it is a true positive. Otherwise, it is a false positive. Although it is easy to understand the frame-level measurement, it does not identify whether the detected anomaly is truly abnormal or not. Based on this, some detected true positive frames may be co-occurrences of true abnormalities and false detections.

- *Pixel-Level.* A detected abnormal frame is true positive if and only if at least 40% of the true abnormal pixels are detected, compared with the pixel-level ground-truth. It is a false positive if any pixel in a negative frame is detected to be abnormal. Compared with frame-level measurement, pixel-level measurement emphasizes more about the detection of truly abnormal pixels.

- *Object-Level.* Although the pixel-level criterion seems pretty good, it contains plenty of falsely detected pixels. The reason is that if all pixels of an abnormal frame are detected as abnormalities, more than 40% of truly abnormal pixels must be detected. Object-level measurement identifies the frames, in which

$$\frac{\text{detected abnormality} \cap \text{true abnormality}}{\text{detected abnormality} \cup \text{true abnormality}} \geqslant Thr, \tag{12}$$

where Thr is a given threshold. It is easy to find that object-level measurement concerns more about the accurate detection of the truly abnormal events.

For both the frame-level and the pixel-level criteria, the receiver operating characteristic (ROC) curve is utilized to measure the detection accuracy. ROC is a curve of true positive rate (TPR) vs false positive rate (FPR):

$$TPR = \frac{\text{number of true positive}}{\text{number of positive}},$$
$$FPR = \frac{\text{number of false positive}}{\text{number of negative}}.$$

Based on the ROC curve, there are three evaluation criteria:

- *Area Under Curve* (AUC): the area under the ROC curve,
- *Equal Error Rate* (EER): the ratio of misclassified frames at which $FPR = 1 - TPR$, and
- *Equal Detected Rate* (EDR): the detected rate at EER, i.e., $EDR = 1 - EER$.

6.2 UCSD Pedestrian Dataset

The UCSD[2] dataset is a dataset recorded on the UCSD campus, overlooking the pedestrian walkways.

[2] http://www.svcl.ucsd.edu/projects/anomaly/dataset.html, Mar. 2017.

The crowd density varies from sparse to extremely crowded. The normal videos contain only normal events, which are comprised with only pedestrians. Abnormal events are caused by either 1) the entities of non-pedestrians, or 2) abnormal motion patterns of pedestrians. Accurate detection on this dataset is hard for that all the abnormal events are not staged or synthesized, but occur naturally.

In this paper, UCSD Ped1 dataset is used to evaluate the performance of the proposed algorithm. There are 34 and 36 video clips in training and testing sets, respectively. Each short video clip is composed of 200 frames with spatial resolution 158×238. In this experiment, frames are divided into $15 \times 15 \times 5$ patches. For each patch, a 16-dimension modified MHOF feature is extracted to describe an event, and a 4-dimension texture and a 1-dimension size feature are extracted from the current frame to discover the relationship between events. Parameters in (5) are set as $\lambda = 0.1$, $\mu = 0.2$, and $\gamma = 0.2$. The training dataset is used to learn the structured dictionary.

Some state-of-the-art algorithms are selected as competitors, which are MDT[43], MPPCA[42], SF[7], Cong *et al.*'s algorithm[1], and Adam *et al.*'s algorithm[41]. Some visual results are shown in Fig.1, in which the abnormal events are highlighted with red masks. These abnormal events include a car, bikes, a skater, people running or walking in the grass. Fig.1(a) is the ground-truth, Fig.1(b) is generated by MDT algorithm[43], and Figs.1(c) and 1(d) are from SF-MPPCA[43] and the sparse algorithm[1], respectively. Fig.1(e) is generated by the proposed algorithm. For the MDT algorithm, the result in the third column of Fig.1 misses people walking through the grass, and is inaccurate since the foreground mask is too large. The SF-MPPCA algorithm completely misses the skater in the second column of Fig.1, person running and people walking through the grass in the third column of Fig.1, and the cyclist in the last column of Fig.1. For sparse[1] and the proposed algorithms, they all fail to detect the runner in the third column of Fig.1.

The frame-level and pixel-level criteria defined above are used for quantitative comparisons. In Fig.2 and Fig.3, ROC curves of state-of-the-art algorithms for anomaly detection are shown, including MDT[43], sparse[1], social force model[7], MPPCA[42], and Adam *et al.*'s work[41]. Fig.2 shows their frame-level performance, and Fig.3 is the pixel-level performance. It can be easily seen that for frame-level measurement, ROC curve of the proposed algorithm is above the other algorithms when false positive rate is high. For pixel-level measurement, ROC curve of the proposed algorithm outperforms those of all the others.

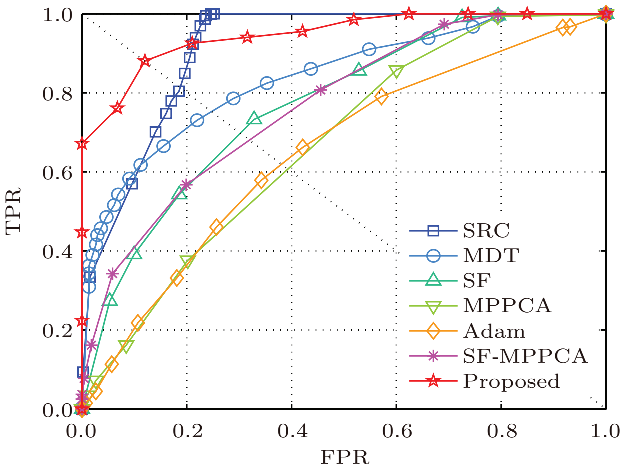


Fig.2. Frame-level ROC for Ped1 dataset.

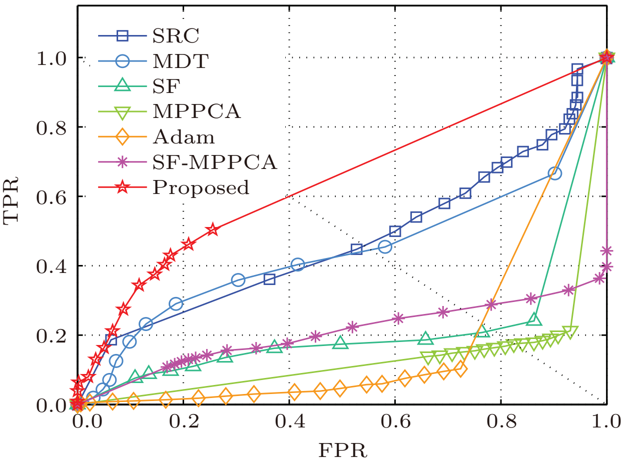


Fig.3. Pixel-level ROC for Ped1 dataset.

The evaluation criteria of the two comparisons are listed in Table 1 and Table 2, respectively. Values in these tables are from existing work. With these criteria, it can be seen that the proposed algorithm is the best in the quantitative comparison.

Table 1. Comparison of Frame-Level EER and AUC on the UCSD Ped1 Dataset

Algorithm	EER (%)	AUC (%)
MDT[43]	25.0	81.8
MPPCA[42]	40.0	67.0
SF-MPPCA[43]	32.0	76.9
SF[7]	31.0	76.8
Adam[41]	38.0	64.9
SRC[1]	19.0	86.0
Ours	12.1	94.6

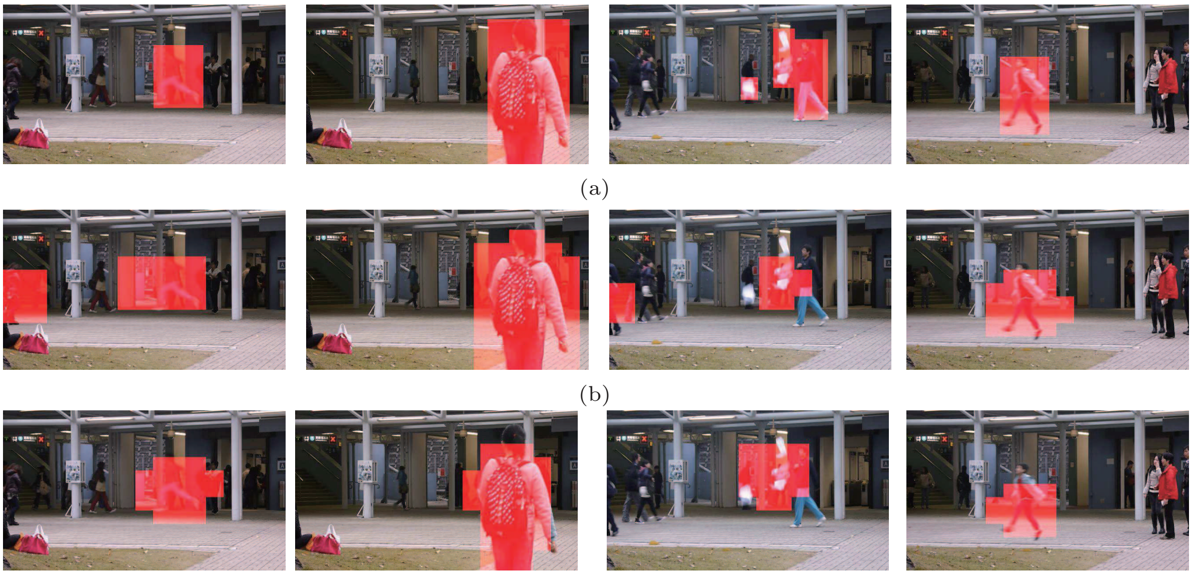
Table 2. Comparison of Pixel-Level EDR and AUC on the UCSD Ped1 Dataset

Algorithm	EDR (%)	AUC (%)
MDT[43]	45.0	44.1
MPPCA[42]	18.0	13.3
SF-MPPCA[43]	28.0	20.5
SF[7]	21.0	21.3
Adam[41]	24.0	19.7
SRC[1]	46.0	46.1
Ours	51.3	52.1

6.3 Avenue Dataset

The Avenue dataset is released by Lu *et al.*[16], containing 16 training and 21 testing video clips. Abnormal events in this dataset include running, loitering, and throwing objects. As indicated by the authors of [16]③, the main challenges are a few outliers in training videos, the slight camera shakes in testing video, and the lack of some normal patterns appearing in testing videos.

Some visual results are shown in Fig.4. For the sake of high detection rate, [16] ignores the relationships among samples, which leads to the difficulty of distinguishing normal and abnormal events by reconstruction errors. In the first column and the third column of Fig.4, some false detections are provided with Lu *et al.*'s algorithm[16]. Since only the object-level ground-truth is provided, the object-level quantitative comparison is used in this subsection. Table 3 lists the object-level measurements under varying overlap threshold *Thr* computed by (12). It is easy to find that the introduction of structural information by the proposed algorithm improves the average detection accuracy by 6.8%.


(a)

(b)

(c)

Fig.4. Examples of abnormal event detections in Avenue dataset. (a) Ground-truth. (b) Results of Lu *et al.*'s algorithm[16]. (c) Results of the proposed algorithm.

Table 3. Comparisons of Detection Accuracy on the Avenue Dataset

Thr	Lu *et al.*'s Algorithm[16] (%)	Proposed (%)
0.2	70.0	75.8
0.3	67.3	72.8
0.4	63.3	69.7
0.5	59.3	66.2
0.6	57.5	64.7
0.7	55.7	63.6
0.8	54.4	62.8

7 Conclusions

In this paper, we proposed a robust collective representation with multi-feature descriptors for abnormal event detection. The proposed method is composed of three steps. First, we extracted the distinguishable and effective features to represent video events and represent these different features in an identical representation, in which different features of the same topic

③http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html, Mar. 2017.

will show more common properties. Second, the intrinsic relation between different feature descriptors is explored and built to capture concept drift in the video sequence. Finally, the proposed collective representation based sparse coding can be exploited as a measurement to identify anomalies.

References

[1] Cong Y, Yuan J, Liu J. Sparse reconstruction cost for abnormal event detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2011, pp.3449-3456.

[2] Zhao B, Li F F, Xing E P. Online detection of unusual events in videos via dynamic sparse coding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2011, pp.3313-3320.

[3] Zhou Y, Bai X, Liu W *et al.* Swarm fusion for visual tracking. *International Journal of Computer Vision*, 2016, 118(3): 337-363.

[4] Li C, Han Z, Ye Q, Jiao J. Abnormal behavior detection via sparse reconstruction analysis of trajectory. In *Proc. the 6th International Conference on Image and Graphics*, August 2011, pp.807-810.

[5] Piciarelli C, Micheloni C, Foresti G L. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2008, 18(11): 1544-1554.

[6] Lu X, Wang Y, Yuan Y. Alternatively constrained dictionary learning for image superresolution. *IEEE Transactions on Cybernetics*, 2014, 44(3): 366-377.

[7] Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp.935-942.

[8] Lu X, Yuan Y, Zheng X. Jointly dictionary learning for change detection in multispectral imagery. *IEEE Transactions on Cybernetics*, 2017, 47(4): 884-897.

[9] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. *ACM Computing Surveys*, 2009, 41(3): 15:1-15:58.

[10] Vishwakarma S, Agrawal A. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 2013, 29(10): 983-1009.

[11] Borges P V K, Conci N, Cavallaro A. Video-based human behavior understanding: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2013, 23(11): 1993-2008.

[12] Bruckstein A, Donoho D, Elad M. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.*, 2009, 51(1): 34-81.

[13] Lu X, Wu H, Yuan Y. Double constrained NMF for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 2014, 52(5): 2746-2758.

[14] Lu X, Wang Y, Yuan Y. Graph regularized low-rank representation for destriping of hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 2013, 51(7-1): 4009-4018.

[15] Song B, Li J, Mura M D, Li P, Plaza A, Bioucas-Dias J M, Benediktsson J A, Chanussot J. Remotely sensed image classification using sparse representations of morphological attribute profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 2014, 52(8): 5122-5136.

[16] Lu C, Shi J, Jia J. Abnormal event detection at 150 FPS in MATLAB. In *Proc. IEEE International Conference on Computer Vision*, December 2013, pp.2720-2727.

[17] Mo X, Monga V, Bala R, Fan Z. Adaptive sparse representations for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014, 24(4): 631-645.

[18] Basharat A, Gritai A, Shah M. Learning object motion patterns for anomaly detection and improved object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2008.

[19] Yuan Y, Fang J, Wang Q. Online anomaly detection in crowd scenes via structure analysis. *IEEE Transactions on Cybernetics*, 2015, 45(3): 562-575.

[20] Itti L, Baldi P. A principled approach to detecting surprising events in video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2005, pp.631-637.

[21] Han J, Zhang D, Hu X, Guo L, Ren J, Wu F. Background prior-based salient object detection via deep reconstruction residual. *IEEE Trans. Circuits and Systems for Video Technology*, 2015, 25(8): 1309-1321.

[22] Han J, Zhang D, Wen S, Guo L, Liu T, Li X. Two-stage learning to predict human eye fixations via SDAEs. *IEEE Trans. Cybernetics*, 2016, 46(2): 487-498.

[23] Qi W, Cheng M, Borji A, Lu H, Bai L. SaliencyRank: Two-stage manifold ranking for salient object detection. *Computational Visual Media*, 2016, 1(4): 309-320.

[24] Cheng M, Mitra N J, Huang X, Torr P H S, Hu S. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(3): 569-582.

[25] Boiman O, Irani M. Detecting irregularities in images and in video. *International Journal of Computer Vision*, 2007, 74(1): 17-31.

[26] Kratz L, Nishino K. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp.1446-1453.

[27] Wu S, Moore B, Shah M. Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp.2054-2060.

[28] Cheng H Y, Hwang J N. Integrated video object tracking with applications in trajectory-based event detection. *Journal of Visual Communication and Image Representation*, 2011, 22(7): 673-685.

[29] Cui X, Liu Q, Gao M, Metaxas D N. Abnormal detection using interaction energy potentials. In *Proc. the 24th IEEE Conference on Computer Vision and Pattern Recognition*, June 2011, pp.3161-3167.

[30] Saligrama V, Chen Z. Video anomaly detection based on local statistical aggregates. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp.2112-2119.

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

[32] Sodemann A A, Ross M P, Borghetti B J. A review of anomaly detection in automated surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 2012, 42(6): 1257-1272.

[33] Li T, Chang H, Wang M, Ni B, Hong R, Yan S. Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2015, 25(3): 367-386.

[34] Zhong H, Shi J, Visontai M. Detecting unusual activity in video. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 2, June 27-July 2, 2004, pp.819-826.

[35] Benezeth Y, Jodoin P M, Saligrama V, Rosenberger C. Abnormal events detection based on spatio-temporal cooccurences. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp.2458-2465.

[36] del Rincon J, Lewandowski M, Nebel J C, Makris D. Generalized Laplacian eigenmaps for modeling and tracking human motions. *IEEE Transactions on Cybernetics*, 2014, 44(9): 1646-1660.

[37] Azhar F, Tjahjadi T. Significant body point labeling and tracking. *IEEE Transactions on Cybernetics*, 2014, 44(9): 1673-1685.

[38] Xie Y, Zhang W, Li C, Lin S, Qu Y, Zhang Y. Discriminative object tracking via sparse representation and online dictionary learning. *IEEE Transactions on Cybernetics*, 2014, 44(4): 539-553.

[39] Yang Y, Hu W, Xie Y, Zhang W, Zhang T. Temporal restricted visual tracking via reverse-low-rank sparse learning. *IEEE Transactions on Cybernetics*, 2016, 47(2): 485-498.

[40] Zhang Y, Chen X, Lin L, Xia C, Zou D. High-level representation sketch for video event retrieval. *Science in China Series F: Information Sciences*, 2016, 59(7): 072103.

[41] Adam A, Rivlin E, Shimshoni I, Reinitz D. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, 30(3): 555-560.

[42] Kim J, Grauman K. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp.2921-2928.

[43] Mahadevan V, Li W, Bhalodia V, Vasconcelos N. Anomaly detection in crowded scenes. In *Proc. the 23rd IEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp.1975-1981.

[44] Li W, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(1): 18-32.

[45] Cong Y, Yuan J, Liu J. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 2013, 46(7): 1851-1864.

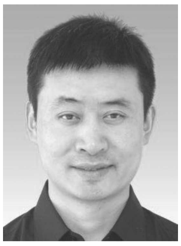
[46] Thida M, Eng H L, Remagnino P. Laplacian eigenmap with temporal constraints for local abnormality detection in crowded scenes. *IEEE Transactions on Cybernetics*, 2013, 43(6): 2147-2156.

[47] Kaltsa V, Briassouli A, Kompatsiaris I, Hadjileontiadis L J, Strintzis M G. Swarm intelligence for detecting interesting events in crowded environments. *IEEE Transactions on Image Processing*, 2015, 24(7): 2153-2166.

[48] Reddy V, Sanderson C, Lovell B C. Improved anomaly detection in crowded scenes via CellBased analysis of foreground speed, size and texture. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2011, pp.55-61.

[49] Censor Y, Zenios S. Parallel optimization: Theory, algorithms and applications. Oxford University Press, 1997.

Renzhen Ye is currently an associate professor with the Department of Mathematics, Huazhong Agricultural University, Wuhan. She is pursuing her Ph.D. degree in the Center for OPTical IMagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an. Her research interests include partial differential equations, mathematical mechanization and mathematical physics, and machine learning.



Xuelong Li is a full professor with the Center for OPTical IMagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an.