Anomaly Detection Based on Stacked Sparse Coding With Intraframe Classification Strategy

Ke Xu, Xinghao Jiang, *Member, IEEE*, and Tanfeng Sun, *Member, IEEE*

***Abstract*—Anomaly detection in videos is still a challenging task among the computer vision community. In this paper, an efficient anomaly detection method based on stacked sparse coding (SSC) with intraframe classification strategy is proposed. Each video is divided into blocks and the Foreground Interest Point (FIP) descriptor is proposed to describe the appearance and motion features for each block. The spatial-temporal features are then encoded with SSC. Specifically, the first stage of SSC encodes the spatial connections among blocks and the second stage of SSC encodes the temporal connections of all frame patches in each block. Finally, an intraframe classification strategy which uses the probabilistic outputs of SVM is proposed to evaluate the abnormality of each block. Contributions of this paper are listed as follows: 1) The FIP descriptor is proposed to describe the features of blocks, which reserves more spatial-temporal information. 2) The SSC encoding method encodes both the spatial and temporal connections of blocks, which makes the features more representative. 3) The intraframe classification strategy keeps the evaluation consistency among blocks and it helps to improve detection performance. The proposed method is examined on four public datasets with different background complexities and resolutions: UCSD Ped1 dataset, UCSD Ped2 dataset, Avenue dataset, and Subway dataset. The results are further compared with previous approaches to confirm the effectiveness and advantages of this method.**

***Index Terms*—Anomaly detection, foreground interest points, stacked sparse coding, intraframe classification.**

I. Introduction

WITH the increasing demand of security, surveillance cameras are widely installed and a huge amount of surveillance videos are generated everyday. So the crowd behaviour analysis is needed to help people to automatically recognize the events in these videos. Anomaly detection is an important application in crowd behaviour analysis which has recently attracted the interest of vision communities. Some challenges of anomaly detection have been summarized and studied by researchers. One of the challenges is lacking clear definition of anomalies. An anomaly in one scene can be taken as a normality in another scene. Thus the anomalies are generally defined as outliers of normal distributions in training samples. Researches in this area commonly follow the line that normal patterns are first learned from training videos, and the patterns are then used to detect events deviated from them [1]. Another challenge is to detect anomalies in different scales. The objects scales in a camera will change due to the viewpoint and the distance between objects and the camera. So researchers generally divide the frames into different sub-regions and detect anomalies in each sub-region separately [2]. The existing approaches for anomaly detection can be mainly classified into three categories: (i) Trajectory based methods, (ii) Global pattern based methods, (iii) Grid pattern based methods.

In a typical trajectory based method, the crowd scene of interest is first segmented into different objects. Objects are then tracked through the video sequences and their behaviours are inferred from the extracted trajectories [3]. For example, researches in this area evaluated the abnormality of trajectories by zone based analysis [4], single-class Support Vector Machine (SVM) [5], string kernels clustering [6], spatial temporal path search [7], semantic tracking [8] or deep learning-based system [9]. This class of methods relies on the tracking of people and objects. The tracking performance will be affected by low video resolution, rapid motion or occlusion.

For the global pattern based methods, the goal is not to separately detect and track individuals in a scene. Instead, these methods try to extract low or medium level features from the video in order to analyze the sequence as a whole entity [10]. Typical features used in these approaches are spatial temporal gradients and optical flow. Typical models used in these approaches are Social Force Model (SFM) [11], Gaussian mixture model (GMM) [12], Principal Component Analysis (PCA) model [13], energy model [14], salient motion map [15], global motion map [16], Gaussian regression [17], stationary map [18], motion influence map [19], etc. This type of work is effective in dealing with crowded groups. But it is not easy to locate the position of anomalies in a global pattern.

The grid pattern based methods do not take the frame as an entirety. These methods attempt to split frames into blocks and focus on the patterns in the blocks separately [20]. The grid patterns are often evaluated by Sparse Coding (SC) [21], local features probabilistic framework [22], mixtures of dynamic textures with GMM [23], spatial and temporal anomaly maps [24], Low-rank and Sparse Decomposition (LSD) [25], joint sparsity model [26], cell-based analysis of texture [27] and deep learning

Manuscript received April 14, 2017; revised December 11, 2017 and February 13, 2018; accepted March 19, 2018. Date of publication March 23, 2018; date of current version April 17, 2018. This work was supported in part by the Nature Science Foundation of China under Grant 61572321 and Grant 61572320. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Balakrishnan Prabhakaran. (*Corresponding author: Xinghao Jiang*).

The authors are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: l13025816@sjtu.edu.cn; xhjiang@sjtu.edu.cn; tfsun@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2018.2818942

network [28]–[30], etc. This type of approaches has less time cost comparing to the other types because the abnormality of each sub-region is evaluated separately and the connections of objects are not considered.

The above existing methods encounter some common problems in anomaly detection area. 1) The methodology in the anomaly detection domain is incomplete so that a single method can not detect all kinds of the anomalies. 2) Most of the existing algorithms have high computational cost and high time cost. So the methods can not satisfy the demand of real world. 3) Since anomaly detection is a relatively new research field, the public datasets are insufficient.

The work of this paper mainly focus on the problem of incompleteness of methodology. To enrich the methodologies in anomaly detection, we make effort to propose novel methods on feature extraction, normal pattern representation and anomaly judgement. For feature extraction, FIPs descriptor is proposed to extract more useful information on foreground. For normal pattern representation, SSC encoding method is proposed to extend the idea of sparse coding to temporal domain. Previous methods mainly focus on increasing spatial scales or increasing encoding depth to improve sparse coding performance. But little work is considered on encoding both the spatial connections of different blocks and the temporal connections of different frames. SSC method utilizes the connection information and generates more representative normal patterns. For anomaly judgement, the intra-frame classification strategy is proposed to transform the anomaly detection task to evaluating the probabilistic outputs of a multi-class SVM classifier, which makes the detection faster and more accurate.

The main contributions of this paper are summarized as follows: 1) An FIPs descriptor is proposed to describe the foreground spatial-temporal features of sub-regions, which makes the detection and localization results more accurate. 2) The proposed SSC encoding method encodes the connections of FIPs in frames in the first stage and encodes the connections of sub-regions between frames in the second stage, which makes the normal patterns more representative. 3) The proposed intra-frame classification strategy takes each sub-region in the training samples as one class. The anomaly detection task is transformed to evaluating the probabilistic outputs of a multi-class SVM classifier and it makes the detection faster and more accurate.

The rest of paper is organized as follows, Section II elaborates the related works and the ideas that inspire our method. Section III explains the FIP descriptor. Section IV presents the SSC strategy proposed in this paper. Section V introduces the use of intra-frame classification strategy in our work. Section VI presents the framework, the experiment details and the experimental analyses. The experimental results are compared with some state-of-the-art approaches. Section VII draws the conclusion.

II. Related Work

In this section, the methods that inspire our work are introduced. In recent years, anomaly detection methods based on grid pattern are popular. These methods split frames into blocks and built a model for each block separately. For example, Xu *et al.* [31] presented an approach via sparse reconstruction of dynamic textures over an overcomplete basis set. Cong *et al.* [32] proposed the motion context descriptor and evaluated the abnormality of region by searching for the best match in the training dataset. Thida *et al.* [33] used a spatial temporal Laplacian eigenmap method to extract different crowd activities from videos. In these methods, Sparse Coding (SC) has been proved to be an effective method in anomaly detection. The main assumption of SC is that normal video events can be represented as sparse linear combinations of normal patterns with small reconstruction errors, while abnormal patterns with large reconstruction errors. Although sparse coding discards some information in comparison with higher-order encoding methods like VLAD and Fisher Vector (FV) [34], it is more sensitive to anomalies than higher-order encoding methods.

To improve the performance of SC, researches tried to find better dictionary learning methods and more robust SC structures. Yu *et al.* [35] proposed the hierarchical sparse coding method for classification. Li *et al.* [24] and Lu *et al.* [21] both followed the idea of Course-to-fine, which split frames in multiscale and compute sparse representation in all scales. Lu *et al.* [36] proposed the adaptive dictionary learning to reduce the size of dictionary and generate better representations. Han *et al.* [37] proposed an online adaptive dictionary learning and weighted sparse coding method for anomaly detection. Zhao *et al.* [38] proposed online detection of unusual events in videos via dynamic sparse coding and sliding window.

These methods tried to make better representations for normal patterns, but they still have some limitations. The improvements are mainly made on increasing spatial scales or encoding depth, little work is considered on encoding both the spatial connections of different blocks and the temporal connections of different frames. Inspired by the idea of hierarchical sparse coding [35], we extend the idea to temporal domain and propose the SSC strategy. Meanwhile, the idea of [21] which determines the abnormality by comparing reconstruction errors is extended to making determination by comparing the probabilistic outputs of SVM. The probabilistic outputs of SVM map the reconstruction errors to relative probability scores. The range of reconstruction errors varies among blocks, while the relative probability scores can keep the evaluation consistency among blocks.

III. Foreground Interest Point Descriptor

In this section, FIP descriptor is proposed to describe the appearance and motion features of frames. FIPs are detected on the foreground edges densely and uniformly. It helps to remove the background noise and make the localization results accurate and clean. The visualized results are shown in Fig. 1 and the FIPs are marked with green color.

As mentioned in [12] and [24], foreground extraction is helpful to localize anomalies. Previous methods used background substraction or optical flow to extract foreground, which are not accurate and bring some deviations. Since the cameras do not move in the datasets, the RPCA method proposed by [39] is used in this paper to extract foreground for training samples.

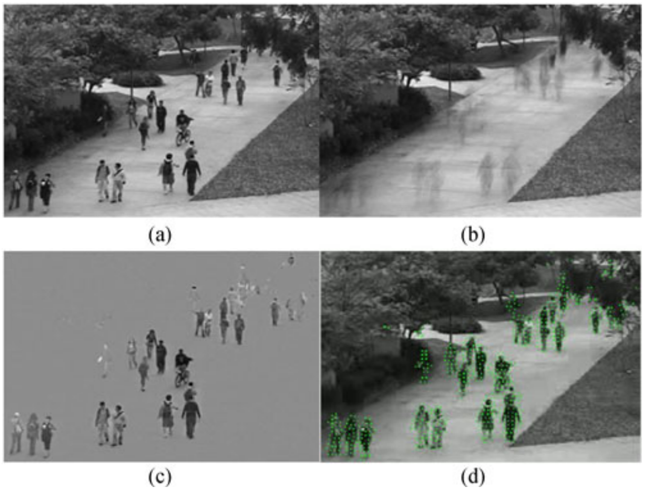

Fig. 1. Samples from UCSD Ped1 dataset. (a) is original frame, (b) and (c) are the background and foreground calculated by RPCA, (d) is the detected FIPs, which are marked with green color.

Supposing that a given video frame is rearranged to a vector and is taken as a row of a large matrix $\boldsymbol{D} \in \mathbb{R}^{m \times n}$, where m is the number of all frames and n is the number of pixels in one frame. The mathematical model for estimating the low-dimensional subspace is built to find a low rank matrix $\boldsymbol{A}$, such that the discrepancy between $\boldsymbol{A}$ and $\boldsymbol{D}$ is minimized. It leads to the following constrained optimization:

$$\min_{\boldsymbol{A},\boldsymbol{E}} \|\boldsymbol{E}\|_F \qquad \text{s.t.} \quad \text{rank}(\boldsymbol{A}) \leq r, \boldsymbol{D} = \boldsymbol{A} + \boldsymbol{E} \tag{1}$$

where $r \ll \min(m, n)$ is the target dimension of the subspace. $\|\cdot\|_F$ is the Frobenius norm, which corresponds to assuming that the data are corrupted by i.i.d. Gaussian noise. The estimated low rank matrix $\mathbf{A}$ is considered as the background matrix and the sparse matrix $\boldsymbol{E}$ is considered as the foreground matrix. The optimization is solved by the Augmented Lagrange Multiplier (ALM) Method.

Then for the i-th frame, the foreground map $\boldsymbol{E_i}$ is rearranged to the original frame size. The gradients of foreground $L_x = \partial_x(\boldsymbol{E_i})$ and $L_y = \partial_y(\boldsymbol{E_i})$ are computed. The FIPs are defined as follows:

$$FIP(x, y) = \left\{\forall p(x, y) \in \boldsymbol{E_i} \mid \sqrt{L_x^2 + L_y^2} > C_0\right\} \tag{2}$$

where $p(x, y)$ is the pixels in $\boldsymbol{E_i}$, C_0 is a constant threshold. In order to obtain the segmentation maps, we fill the holes in gradients maps and find connected regions on it. Then the connected regions are segmented and marked by colors. The results of foreground and the segmentation map are shown in Fig. 2. The segmentation maps are used to make the localization results more accurate. In this paper, detecting FIPs on edges intends to excludex the all-zero feature vectors. The FIPs are also suppressed with a 3×3 window to keep the number of FIPs moderately.

After FIPs are detected, the FIP descriptor is applied to describe the features of FIPs. FIP descriptor consists of 3DSIFT descriptor [40] and HOG/HOF descriptor [41], which both describe points in spatial and temporal domain. For an FIP (x, y, t), the gradients L_x, L_y, L_t are first calculated in the surrounding cube. Then the magnitude of point $mag(x, y, t)$, the spatial orientation $\theta(x, y, t)$ and the temporal orientation $\phi(x, y, t)$ are defined as:

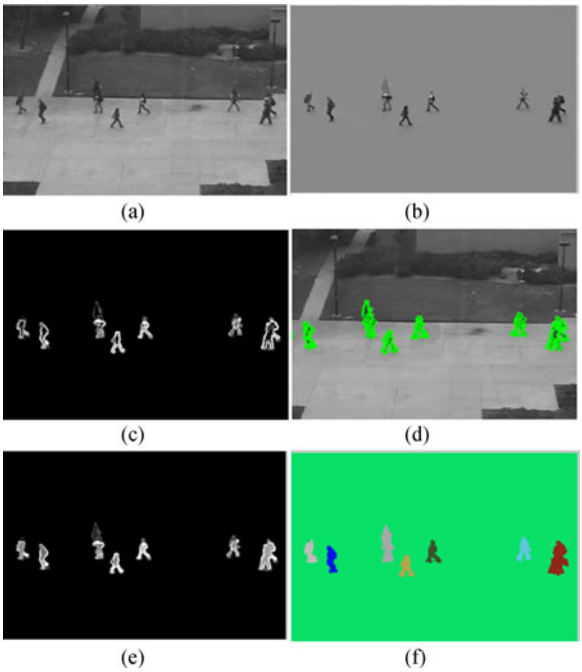

Fig. 2. Samples from UCSD Ped2 dataset. (a) is the original frame, (b) is the extracted foreground, (c) is the gradient of foreground, (d) is the FIPs in the frame, (e) is the filled gradient map, (f) is the segmentation map.

$$mag(x, y, t) = \sqrt{L_x^2 + L_y^2 + L_t^2} \tag{3}$$

$$\theta(x, y, t) = \tan^{-1}\left(\frac{L_y}{L_x}\right) \tag{4}$$

$$\phi(x, y, t) = \tan^{-1}\left(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}\right). \tag{5}$$

Then the spatial and temporal bins are split by the meridians and parallels method, which leads to the spatial bin size $\Delta\theta$ and the temporal bin size $\Delta\phi$. When computing $mag(x, y, t)$, the magnitude will also need to be normalized by the solid angle ω, which is defined as:

$$\omega = \Delta\phi(\cos\theta - \cos(\theta + \Delta\theta)). \tag{6}$$

For a point (x', y', t') in the surrounding cube, the differences between the FIP and the neighbour point $\{\Delta x, \Delta y, \Delta t\}$ are computed. The actual value added to the histogram F_{3d} is shown as below:

$$F_{3d}(\bar{\theta}, \bar{\phi}) += \frac{1}{\omega} mag(x', y', t') e^{-(\Delta x^2 + \Delta y^2 + \Delta t^2)} \tag{7}$$

where $\bar{\theta}$ and $\bar{\phi}$ are the bin indexes that θ and ϕ are located in. The orientations (θ, ϕ) of the largest $F_{3d}(\theta, \phi)$ is taken as the dominant orientation and a rotation matrix $\boldsymbol{R}$ is applied on this surrounding cube of FIP to find the rotated surrounding cube

$$\boldsymbol{R} = \begin{bmatrix} \cos\theta\cos\phi & -\sin\phi & -\cos\theta\sin\phi \\ \sin\theta\cos\phi & \cos\theta & -\sin\theta\sin\phi \\ \sin\phi & 0 & \cos\phi \end{bmatrix}. \tag{8}$$

The final histogram F_{3d} is computed on the new rotated surrounding cube with (7) to keep the rotation invariance. For the HOG descriptor, the bins are split in spatial domain with size $\Delta\theta$. θ is computed by (4) and $\bar{\theta}$ denotes the corresponding bin index. The histogram F_{hog} is calculated by Eq. (9). For the HOF descriptor, the optical flows $\{u, v\}$ between the adjacent two frames are computed first, where u denotes the horizontal optical flow map and v denotes the vertical optical flow map. The bins are split with size $\Delta\phi$. The angle $\phi = \tan^{-1}(v(x,y)/u(x,y))$ and $\bar{\phi}$ denotes the corresponding bin index. The histogram F_{hof} is calculated by (10). The descriptors are then concatenated to form the FIP descriptor, which is denoted as (11)

$$F_{hog}(\bar{\theta}) + = \sqrt{L_x^2 + L_y^2} \tag{9}$$

$$F_{hof}(\bar{\phi}) + = \sqrt{u(x,y)^2 + v(x,y)^2} \tag{10}$$

$$F_{fip} = [F_{3d}, F_{hog}, F_{hof}]. \tag{11}$$

The concatenated descriptor F_{fip} is adopted in this paper mainly because the anomalies contain abnormal objects and abnormal directions. Since F_{3d} and F_{hog} are sensitive to appearance and F_{hof} is sensitive to direction and motion intensity, F_{fip} is an effective descriptor to capture anomalies.

In this paper, following [12], the FIP descriptor is computed in a $11 \times 11 \times 3$ cube surrounding the FIPs. $\Delta\theta$ and $\Delta\phi$ are both set to $\pi/4$. The HOG has 8 bins and the HOF has 9 bins (1 bin for static points). To address the effectiveness of FIP descriptor, the comparison results between using FIP descriptor and using other descriptors are shown in the experiment section.

IV. Stacked Sparse Coding Strategy

In this section, the SSC strategy is introduced in detail. The SSC strategy contains two stages. In the first stage, the spatial connections are encoded. A point-level dictionary is trained with all FIP descriptors in training samples. The descriptors are reconstructed with sparse representations. In the second stage, the temporal connections are encoded. The Maxpooling of Sparse Coding (MSC) features for each block are computed and frame-level dictionaries are trained with all the frame patches in each block.

A. First Stage Encoding Process

In the first stage of SSC, formula (12) is used to train the point-level dictionary. In the training phase, given all the FIP descriptors $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ in training videos, a point-level dictionary $\boldsymbol{D}_p$ is learned with a sparsity prior.

$$\min_{\boldsymbol{D}_p} E(\boldsymbol{D}_p, \boldsymbol{A}) \triangleq \frac{1}{2k} \sum_{i=1}^{k} \left\{ \|\boldsymbol{x}_i - \boldsymbol{D}_p \boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1 \right\} \tag{12}$$

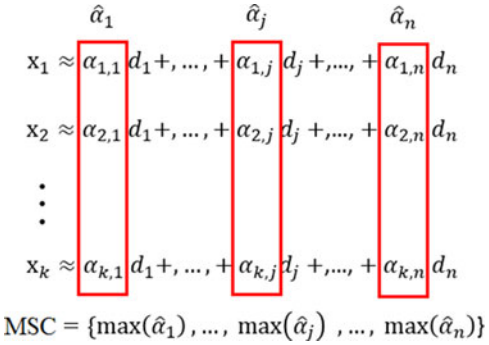
$\hat{\alpha}_1$
$\hat{\alpha}_j$
$\hat{\alpha}_n$
$x_1 \approx \alpha_{1,1} d_1 +, \ldots, + \alpha_{1,j} d_j +, \ldots, + \alpha_{1,n} d_n$
$x_2 \approx \alpha_{2,1} d_1 +, \ldots, + \alpha_{2,j} d_j +, \ldots, + \alpha_{2,n} d_n$
$x_k \approx \alpha_{k,1} d_1 +, \ldots, + \alpha_{k,j} d_j +, \ldots, + \alpha_{k,n} d_n$
MSC = $\{\max(\hat{\alpha}_1), \ldots, \max(\hat{\alpha}_j), \ldots, \max(\hat{\alpha}_n)\}$

Fig. 3. MSC is computed by maxpooling the AIVs.

where k is the number of training FIPs and λ is the balance parameter. $A = \{\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_k}\}$ is the corresponding sparse coefficients set of $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$, which is used to generate MSCs in the second stage of SSC. The formula can be solved via alternatively solving for $\boldsymbol{D}_p$ and $\boldsymbol{A}$ until they minimize both the data fitting term $\|\boldsymbol{x}_i - \boldsymbol{D}_p\boldsymbol{\alpha}_i\|_2^2$ and the regularization term $\|\boldsymbol{\alpha}_i\|_1$. L1-norm is used for sparsity constraint. Since it is possible to make the sparsity penalty arbitrarily small by scaling down $\boldsymbol{\alpha}_i$ and scaling $\boldsymbol{D}_p$ up by some large constant, dictionary $\boldsymbol{D}_p$ need to be restricted to a closed convex set as in (13) following the setting in [42]:

$$\boldsymbol{D}_p \triangleq \{\boldsymbol{D}_p \in \mathbb{R}^{m \times n} \quad \text{s.t.}\ \forall j = 1, \ldots, k, \boldsymbol{d}_j^T \boldsymbol{d}_j \leq 1\} \tag{13}$$

where m is the FIP descriptor length and n is the dictionary size. $\boldsymbol{d}_j$ denotes the j-th column of $\boldsymbol{D}_p$.

The size of dictionary n is a free parameter. In order to get a compact dictionary, $\boldsymbol{D}_p$ has to be an over-complete dictionary and the dictionary size n is supposed to be as small as possible at the same time. So the scale penalty is added to the formula as [36]. The new object function is:

$$\min_{\boldsymbol{D_p}} \boldsymbol{E}(\boldsymbol{D}_p, \boldsymbol{A}) + \mu \sum_{j=1}^{k} \boldsymbol{I}(\hat{\boldsymbol{\alpha}}_j) \tag{14}$$

where $\sum_{j=1}^{k} \boldsymbol{I}(\hat{\boldsymbol{\alpha}}_j)$ imposes dictionary scale penalty and μ is a balance parameter. The indicator function is defined as:

$$\boldsymbol{I}(\hat{\boldsymbol{\alpha}}_j) = \begin{cases} 0 & \text{if} \quad \|\hat{\boldsymbol{\alpha}}_{\boldsymbol{j}}\|_1 \leq \xi \\ 1 & \text{otherwise} \end{cases} \tag{15}$$

where ξ is a constrain parameter with a small value close to zero. As shown in Fig. 3, $\hat{\boldsymbol{\alpha}}_{\boldsymbol{j}} = \{\alpha_{1,j}, \ldots, \alpha_{k,j}\} \quad (1 \leq j \leq n)$ is defined as the Atom Indicator Vectors (AIVs) where $\alpha_{i,j}$ is the j-th element of α_i. $\boldsymbol{I}(\hat{\boldsymbol{\alpha}}_j)$ outputs 1 for non-zero vectors. Hence, the sum of the indicator functions for all AIV elements, expressed as $\sum_{j=1}^{k} \boldsymbol{I}(\hat{\boldsymbol{\alpha}}_j)$ can represent the number of the atoms that are indeed used. The objective function (14) can control the number of non-zero AIVs in optimization and it can be solved by the scale adaptive dictionary learning method in [36].

In this stage, the initial size of dictionary is 1000. After solving the object function (14), the size of dictionary is reduced by around 30% as in Table I. It shows that the method makes the dictionaries more compact and leads to a reduction in computational cost of reconstruction step.

TABLE I
DICTIONARY SIZES WITH SCALE PENALTY

Initial	UCSD Ped1	UCSD Ped2	Avenue	Subway
1000	677	535	762	730

In the reconstruction phase, the FIP descriptors $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ are reconstructed by the point-level dictionary $\boldsymbol{D}_p$ by

$$\min_{\boldsymbol{\beta}_i} \|\boldsymbol{x}_i - \boldsymbol{D}_p \boldsymbol{\beta}_i\|_2^2 \qquad \text{s.t.} \ \|\boldsymbol{\beta}_i\|_1 \leq T_p \tag{16}$$

where $\boldsymbol{\beta}_i$ is the sparse coefficients of $\boldsymbol{x}_i$. T_p is the sparsity upper bound which is set to 50 in order to encode more spatial connections in the dictionary. The equation is solved by LARS algorithm. The coefficients set $\boldsymbol{\beta}$ is further used for encoding in the second stage of SSC.

In the first stage encoding process, a point-level dictionary is trained with FIPs descriptors of all the blocks in training samples. The words in the point-level dictionary contain the features in different locations, so the dictionary contains the connections of FIPs descriptors. Each descriptor is reconstructed by the words in the dictionary, which means the reconstructed output vectors encode the spatial connections of frames.

B. Second Stage Encoding Process

In the second stage of SSC, the coefficients set obtained in the first stage is processed. Specifically, for the i-th block in the t-th frame, the FIPs located in this block are described and the features set is denoted as $X_{i,t} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$, where $\boldsymbol{x}_j, j \in [1, k]$ is an FIP descriptor and k is the number of FIPs. The sparse representations set $A_{i,t} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k\}$ is computed by dictionary reconstruction. As shown in Fig. 3, in a block, k FIP descriptors are reconstructed by the dictionary. The dictionary size is n. The reconstruction coefficients form a $k \times n$ matrix. The n-th column of the matrix is denoted as $\hat{\boldsymbol{\alpha}}_n$. The column set is denoted as $AIV_{i,t} = \{\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_n\}$. By maxpooling the AIVs, the MSC feature is taken as the block descriptor, which is denoted as $h_{i,t}$:

$$h_{i,t} = \{\max(\hat{\boldsymbol{\alpha}}_1), \ldots, \max(\hat{\boldsymbol{\alpha}}_n)\}. \tag{17}$$

In each block, for all the frames, the MSC features are used to build the frame-level representation for events. In the training phase, supposing the entire number of training frames is t and a frame is split into p blocks, the i-th block's MSC set $H_i = \{\boldsymbol{h}_{i,1}, \ldots \boldsymbol{h}_{i,t}\}$ is used to train a block dictionary $\boldsymbol{D}_{bi}$ with formula (18), $i \in [1, p]$. $\boldsymbol{D}_{bi}$ integrates the temporal connections of all training frames in the i-th block

$$\min_{\boldsymbol{D}_{bi}} \|H_i - \boldsymbol{D}_{bi} S_i\|_2^2 + \lambda \|S_i\|_1 \tag{18}$$

where $\boldsymbol{D}_{bi} \in \mathbb{R}^{n \times q}$, n is the size of point dictionary $\boldsymbol{D}_p$, q is the size of block dictionary $\boldsymbol{D}_{bi}$. $S_i = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_t\}$ is the sparse coefficients set in the i-th block of all t frames.

Totally p dictionaries are generated after training. Same as in the first stage of SSC process, the dictionary scale penalty is added to each block dictionary separately to constrain the size of $\boldsymbol{D}_{bi}$. So the dictionary size of each block will be smaller than the initial size. The training set of all blocks in the second stage is denoted as $H = \{H_1, \ldots H_p\}$, where $H_i, i \in [1, p]$ is the MSCs set of block i. The corresponding representations set $S = \{S_1, \ldots, S_p\}$ will be further used in training SVM. The block indexes are taken as the class labels.

In the testing phase, the MSCs set of the i-th block H'_i with $i \in [1, p]$ is calculated. H'_i is reconstructed by the corresponding $\boldsymbol{D}_{bi}$ following formula (19). For a testing frame t, the testing set is $H' = \{H'_1, \ldots H'_p\}$. The i-th block representation set is also computed by (17) and is denoted as $H'_i = \boldsymbol{h}'_{i,t}$. Reconstructing H' with dictionaries set $D_b = \{\boldsymbol{D}_{b1}, \ldots, \boldsymbol{D}_{bp}\}$ is expressed as

$$\min_{S'_i} \|H'_i - \boldsymbol{D}_{bi} S'_i\|_2^2 \qquad \text{s.t.} \ \|S'_i\|_1 \leq T_i \tag{19}$$

where $i \in [1, p]$, $S'_i \in \mathbb{R}^{q \times t}$ contains sparse coefficients. q is the size of block dictionary $\boldsymbol{D}_{bi}$. t is the frame number in the testing samples. $\|H'_i - \boldsymbol{D}_{bi} S'_i\|_2^2$ is the data fitting term and $\|S'_i\|_1$ is the sparsity regularization term. T_i is the upper bound parameter to control sparsity. The block representations set $S' = \{S'_1, \ldots, S'_p\}$ is then classified by SVM. To address the effectiveness of SSC, the comparison between using different SSC stages is given in Section VI.

V. Intra-Frame Strategy With Multi-Class SVM

In this paper, the intra-frame strategy is applied for anomaly detection. It is called intra-frame mainly because each subregion inside the frame is taken as a single class. A multi-class SVM classifier is trained to classify the sub-regions. Previous methods used clustering models such as one-class SVMs to evaluate the deviation between testing samples and the centers of normal samples. The problem of these models is that the radius in each block is different and the outliers in each block will cause false alarms. By using the multi-class SVM, the blocks inside the frame will have the max margin with each other due to the principle of SVM classifier. So the detection performance of using intra-frame strategy will be better than using previous models.

In the training phase, given the sparse representation $S = \{S_1, \ldots, S_p\}$ and the corresponding index from 1 to p, p one-versus-all SVMs are trained as a multi-class SVM classifier for modelling block patterns. Meanwhile, following the method in [43], the probabilistic outputs of SVM are used to evaluate the abnormality of each block. For a single SVM classifier, the probabilistic outputs of SVM are defined as below

$$Pr(y = 1|x) \approx P_{A,B}(f) \equiv \frac{1}{1 + exp(Af + B)} \tag{20}$$

where x is a sample feature and y is the label. $f = f(x)$ is a decision function such that $sign(f)$ can be used to predict the label of sample x. A posterior class probability $Pr(y = 1|x)$ is approximated by a sigmoid function instead of predicting the label.

Solving the regularized maximum likelihood problem in Eq. (21) and Eq. (22), the best parameter setting $z^* = (A^*, B^*)$

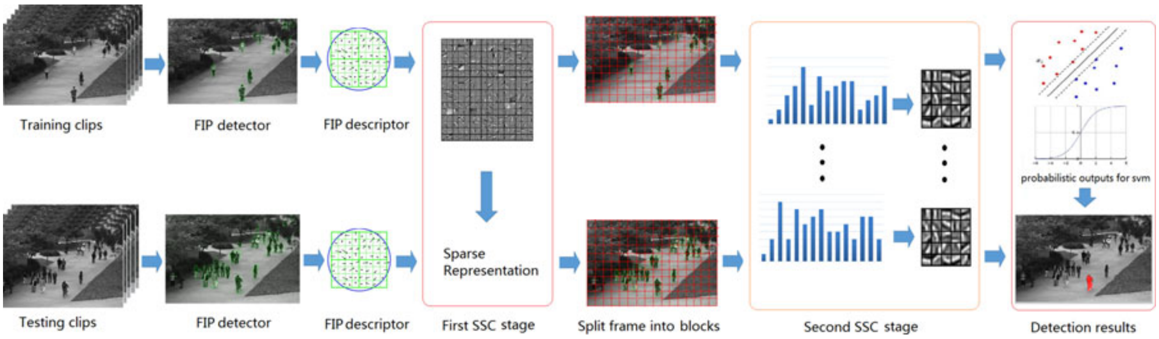

Fig. 4. The framework of our method. In the training phase, a point-level dictionary $\boldsymbol{D}_p$ is trained with FIP descriptors in the first stage of SSC. Frames are then split and the MSCs are generated to train frame-level dictionaries $\boldsymbol{D}_b$ for each block in the second stage of SSC. In the testing phase, the FIP descriptors are reconstructed by $\boldsymbol{D}_p$ in the first stage of SSC. MSCs are generated and reconstructed by the corresponding $\boldsymbol{D}_b$ in the second stage of SSC.

is determined

$$\min_{z=A,B} F(z) = -\sum_{i=1}^{N} (t_i \log(p_i) + (1-t_i)\log((1-p_i)) \tag{21}$$

$$p_i = P_{A,B}(f_i), \quad t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2} & \text{if} \quad y_i = +1 \\ \frac{1}{N_- + 2} & \text{if} \quad y_i = -1 \end{cases} \tag{22}$$

where N is the number of training samples, N_+ is the number of positive y_i, N_- is the number of negative y_i. f_i is an estimation of $f(x_i)$.

In the testing phase, instead of directly using reconstruction errors to evaluate the abnormality of event, the probabilistic outputs of SVM as in (20) are used to compute the scores of blocks. The score of a block is used to evaluate the abnormality of this block as below

$$\text{score} = 1 - Pr(y = i|x) \tag{23}$$

where $i \in [1, p]$ is the corresponding block label. If the score is higher than the threshold, the block is recognized as an abnormal block. The method improves the detection accuracy because different testing samples may have different appropriate thresholds. By using SVM classifier, the reconstruction errors are transformed to respective scores to evaluate the similarities between the testing blocks and the training blocks. Meanwhile, different from traditional supervised classifiers which need to use both normal set and abnormal set, only the blocks within normal set are used for training in this paper. To evaluate the effectiveness, comparison of using intra-frame classification strategy and reconstruction error strategy is given in Section VI.

VI. Experiments and Analysis

In this section, the framework used in the experiments, the datasets descriptions, the implement details and the experimental results are presented. The experiment results are further compared with some state-of-the-art approaches and the analyses of these comparisons are presented. In this paper, four public datasets are used to evaluate the performance of the proposed method: UCSD Ped1 dataset [23], UCSD Ped2 dataset [23], Avenue dataset [21] and Subway dataset [44]. The receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC), and the equal error rate (EER) are calculated on frame-level and pixel-level. Furthermore, the experiments to verify the effectiveness of FIP descriptor, SSC encoding and intra-frame classification strategy are introduced and analysed.

A. Framework

The framework of our method is shown in Fig. 4, which contains the steps of training and testing phases described in this paper. In Fig. 4, in the training phase, the FIPs are detected from the training samples and described by FIP descriptor. The FIP descriptors are then used to train the point-level dictionary. The FIP descriptors are encoded by SSC to generate representations for blocks. The representations are used to train the frame-level dictionaries. Then a multi-class SVM is trained for abnormality evaluation. In the testing phase, for the testing samples, the FIP descriptors are encoded and the representations of split frames are generated by SSC. Each block representation is sent to SVM classifier to evaluate the abnormality. The abnormal block is masked by the foreground map obtained in FIPs detection phase. So only the foreground pixels in the abnormal block are marked with red color.

B. Datasets Description

1) *UCSD Ped1 Dataset:* The dataset contains samples of groups of people walking towards and away from the camera. The training dataset is composed of normal events that contains only pedestrians. Abnormal events in the testing dataset are caused by either: 1) the entities of non-pedestrians (wheelchairs,carts or cars) or 2) abnormal motion patterns of pedestrians (skating or bicycling). This dataset is challenging because all abnormal events occur naturally and are not staged or synthesized. There are 34 video samples in the training sets and 36 samples in the testing sets. In the testing dataset, all samples have frame-level ground-truth labels and 10 samples are provided with pixel-level ground-truth labels. Each video sample is composed of 200 frames and the frame resolution is 158 × 238.

2) *UCSD Ped2 Dataset:* In this dataset, the training set includes 16 normal video samples and the testing set contains 12 video samples with the frame resolution 320 × 240. All testing samples are captured in the scenes with pedestrian

movement parallel to the camera and are associated with a manually-collected frame-level abnormal events annotation ground-truth list. The 12 testing samples are all provided with pixel-level ground-truth masks.

3) *Avenue Dataset:* The Avenue dataset is also recorded on the campus. The abnormal events in the Avenue dataset include running, loitering, and throwing objects, etc. There are 16 video samples for training, and 21 samples for testing. All these testing samples have object-level ground-truth labels, i.e., labelling abnormalities in spatial location with rectangle regions. This dataset has 30652 frames totally, with spatial resolution of 360 × 640.

4) *Subway Dataset:* The Subway dataset is taken from surveillance cameras at a subway station. There are two types of videos: ‘exit gate’ and ’entrance gate. Both video sequences have the resolution of 512 × 384. The exit gate video is 43 minutes long with 64901 frames in total, and video frames of the first 5 minutes are used for training. The entrance gate video is 96 min long with 144249 frames in total, and the first 15 minutes are used for training. These settings are the same as [38] and [45]. To ensure a fair comparison, we use the same definitions of abnormal events as [45]. The main abnormal events include wrong direction, loitering, no payment, and irregular interaction, etc. This dataset is provided by Adam *et al.* [44] and detection results are compared against previous methods quantitatively.

C. Implementation Details

In this section, the overall experiment settings and the evaluation criteria are explained. Referring to the block sizes in previous methods [22], [28], the block sizes in this paper are set to approximately 1/10 of the larger value between frame width and frame height. Specifically, in UCSD Ped1 dataset, the block size is set to 20×20 with 10 pixels overlapping. In UCSD Ped2 datasets, the block size is set to 30×30 with 15 pixels overlapping. In Avenue dataset and Subway dataset, the block size are both set to 60×60 with 30 pixels overlapping. The FIP descriptors in blocks are all computed in a $11 \times 11 \times 3$ cube surrounding the points referring to [12].

For all datasets, the first SSC stage dictionary size is originally set to 1000 and the sparsity constrain T_p is set to 50. The size drops around 30% with scale penalty. The coefficient of sparse regularization term λ is set to 0.15. The second SSC stage dictionary size is originally set to 500 and the sparsity constrains T_i are all set to 50. The size drops around 10% with scale penalty. The coefficient of sparse regularization term λ is also set to 0.15. Besides, in Avenue dataset and Subway dataset, the samples last long so that we sample 5 frames per second in these video samples. The FIP threshold C_0 is set to 5 and the balance parameter μ is set to 0.5. In the SVM, the parameter C is set to 300 and g is set to 0.01. A score above threshold is taken as an anomaly score and the threshold is changed to form an ROC curve. The experiments are all conducted on a laptop with an 3.2 GHz Intel i5-4570 CPU and an 8G memory.

For visualization, in a detected abnormal block, a pixel is marked with red color if it is also a foreground pixel in the RPCA foreground map. The RPCA foreground map is calculated in FIPs detection phase as described in Section III by (1).

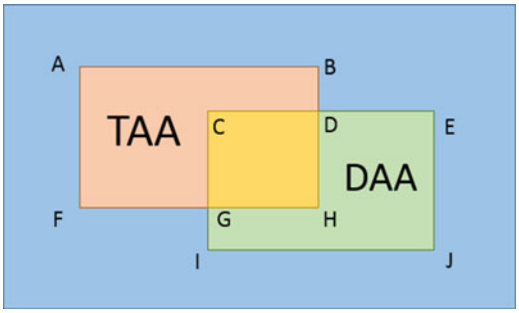

Fig. 5. Area of TAA is denoted as S_{AFHB}, area of DAA is denoted as S_{CIJE}. The object-level measurement function is $S_{CGHD}/(S_{AFHB} + S_{CIJE} - S_{CGHD})$.

In this paper, three commonly used measurements are adopted to evaluate the accuracy of abnormality detection: 1) Frame-level; 2) Pixel-level; 3) Object-level. All measurements consider the matching between the evaluated result and the ground-truth.

1) *Frame-level:* If one or more blocks are detected as abnormal blocks in a testing frame, it is labelled as an abnormal frame. If the ground-truth of this frame is abnormal, it is a True Positive (TP). Otherwise, it is a False Positive (FP). If the ground-truth frame is normal and it is labelled as abnormal, it is a True Negative (TN). Otherwise it is a False Negative (FN). In the real world surveillance anomaly detection, people want to quickly find the starting frame and ending frame of an abnormal event. So the frame-level measurement result is important and meaningful.

2) *Pixel-level:* In pixel-level measurement, following the parameters in [12] and [46], a detected abnormal frame is TP if more than 40% truly abnormal pixels are detected. A normal frame is FP as long as one pixel is detected as abnormal. Compared with frame-level measurement, pixel-level measurement emphasizes the correct detection of abnormal objects.

3) *Object-level:* Although the pixel-level measurement seems reasonable, the detected abnormal frames may contain plenty of FP pixels. Since if all pixels of an abnormal frame are detected as abnormalities, there must be more than 40% truly abnormal pixels being correctly detected. Object-level measurement aims at finding the frames which the Detected Abnormality Area (DAA) is closed to the True Abnormality Area (TAA) as shown in Fig. 5

$$\frac{DAA \bigcap TAA}{DAA \bigcup TAA} \geq \vartheta \tag{24}$$

where $\bigcap$ and $\bigcup$ stand for the intersection and union operators, respectively. ϑ is a given threshold. It is shown that the object-level measurement concerns more about the accurate detection of abnormal objects.

For both the frame-level and pixel-level measurements, the ROC curve is adopted to measure the detection accuracy. ROC is a curve of true positive rate (TPR) versus false positive rate (FPR). The equations to calculated TPR and FPR are given in

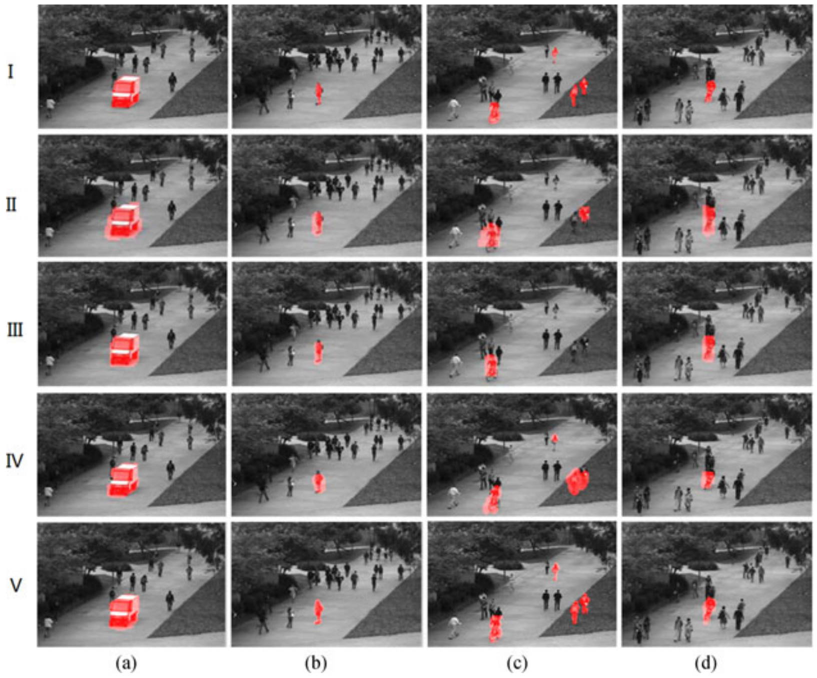

Fig. 6. Examples of anomaly detection results on the UCSD Ped1 dataset. Detected anomalies are marked with red color. (I) Ground-truth. (II) SRC [47]. (III) Cong *et al.*'s work [32]. (IV) Yuan *et al.*'s work [12]. (V) Proposed method in this paper.

(25) and (26)

$$TPR = \frac{TP}{TP + FN} \tag{25}$$

$$FPR = \frac{FP}{TN + FP}. \tag{26}$$

Based on ROC curves, three evaluation criteria are taken: 1) Area Under Curve (AUC): Area under the ROC curve. 2) Equal Error Rate (EER): The ratio of misclassified frames when the FPR equals to the miss rate, i.e., the FPR at which FPR $=$ $1 - \text{TPR}$. 3) Equal Detected Rate (EDR): The detection rate at EER, i.e., EDR $= 1 -$ EER. As EDR can be computed by EER, only the AUC and EER are listed in this paper.

D. Experiments on UCSD Ped1 Dataset

In this section, the detection results and the analyses on UCSD Ped1 dataset are given. Some visualized comparison results are shown in Fig. 6 and the detected anomalies are marked with red color. The anomalies include car, skater, biker, and so on. (I) is the ground-truth. (II) and (III) are the results of Cong *et al.*'s work, they fail to detect people who walk on grass (column c). This is because the movements are almost the same with normal pedestrians. This anomaly is detected by our method because the grass blocks have little motion features in the training set, which makes the grass blocks thresholds very small. For the Yuan *et al.*'s work in (IV), the detection of the abnormal pixels is less accurate than our method because the RPCA based segmentation map is applied in this paper.

Frame-level ROC curves are shown in Fig. 7 and pixel-level ROC curves are shown in Fig. 8. Based on these ROC curves, AUC and EER are computed and listed in Tables II and III. For the frame-level comparison in Table II, the EER of our work is 11.4% and the AUC of our work is 94.1%. For the pixel-level comparison in Table III, the EER of our work is 28.8% and the AUC of our work is 74.8%. The previous methods have some limitations. In [21] and [22], the anomaly is evaluated by reconstruction errors and the thresholds could vary in different testing samples. In Yuan *et al.*'s work [12], the blocks are simply represented by HOG/HOF feature. Our results outperform the previous methods because in our work, the FIP descriptor and the SSC encoding catch more spatial and temporal connection information than HOG/HOF, which make the block representations more robust. Meanwhile, the blocks are evaluated by probabilistic outputs of SVM. The method maps different score ranges of testing samples to the same range. The results

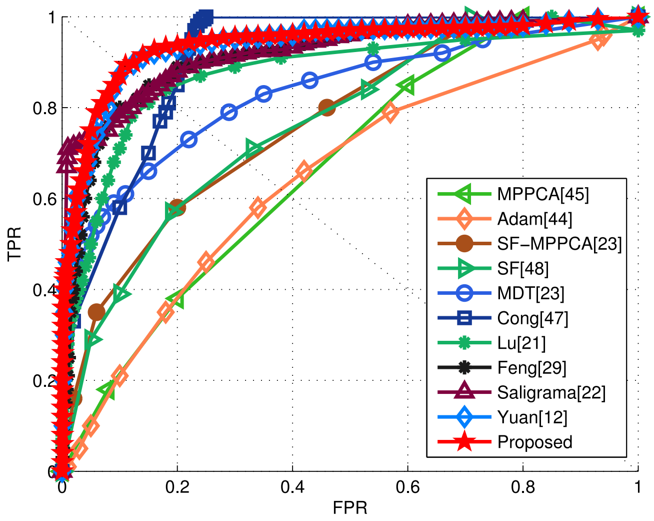

Fig. 7. Frame-level ROC curves for the UCSD Ped1 dataset.

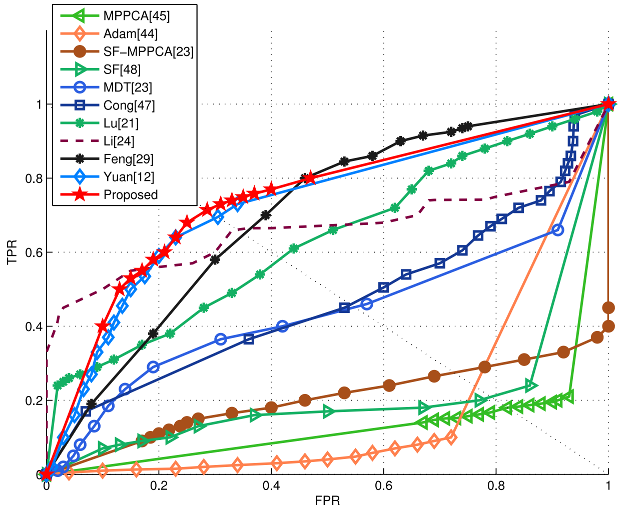

Fig. 8. Pixel-level ROC curves for the UCSD Ped1 dataset.

TABLE II
COMPARISON OF FRAME-LEVEL EER AND AUC ON THE UCSD PED1 DATASET

METHOD	EER	AUC
MPPCA [45]	40.0%	67.0%
Adam [44]	38.0%	64.9%
SF-MPPCA [23]	32.0%	76.9%
SF [48]	31.0%	76.8%
MDT [23]	25.0%	81.8%
Cong [47]	19.0%	86.0%
Lu [21]	15.0%	91.8%
Feng [29]	15.1%	92.5%
Saligrama [22]	16.0%	92.7%
Yuan [12]	12.1%	93.7%
ours	**11.4%**	**94.1%**

TABLE III
COMPARISON OF PIXEL-LEVEL EDR AND AUC ON THE UCSD PED1 DATASET

METHOD	EER	AUC
MPPCA [45]	82.0%	13.3%
Adam [44]	76.0%	19.7%
SF-MPPCA [23]	72.0%	20.5%
SF [48]	79.0%	21.3%
MDT [23]	55.0%	44.1%
Cong [47]	54.0%	46.1%
Lu [21]	40.9%	63.8%
Li [24]	35.2%	66.2%
Feng [29]	35.1%	69.9%
Yuan [12]	30.5%	73.1%
ours	**28.8%**	**74.8%**

illustrate that our algorithm has competitive advantages on both frame-level and pixel-level measurements.

E. Experiments on UCSD Ped2 Dataset

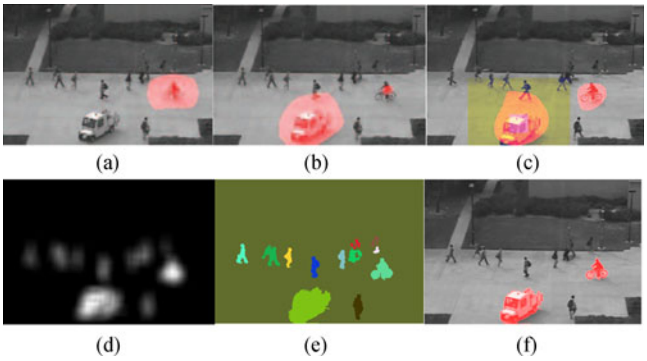
(a) (b) (c)
(d) (e) (f)

Fig. 9. Examples of anomaly detection results on the UCSD Ped2 dataset. Detected anomalies are marked with red color. (a) and (b) are from MDT [23], (c) is from Li *et al.*'s work [24], (d) is the scores map in this paper, (e) is the segmentation map. (f) is the detection result.

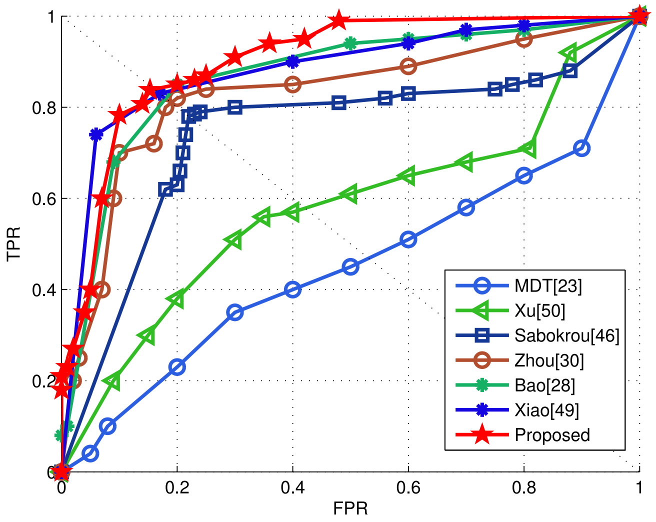

Fig. 10. Pixel-level ROC curves for the UCSD Ped2 dataset.

TABLE IV
PIXEL-LEVEL COMPARISON OF EER AND AUC ON THE UCSD PED2 DATASET

METHOD	EER	AUC
MDT [23]	54.0%	49.8%
Xu [50]	42.0%	61.9%
Sabokrou [46]	24.0%	87.1%
Zhou [30]	18.1%	88.0%
Bao [28]	18.0%	88.0%
Xiao [49]	17.0%	88.2%
ours	**16.7%**	**89.2%**

In this section, the proposed method is compared with state-of-the-art algorithms on UCSD Ped2 dataset. Some visualized comparison results are shown in Fig. 9 and the detected abnormal events are marked with red color. The areas labelled by MDT method and Li *et al.*'s work are wider than ours and contain some normal patterns. It can be inferred from Fig. 9 that our algorithm can correctly locate the anomalies and is more accurate than previous methods. As all the testing samples are provided with pixel-level ground-truth masks, only the pixel-level evaluation is taken in this paper to explain the effectiveness of our work. The pixel-level ROC curves are shown in Fig. 10 and the comparisons with other methods are shown in Table IV.

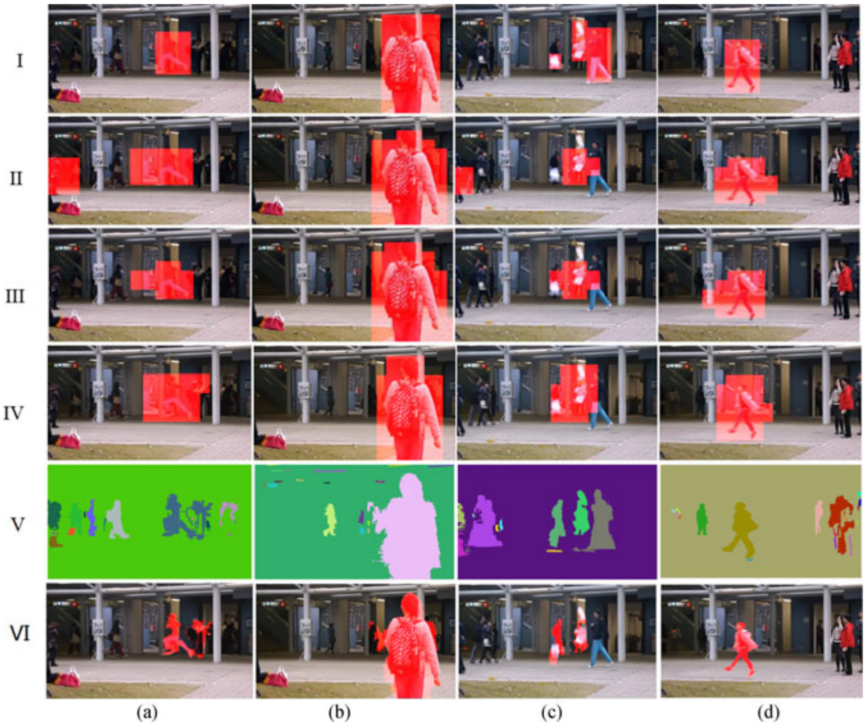

Fig. 11. Examples of anomaly detection results on the Avenue dataset. Detected anomalies are marked with red color. I is the ground-truth, II is Lu *et al.*'s work [21], III is Yuan *et al.*'s work [12]. IV is the detection results of abnormal block using our method. V is the foreground mask. VI is the masked results.

It is inferred from Table IV that our algorithm reaches an EER of 16.7% and an AUC of 89.2%, which outperforms the previous methods on UCSD Ped2 dataset and reaches a state-of-the-art. In our analysis, the proposed method performs well mainly because we use the probabilistic outputs of SVM as model. The training set of UCSD Ped2 is small. So the methods with probabilistic model [24], [49], Gaussian distributions [46] and one-class classifier [28] are easily affected by outliers and noises in normal patterns. In our intra-frame classification strategy, the multi-class SVM has advantages in small sample classification and is relatively robust to outliers.

F. Experiments on Avenue Dataset

In this section, the proposed method is compared with Lu *et al.*'s work, Yuan *et al.*'s work and Feng *et al.*'s work on the Avenue dataset. Some visualized results are shown in Fig. 11. In Fig. 11, (I) is the ground-truth. (II) is Lu *et al.*'s work [21]. (III) is Yuan *et al.*'s work [12]. (IV) is the block marked results of our method. Although the pixel-level evaluation is not required in this dataset, the segmentation maps are still given in this paper as row (V) and the masked detection results are in row (VI). It is shown that our results is more accurate than the previous methods. The detected area of Lu *et al.*'s work (II) and Yuan *et al.*'s work (II) are wider than ours in both the block marked results (IV) and the pixel marked results (VI). It is shown that the proposed algorithm can correctly detect anomalies and can locate the anomalies more accurately than previous methods.

TABLE V
COMPARISONS OF OBJECT-LEVEL DETECTION ACCURACY ON THE AVENUE DATASET

ϑ	Lu [21]	Yuan [12]	Feng [29]	ours
0.2	70.0%	74.2%	75.4%	75.5%
0.3	67.3%	70.3%	73.4%	73.7%
0.4	63.3%	66.5%	70.6%	71.2%
0.5	59.3%	64.0%	67.5%	68.1%
0.6	57.5%	62.1%	64.6%	66.8%
0.7	55.7%	61.2%	63.5%	65.3%
0.8	54.4%	60.8%	63.0%	64.1%
$\vartheta_{0.8}$-$\vartheta_{0.2}$	15.6%	13.4%	12.4%	**11.4%**

Since only the object-level ground-truth is provided by the authors, object-level measurement is adopted for quantitative comparison. Average accuracies under different value of ϑ are listed in Table V. ϑ evaluates the similarity between the detected abnormal region (DAR) and the ground-truth region (GTR), which ranges from 0 to 1. A large value of ϑ means the detection is correct only when the DAR and the GTR are nearly the same. A small value of ϑ means the detection is correct as long as the DAR covers the GTR. It is shown in Table V that our work outperforms previous work in all ϑ values. Meanwhile, the value of $\vartheta_{0.8}$ in our work is 64.1% and the $\vartheta_{0.2}$ in our work is 75.5%. Comparing the difference between $\vartheta_{0.8}$ and $\vartheta_{0.2}$, the result of Lu *et al.*'s work is 15.6%, the result of Yuan *et al.*'s work is 13.4%, the result of Feng *et al.*'s work is 12.4% and the result of our work is 11.4%. It indicates that the DAR of our method is more similar to GTR than the other methods. It also addresses that our result outperforms the previous methods.

G. Experiments on Subway Dataset

In this section, the proposed method is compared on Subway Exit and Subway Entrance with different methods. Quantitative

TABLE VI
COMPARISON OF ABNORMAL EVENTS DETECTION RATE AND FALSE ALARM RATE ON THE SUBWAY DATASET

	Exit				
	WD	LT	MISC	Total	FA
GT	9	3	7	19	0
Cong [47]	9	–	–	–	2
Kim [45]	9	3	7	19	3
Yuan [12]	9	3	7	19	3
Lu [21]	9	3	7	19	2
Zhao [38]	9	3	7	19	2
Ours	9	3	7	19	2

TABLE VII
COMPARISON OF ABNORMAL EVENTS DETECTION RATE AND FALSE ALARM RATE ON THE SUBWAY DATASET

	Entrance						
	WD	NP	LT	II	MISC	Total	FA
GT	26	13	14	4	9	66	0
Cong [47]	21	6	–	–	–	–	4
Kim [45]	24	8	13	4	8	57	6
Lu [21]	25	7	13	4	8	57	4
zhao [38]	25	9	14	4	8	60	5
Yuan [12]	24	9	13	4	8	58	4
Ours	25	9	14	4	8	60	4

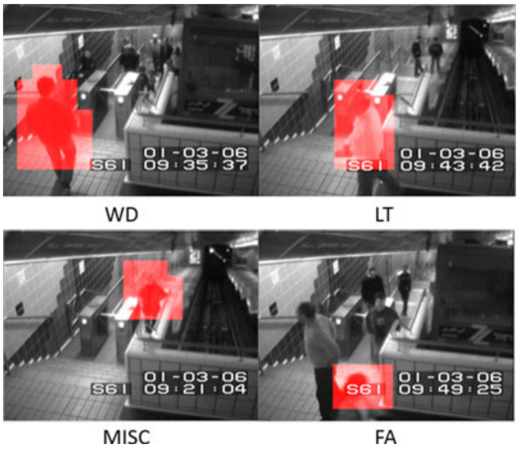

Fig. 12. Detection results on Subway dataset (Exit).

comparison results are listed in Tables VI and VII. The meanings of symbols in Tables VI and VII are listed as follows: GT: ground truth, WD: wrong direction, NP: no payment, LT: loitering, II: irregular interaction, MISC: misc, and FA: false alarm. '–' denotes the result not provided.

For the performance on Subway Exit, the 19 anomalies are all detected by our method and 2 false alarms are detected. The detection results are comparable to the other methods. Some detection results are shown in Fig. 12. It is shown that the anomalies can be correctly detected and located. The false alarms occur when people move fast near the camera although they are moving in normal directions.

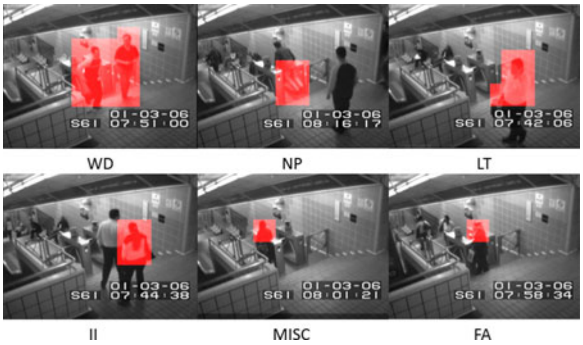

Fig. 13. Detection results on Subway dataset (Entrance).

TABLE VIII
COMPARISON WITH DIFFERENT STRATEGY ON UCSD PED2 DATASET

METHOD	EER	AUC
MHOF + SC + RE [32]	25.0%	86.1%
Dense Sample + SC + RE	24.3%	86.8%
FIP + SC + RE	23.0%	87.4%
FIP + SSC + one-class SVM	18.2%	88.2%
FIP + SSC +RE	17.5%	88.5%
FIP + SSC + intra-frame SVM	16.7%	89.2%

For the Subway Entrance, the results are shown in Table VII. It is shown that our total number of detected anomalies identical to Zhao *et al.*'s work. But our false alarms number is less than Zhao *et al.*'s work. Meanwhile, the performance of detection accuracy and false alarms is better than the other methods. In the experiments, irregular interactions are easily detected. Unfortunately, false alarms are generally raised by normal interactions because the interactions last in different periods. Some people spending more time standing before the ticket checking machine also cause false alarms. Fig. 13 shows some visualized results on Subway Entrance. It is shown that the anomalies can be correctly detected and located. The false alarms occur when people standing before the machine for a long time.

H. *Discussion of Effectiveness*

In this section, the effectiveness of FIP descriptor, SSC encoding and intra-frame classification are evaluated. The experiments are conducted on UCSD Ped2 dataset and the comparisons of using different descriptors and encoding methods are shown in Table VIII. The EER and AUC are calculated for evaluation. In Table VIII, Cong *et al.*'s method, which used MHOF feature, SC encoding and Reconstruction Errors (RE) to evaluate abnormality, is taken as the baseline. It obtains an EER of 25.0% and an AUC of 86.1%, which is reported in [32]. Using dense sample method with HOG/HOF descriptor enhances the AUC to 86.8%. Using FIP descriptor decreases the EER to 23.0% and enhances the AUC to 87.4%. As MHOF describes a whole sub-region and dense sample method densely samples the raw pixels, the two method both bring some background noise and some zero vectors in flat region. Meanwhile, our method extracts and describes FIPs separately, which increases the feature number and excludes zero vectors existing in features. The

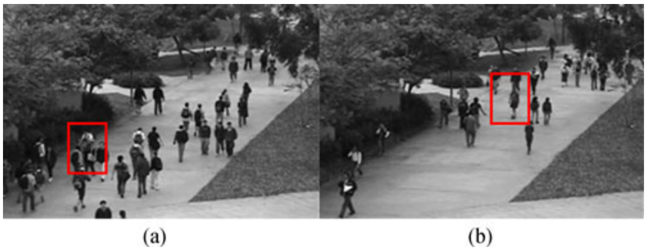

Fig. 14. The anomalies missed in UCSD ped1 dataset. (a) is an obscured bicycle, (b) is a skater moving towards the camera.

results demonstrate the contribution of FIP descriptor. In addition, the use of SSC encoding method with FIP decreases the EER to 17.5% and increases the AUC to 88.5%. It demonstrates that using SSC to encode both the spatial connections and temporal connections of block is better than only using SC. To address the effectiveness of intra-frame strategy, one-class SVMs are trained for each sub-regions to evaluate the anomaly. It achieves an 18.2% EER and an 88.2% AUC. It mainly because when training one-class SVMs, some training samples will be taken as negative samples, which causes some false alarms. The use of probabilistic outputs of SVM with FIP and SSC enhances the AUC to 89.2% and decreases the EER to 16.7%. It shows that using intra-frame SVM performs better than directly using RE and using one-class SVMs, which illustrates that the use of intra-frame classification strategy helps to increase the detection accuracy.

The computational complexity of our SSC method is higher than the single layer sparse coding because SSC encodes both the connections of FIPs and the connections of frames. However, the use of scale penalty decreases the dictionary size which lowers the feature dimensions and makes the detection faster. In addition, the intra-frame strategy trains a multi-class SVM rather than training multiple one-class SVMs, which also lower the computational complexity.

The limitation and some failure cases are also discussed in this section. The limitation of our method is that the anomaly is detected frame by frame, which makes the detection discontinuous. It only evaluates the anomaly based on the current sub-region pattern. If an abnormal object occupies several blocks and several frames, only the detected anomaly blocks will be marked. In other words, if an abnormal object is split into sub-regions and a part of the abnormal object moves normally in its sub-region, it will cause an incomplete marking result of the abnormal object. Some failures are given in Fig. 14. It shows the anomalies that are missed by our method. (a) is a moving bicycle, in this frame, the bicycle is obscured by pedestrians and the moving speed is low, so the anomaly is not detected. (b) is a moving skater, in this frame, the skater moves towards the camera, the appearance and the speed of the skater is normal as other pedestrians, so the anomaly in this frame is missed.

VII. Conclusion

In this paper, an efficient anomaly detection method based on Stacked Sparse Coding (SSC) with intra-frame classification strategy is proposed. An FIPs descriptor is proposed to describe the spatial and temporal features in sub-regions, which makes the detection and localization results more accurate. The SSC encoding method encodes both the spatial connections and temporal connections of sub-regions, which makes the features more representative. The intra-frame classification strategy takes the advantage of multi-class SVM and it helps to improve the detection results.

The proposed method is examined on four public datasets with different background complexities and resolution: UCSD Ped1 dataset, UCSD Ped2 dataset, Avenue dataset and Subway dataset. The results are compared with previous approaches and it shows the effectiveness and advantages of our method.

Furthermore, deep learning network shows great power in texture description. So the CNN based anomaly detection methods may improve the detection performance. Since the scale of abnormal event dataset is relatively small, it is a good idea to use fine-tuned CNN networks to prevent overfitting. We will validate this idea in our future work.

References

[1] T. Li *et al.*, "Crowded scene analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015.

[2] A. A. Sodemann, M. P. Ross, and B. J. Borghetti, "A review of anomaly detection in automated surveillance," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1257–1272, Nov. 2012.

[3] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1114–1127, Aug. 2008.

[4] S. Cosar *et al.*, "Towards abnormal trajectory and event detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 683–695, Mar. 2017.

[5] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1544–1554, Nov. 2008.

[6] L. Brun, A. Saggese, and M. Vento, "Dynamic scene understanding for behavior analysis based on string kernels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1669–1681, Oct. 2014.

[7] D. Tran, J. Yuan, and D. Forsyth, "Video event detection: From subvolume localization to spatiotemporal path search," *IEEE Trans. Softw. Eng.*, vol. 36, no. 2, pp. 404–416, Feb. 2014.

[8] X. Song *et al.*, "A fully online and unsupervised system for large and high-density area surveillance: Tracking, semantic scene learning and abnormality detection," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 2, 2013, Art. no. 35.

[9] A. R. Revathi and D. Kumar, "An efficient system for anomaly detection using deep learning classifier," *Signal Image Video Process.*, vol. 11, pp. 291–299, 2017.

[10] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—A review," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 865–878, Nov. 2012.

[11] Y. Zhang, L. Qin, R. Ji, H. Yao, and Q. Huang, "Social attribute-aware force model: Exploiting richness of interaction for abnormal crowd detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 7, pp. 1231–1245, Jul. 2015.

[12] Y. Yuan, Y. Feng, and X. Lu, "Statistical hypothesis detector for abnormal event detection in crowded scenes," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3597–3608, Nov. 2017.

[13] Y. J. Lee, Y. R. Yeh, and Y. C. F. Wang, "Anomaly detection via online oversampling principal component analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 7, pp. 1460–1470, Jul. 2013.

[14] G. Xiong *et al.*, "An energy model approach to people counting for abnormal crowd behavior detection," *Neurocomputing*, vol. 83, no. 7, pp. 121–135, 2012.

[15] C. L. Chen, X. Tao, and S. Gong, "Salient motion detection in crowded scenes," in *Proc. 5th Int. Symp. Commun. Control Signal Process.*, 2012, pp. 1–4.

[16] B. Krausz and C. Bauckhage, "Loveparade 2010: Automatic video analysis of a crowd disaster," *Comput. Vis. Image Understanding*, vol. 116, no. 3, pp. 307–319, 2012.

[17] K. W. Cheng, Y. T. Chen, and W. H. Fang, "Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2909–2917.

[18] S. Yi, X. Wang, C. Lu, and J. Jia, "L0 regularized stationary time estimation for crowd group analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2219–2226.

[19] D. Lee, H. I. Suk, S. K. Park, and S. W. Lee, "Motion influence map for unusual human activity detection and localization in crowded scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 10, pp. 1612–1623, Oct. 2015.

[20] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *Visual Comput.*, vol. 29, no. 10, pp. 983–1009, 2013.

[21] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2720–2727.

[22] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2112–2119.

[23] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1975–1981.

[24] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[25] X. Cui, Y. Tian, L. Weng, and Y. Yang, "Anomaly detection in hyperspectral imagery based on low-rank and sparse decomposition," in *Proc. Int. Conf. Graph. Image Process.*, 2014, pp. 90 690R-1–90 690R-7.

[26] X. Mo, V. Monga, R. Bala, and Z. Fan, "Adaptive sparse representations for video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 631–645, Apr. 2014.

[27] V. Reddy, C. Sanderson, and B. C. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2011, pp. 55–61.

[28] T. Bao, S. Karmoshi, C. Ding, and M. Zhu, "Abnormal event detection and localization in crowded scenes based on PCANet," *Multimedia Tools Appl.*, vol. 75, no. 22, pp. 14617–14639, Nov. 2016.

[29] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, 2017.

[30] S. Zhou *et al.*, "Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Process. Image Commun.*, vol. 47, pp. 358–368, 2016.

[31] 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 *Proc. Joint ACM Workshop Modeling Representing Events*, 2011, pp. 25–30.

[32] Y. Cong, J. Yuan, and Y. Tang, "Video anomaly search in crowded scenes via spatio-temporal motion context," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1590–1599, 2013.

[33] M. Thida, H. L. Eng, and P. Remagnino, "Laplacian eigenmap with temporal constraints for local abnormality detection in crowded scenes," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2147–2156, Dec. 2013.

[34] J. Wu, Y. Zhang, and W. Lin, "Towards good practices for action video encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2577–2584.

[35] K. Yu, Y. Lin, and J. Lafferty, "Learning image representations from the pixel level via hierarchical sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1713–1720.

[36] C. Lu, J. Shi, and J. Jia, "Scale adaptive dictionary learning," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 837–847, Feb. 2014.

[37] S. Han, R. Fu, S. Wang, and X. Wu, "Online adaptive dictionary learning and weighted sparse coding for abnormality detection," in *Proc. IEEE Int. Conf. Image Process.*, 2013, pp. 151–155.

[38] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3313–3320.

[39] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *CoRR*, vol. abs/1009.5055, 2010.

[40] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. 15th Int. Conf. Multimedia*, 2007, pp. 357–360.

[41] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[42] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 19–60, 2010.

[43] H. T. Lin, C. J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Mach. Learn.*, vol. 68, no. 3, pp. 267–276, 2007.

[44] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[45] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2921–2928.

[46] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowded scenes," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 56–62.

[47] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3449–3456.

[48] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 935–942.

[49] T. Xiao, C. Zhang, and H. Zha, "Learning to detect anomalies in surveillance video," *IEEE Signal Process. Lett.*, vol. 22, no. 9, pp. 1477–1481, Sep. 2015.

[50] D. Xu *et al.*, "Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts," *Neurocomputing*, vol. 143, no. 16, pp. 144–152, 2014.

Ke Xu received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 2013. He is currently working toward the Ph.D. degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai. His research interests include action recognition and abnormal events detection.

Xinghao Jiang received the Ph.D. degree in electronic science and technology from Zhejiang University, Hangzhou, China, in 2003. He was a Visiting Scholar with the New Jersey Institute of Technology, Newark, NJ, USA, from 2011 to 2012. He is currently a Professor with the School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include multimedia security and image retrieval, intelligent information processing, cyber information security, information hiding, and watermarking.

Tanfeng Sun received the Ph.D. degree in information and communication system from Jilin University, Changchun, China, in 2003. He was a Visiting Scholar with the New Jersey Institute of Technology, Newark, NJ, USA, from 2012 to 2013. He is currently an Associate Professor with the School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include digital forensics on video forgery as well as videos content recognition and understanding.