




Abnormal event detection via covariance matrix for optical flow based feature

Tian Wang¹ · Meina Qiao¹ · Aichun Zhu²  ·
Yida Niu¹ · Ce Li³ · Hichem Snoussi⁴

Received: 7 March 2017 / Revised: 19 August 2017 / Accepted: 13 October 2017 /

Published online: 11 November 2017

© Springer Science+Business Media, LLC 2017

Abstract Abnormal event detection is one of the most important objectives in security surveillance for public scenes. In this paper, a new high-performance algorithm based on spatio-temporal motion information is proposed to detect global abnormal events from the video stream as well as the local abnormal event. We firstly propose a feature descriptor to represent the movement by adopting the covariance matrix coding optical flow and the corresponding partial derivatives of multiple connective frames or the patches of the frames. The covariance matrix of multi-RoI (region of interest) which consists of frames or patches can represent the movement in high accuracy. For public surveillance video, the

✉ Aichun Zhu
aichun.zhu@njtech.edu.cn

Tian Wang
wangtian@buaa.edu.cn

Meina Qiao
meinaqiao@buaa.edu.cn

Yida Niu
2014niuyida@buaa.edu.cn

Ce Li
xjtulice@gmail.com

Hichem Snoussi
hichem.snoussi@utt.fr

¹ School of Automation Science and Electrical Engineering, Beihang University, Beihang, China

² School of Computer Science and Technology, Nanjing Tech University, Nanjing, China

³ College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, China

⁴ Institut Charles Delaunay-LM2S-UMR STMR 6279 CNRS, University of Technology of Troyes, Troyes, France

normal samples are abundant while there are few abnormal samples. Thus the one-class classification method is suitable for handling this problem inherently. The nonlinear one-class support vector machine based on a proposed kernel for Lie group element is applied to detect abnormal events by merely training the normal samples. The computational complexity and time performance of the proposed method is analyzed. The PETS, UMN and UCSD benchmark datasets are employed to verify the advantages of the proposed method for both global abnormal and local abnormal event detection. This method can be used for event detection for a surveillance video and outperforms the state-of-the-art algorithms. Thus it can be adopted to detect the abnormal event in the monitoring video.

Keywords Global abnormal event · Local abnormal event · Multi-RoI · Covariance matrix · Optical flow

1 Introduction

The visual surveillance is one of the major research areas in computer vision. The objective in this field includes picking up the detailed information of the groups as well as individuals. The change detection [4, 28, 37, 42], object tracking [19, 23, 59] have attracted many researchers. Scene classification [8], object segmentation [9], human detection [13], face detection [27, 33, 39], face recognition [34, 45, 47, 57] are also the hottest research fields about the surveillance system. These studies focus on the static biology information recognition, such as the face, appearance or contour, etc. With the development of the technology of video recording and transmission, the dynamic information analysis becomes gradually more and more important. In order to understand the relationship between the individuality and the scene, some researchers have devoted themselves to the research about the behavior. In [15], the behavior of one single person, such as boxing, running, skipping is recognized. In [5, 11, 25], the behavior of persons in the real life scene such as movies and sport videos is researched. The research about human detection and action analysis provides useful technologies and theories for visual information processing, thus the abnormal event detection in the public scene becomes an important research topic both in theory and reality in video surveillance.

The abnormal event detection is also important for the social security of the public scene. The public scene is an important place which involves the daily lives and social activities. The security surveillance problem in the public scene is related to livelihood matters. Since the security guards have to deal with many surveillance videos, it is easy for them to visual and mental fatigue. Thus, they cannot react quickly to the unexpected or abnormal events. Because the monitors are widely used in public scene to ensure safety, the automatic alarm system which aids the security guard to find the emergency events from a large number of videos is crucial. In our work, the global abnormal event detection and the local event detection problems from the surveillance videos are researched. In this paper, we propose an algorithm that simultaneously detects global abnormal events in video streams, and locates the local abnormal events on the frame. The proposed feature descriptor characterizing the spatio-temporal movement information is calculated via the same algorithm in our work. By the covariance matrix descriptor, the optical flow and the corresponding partial derivatives are fused and projected into a lower dimensional feature space. Different from the general binary classification problem, the abnormal event detection is a one-class classification problem which only has the normal samples and few or none of the abnormal samples

are adopted for training. So we use one-class SVM method rather than other classification methods in this paper. As the covariance matrix is in a Lie Group, the designed kernel is suited for calculating the dissimilarity of two elements with the one-class SVM framework. The abnormal event is detected based on the one-class SVM with the proposed proper kernel function for the covariance matrix descriptor. Experiments on the benchmark datasets validate that our model outperforms the state-of-the-art algorithms. The research of this paper will not only develop the theory of abnormal activity detection problem in complex public scenes, but also provide more technologies in the intelligent security surveillance system.

The rest of the paper is organized as the following. Section 2 reviews related works briefly. Section 3 elaborates the proposed motion descriptor, i.e. the multi-RoI covariance matrix descriptor. And then, the suitable kernel for the descriptor in the nonlinear one-class support vector machine is proposed. In Section 4, the proposed global abnormal event and local abnormal event algorithms are further clarified. In Section 5, the experimental results of datasets PETS [35], UMN [50] and UCSD [49] are illustrated and discussed. Finally, Section 6 concludes the paper and gives a perspective of future work.

2 Related work

For abnormal event detection, some researchers focused on the tracking based method. In [36], by analyzing the trajectory, the abnormal event was detected. As the trajectory of the object is hard to be detected due to the occlusion, other feature based methods without trajectory extraction, such as motion history images (MHI), motion energy images (MEI), pixel change history (PCH) [2, 22, 24] were used to present the movement from the original images. In [55], the motion feature, namely expanded relative motion histogram of bag-of-visual-words (ERMH-BoW) was proposed for event detection. In [38], the event saliency concept was introduced, and the strategy which was based on gamification was defined to extract reliable event saliency maps from images by exploiting human interaction. In [12], the video sequences were compressed into two-dimensional action matrices, and then the action categorization problem was handled by a clustering method. These methods were proposed based on the texture or pixel intensity information. And the movement is represented by the pixel intensity arrangement, but the pixels with similar intensity influence the accuracy. Besides, optical flow shows the instantaneous velocity of the motion of pixels in the image plane. Quantized optical flow directions and intensity were used as the basic feature to describe the movement [1, 51–53, 60]. The wavelet transform which was used in image processing can also be adopted to analyze the movement [3, 58]. The statistics property of optical flow of the pixels in the frame or a chosen block was calculated. As deep learning developed quickly, it has been used for the abnormal event detection in recent years. Different from the hand-crafted feature, the feature in the deep learning network was learnt across the layers [18, 61]. The paper [18] used the conventional handcrafted spatio-temporal features as the input of the fully-connected autoencoder to detect the abnormal event. Then they also built a fully-convolutional autoencoder to learn features across layers in a 3D form. Although the deep learning outperformed the hand-crafted feature, it suffered from the problem of computation and storage. It took a long time to train an available model. The probabilistic framework has been applied to model the temporal structure of videos and text [46], such as Hidden Markov Model (HMM) [43]. The authors in [46] learnt the latent temporal structure of complex events in highly varying Internet videos. They trained the variable-duration HMM in a max-margin framework to discover the segments of videos

and achieved different detection and recognition tasks. In this paper, the optical flow is also adopted as the basic feature, but the relations between the optical flows of the pixels are fused with other components of a video to construct a movement descriptor.

The covariance matrix descriptor was firstly introduced for human detection and tracking [48]. It has the advantages of overcoming the illumination instability and rotation. The covariance matrix descriptor is computed based on the extracted multi-modal features from the object, thus retains much more information than a single-modal feature. With the discrimination property, the article [17] adopted the covariance feature for texture classification. The paper [21] recognized the human action based on covariance feature modeling the human joint locations. The paper [14] detected the abnormal event by the representation of trajectories via covariance features. Because the trajectories were needed to be extracted firstly, the occlusion problem also influenced the detection performance. The paper [54] adopted covariance to detect abnormal frame event, but the local abnormal event was not detected. Because the covariance matrix of the optical flow of one frame is influenced by the noise which is induced by the optical flow calculated method. The covariance matrix of multi-RoI can fuse the movement information of several consecutive and related RoI, thus it represents a higher accurate movement information. With the advantage of covariance matrix which can fuse different features, in this paper the covariance matrix which encodes the optical flow to a treatable descriptor is proposed for the global and local abnormal event detection. The descriptor is proposed without object tracking for avoiding the occlusion problem in crowd scenes. Moreover, the one-class SVM classifies the covariance which models the relations of the optical flow in the RoI.

3 Motion descriptor

For the objective of detecting abnormal action in a surveillance video, the information of the movement should be described by a feasible descriptor which can be handled by the machine learning method. Optical flow is the pattern of apparent motion of objects in a visual scene caused by the relative motion between an observer and a scene [6, 56]. In the formulation of Horn and Schunck (HS), the algorithm optimizes an objective function which combines brightness constancy constraint with a spatial term modeling how the flow is expected to vary across the image [20, 44]. The Lucas-Kanade method assumes that the flow is essentially constant in a local neighborhood of the pixel under consideration [29]. As the optical flow information in the interior of uniform regions of the image needs to be considered, the HS method used to provide both amplitude and direction of a pixel. Optical flow is then chosen as the basic feature to represent the motion between two image frame. Optical flow method optimizes an objective function which combines brightness constancy constraint with a spatial term modeling how the flow is expected to vary across the image [20, 44]. The aperture problem is solved by introducing a global constraint of smoothness. Optical flow is formulated as:

$$E = \int \int [(I_x u + I_y v + I_t)^2 + \alpha^2 (\|\nabla u\|^2 + \|\nabla v\|^2)] dx dy, \quad (1)$$

where E is the global energy. I_t , I_x and I_y are the intensities along the time t direction, x (width) direction and y (height) direction. α is the hyperparameter representing the weight of the regularization term. u and v are the horizontal and vertical components of the optical flow.

Based on the optical flow, we propose a covariance matrix based feature descriptor fusing the spatio-temporal motion information of several consecutive RoI (Region of Interest).

Firstly, a video stream is split into groups containing n frames, and the optical flow is calculated on every two consecutive frames. In each group, the features of the pixels in the RoI from 1-st to n -th are arranged. Taking the k -th group as an example, the pixels from position $(1, 1)$ to (h, w) are sampled in a RoI with height h and width w . The movement information is coded by the matrix with $2 + l$ columns and $(h \times w)$ rows, where l is the length of the optical flow based feature, h and w are the height and width of the RoI, respectively. The feature matrix of one RoI is in the form as presented in (2):

$$\left[\begin{array}{cc|c} 1 & 1 & \text{feature} \\ 1 & 2 & \text{feature} \\ \vdots & \vdots & \vdots \\ h & w & \text{feature} \end{array} \right] \left. \vphantom{\begin{array}{cc|c} 1 & 1 & \text{feature} \\ 1 & 2 & \text{feature} \\ \vdots & \vdots & \vdots \\ h & w & \text{feature} \end{array}} \right\} h \times w$$

$\underbrace{\hspace{1.5cm}}_2 \quad \underbrace{\hspace{1.5cm}}_l$

(2)

For the multi-RoI covariance feature, each RoI in a group is labeled with the serial numbers from 1-st to n -th. Then, the feature matrix of one RoI group is in the form as shown in (3), where $\mathbf{1}$ is a vector with all the elements 1. The dimension of the matrix of one group is $(n \times h \times w) \times (3 + l)$. The first column is the RoI position in the group, the second and third columns are the pixel positions in one RoI. The optical flow based feature is defined according to the objective. For example, the feature defined as the intensity might be feasible for the object recognition or tracking problem. For the abnormal event detection problem, the optical flow based feature is organized, as Table 1.

$$\left[\begin{array}{cc|c} \mathbf{1} & 1 & 1 & \text{feature} \\ & 1 & 2 & \text{feature} \\ & \vdots & \vdots & \vdots \\ & h & w & \text{feature} \\ \hline 2 \times \mathbf{1} & 1 & 1 & \text{feature} \\ & 1 & 2 & \text{feature} \\ & \vdots & \vdots & \vdots \\ & h & w & \text{feature} \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline n \times \mathbf{1} & 1 & 1 & \text{feature} \\ & 1 & 2 & \text{feature} \\ & \vdots & \vdots & \vdots \\ & h & w & \text{feature} \end{array} \right] \left. \vphantom{\begin{array}{cc|c} \mathbf{1} & 1 & 1 & \text{feature} \\ & 1 & 2 & \text{feature} \\ & \vdots & \vdots & \vdots \\ & h & w & \text{feature} \\ \hline 2 \times \mathbf{1} & 1 & 1 & \text{feature} \\ & 1 & 2 & \text{feature} \\ & \vdots & \vdots & \vdots \\ & h & w & \text{feature} \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline n \times \mathbf{1} & 1 & 1 & \text{feature} \\ & 1 & 2 & \text{feature} \\ & \vdots & \vdots & \vdots \\ & h & w & \text{feature} \end{array}} \right\} h \times w$$

$\underbrace{\hspace{1.5cm}}_3 \quad \underbrace{\hspace{1.5cm}}_l$

(3)

Table 1 Optical flow based *feature* to construct the covariance descriptor

Feature vector F	
$F_1(4 \times 4)$	$[y \ x \ u \ v]$
$F_2(8 \times 8)$	$[y \ x \ u \ v \ u_x \ u_y \ v_x \ v_y]$
$F_3(14 \times 14)$	$[y \ x \ u \ v \ u_x \ u_y \ v_x \ v_y \ u_{xx} \ u_{yy} \ v_{xx} \ v_{yy} \ u_{xy} \ v_{xy}]$

For the abnormal detection problem, the feature is defined as:

$$F(x, y) = \phi(x, y, u, v), \quad (4)$$

where u, v are the horizontal and vertical optical flow, ϕ relates to the optical flow with the i -th feature. F is the feature extracted, for one group of RoIs it is in dimension $(h \times w) \times (3 + l)$, as described in (2) and (3). Based on a feature F describing the feature, a $(3 + l) \times (3 + l)$ dimension covariance matrix [48] is defined:

$$C = \frac{1}{m-1} \sum_{i=1}^m (z_i - \mu)(z_i - \mu)^T, \quad (5)$$

where m is the number of the pixels sampled. It is represented by a $n \times h \times w$ for a group of RoIs. μ is the mean of the m feature vectors. z_i is the feature vector of the i -th point. Thus, numerous features are fused by the covariance matrix. The calculation progress of the proposed feature is shown in Fig. 1. In a video stream V , each group G containing n frames is represented by a feature matrix F , and then the multi-frame covariance matrix C is calculated. The pixel position, optical flow and corresponding partial derivative characterize the intra-RoI information. Moreover, the RoI position in one group stamps the inter-RoI knowledge. Thus, the covariance descriptor proposed reveals the spatio-temporal feature fusing the RoI movement property. For a covariance matrix C which fuses l features, due to symmetry it has only $\frac{l^2+l}{2}$ different values. The computational complexity of constructing a covariance of a rectangular region is $O(l^2)$ with the aid of the integral image. While if the same region is represented by raw values, the dimension of the feature is $n \times l$, where n is the number of the pixels. And if the joint feature histograms are used, the dimension of the feature is b^l , where b is the number of the histogram bins. Thus, the covariance matrices are low-dimensional and efficient compared to other region descriptors [48].

In video surveillance, a large number of normal scenes are captured, but the abnormal scenes happened with quite few frequencies. Thus, in our work the abnormal detection objective relates to a one-class classification problem. One-class SVM method is adopted

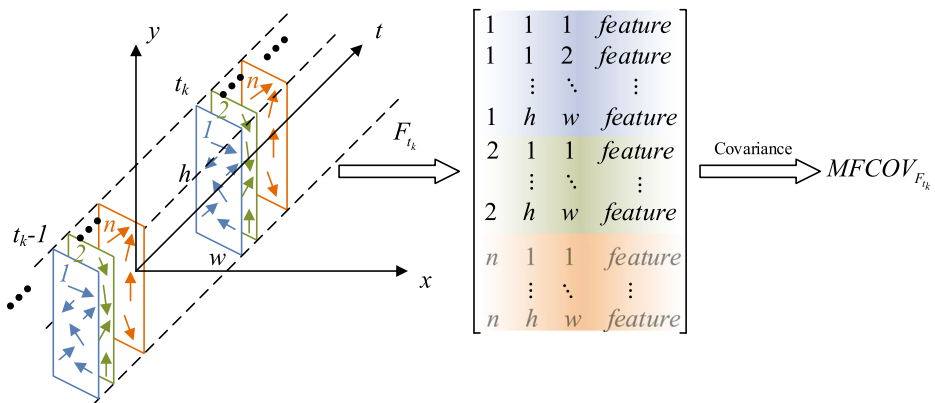


Fig. 1 The calculation progress of multi-RoI covariance matrix feature. RoI: Region of Interest. The arrows demonstrate the optical flow of the RoI. Different RoIs are distinguished by blue, green and red colors

basically in this paper. The problem of non-linear one-class SVM (OCSVM) [7, 40] can be presented as a constrained minimization one:

$$\min_{\omega, \xi, \rho} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{\nu n} \sum_{i=1}^n \xi_i - \rho, \quad (6)$$

$$\text{subject to: } \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle \geq \rho - \xi_i, \xi_i \geq 0. \quad (7)$$

The decision function in the data space \mathcal{X} is defined as:

$$f(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^n \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}) - \rho \right), \quad (8)$$

where \mathbf{x} is a vector in the input data space \mathcal{X} . $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ maps datum \mathbf{x}_i into the feature space \mathcal{H} , $\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle - \rho = 0$ is the decision hyperplane, ξ_i is the slack variable for penalizing the outliers. The hyperparameter $\nu \in (0, 1]$ is the weight for restraining slack variable. κ is the kernel function implicitly mapping the data into a higher dimensional feature space where a linear classifier can be designed.

For one-class SVM, the kernel functional κ of two covariance matrices must be computed. We choose the Gaussian kernel which is usually defined by the following expression:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right), (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{X} \times \mathcal{X}, \quad (9)$$

where the parameter σ indicates the scale factor where the data should be clustered, \mathbf{x}_i and \mathbf{x}_j are two vectors.

Because the covariance matrix is an element in a Lie Group G , the distance measuring the dissimilarity of two elements is defined as:

$$d(\mathbf{X}_1, \mathbf{X}_2) = \|\log(\mathbf{X}_1^{-1} \mathbf{X}_2)\|, \quad (10)$$

$$\text{with } \|A\| = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}, \quad (11)$$

where $\|\cdot\|$ is the Frobenius norm, a_{ij} is an element in the matrix A , \mathbf{X}_i and \mathbf{X}_j are the matrices in a Lie Group G . Thus, the Gaussian kernel in one-class SVM for the covariance matrix which is the element in a Lie Group G is:

$$\kappa(\mathbf{X}_i, \mathbf{X}_j) = \exp \left(-\frac{\|\log(\mathbf{X}_i^{-1} \mathbf{X}_j)\|}{2\sigma^2} \right), \quad (\mathbf{X}_i, \mathbf{X}_j) \in G \times G. \quad (12)$$

The Baker Campbell Hausdorff formula [16] in the theory of Lie Group is:

$$\begin{aligned} \log(\exp X \exp Y) = & \sum_{n>0} \frac{(-1)^{n-1}}{n} \sum_{\substack{r_i+s_i>0 \\ 1 \leq i \leq n}} \frac{(\sum_{i=1}^n (r_i + s_i))^{-1}}{r_1! s_1! \cdots r_n! s_n!} \\ & [X^{r_1} Y^{s_1} X^{r_2} Y^{s_2} \cdots X^{r_n} Y^{s_n}]. \end{aligned} \quad (13)$$

By using the first term of (13), the approximate form of the Gaussian kernel in Lie Group is:

$$\kappa(X_i, X_j) = \exp\left(-\frac{\|\log(X_i) - \log(X_j)\|^2}{2\sigma^2}\right),$$

$$(X_i, X_j) \in G \times G, \quad (14)$$

where $\log(X)$ is a symmetrical matrix. The covariance descriptor C is of size $d \times d$, and has only $\frac{d^2+d}{2}$ different features. By choosing the $\frac{d^2+d}{2}$ upper triangular and the diagonal elements of the matrix $\log(X)$ to construct a vector \bar{x} , the Gaussian kernel can be written as:

$$\kappa(X_i, X_j) = \exp\left(-\frac{\|\bar{x}_i - \bar{x}_j\|^2}{2\sigma^2}\right). \quad (15)$$

4 Abnormal event detection based on classification

Using the proposed descriptor and the classification method presented in Section 3, we propose an abnormal event detection method based on the offline training stage. The flowchart of the whole procedure is shown in Fig. 2, and described in details in Algorithm 1.

Algorithm 1 Abnormal event detection based on multi-RoI covariance feature

Require:

Image training set $S_a = \{I_1^{N_a}\}$, image testing set $S_e = \{I_1^{N_e}\}$

- 1: Compute optical flow of training images,

$$\{I_1^{N_a}\} \rightarrow \{O_1^{N_a}\}.$$

- 2: Feature matrix is calculated based on each consecutive n RoIs,

$$\{O_1^{N_a}\} \rightarrow \{O_1^{P_a}\},$$

where $P_a = \lfloor N_a/n \rfloor$.

- 3: Covariance feature descriptor is calculated

$$\{O_1^{P_a}\} \rightarrow \{C_1^{P_a}\}.$$

- 4: Training samples $\{C_1^{P_a}\}$ are learnt based on one-class SVM, the support vector and decision hyperplane are derived.

$$\{C_1^{P_a}\} \xrightarrow{OCSVM} \{C_1^{P'}\}$$

- 5: Testing samples $\{C_1^{P_e}\}$ is calculated from testing frames $\{S_e\} = \{I_1^{N_e}\}$ online in each n consecutive RoIs.

- 6: Based on the support vector $C_j \in \{C_1^{P'}\}$ and hyperplane, each test sample $C_i \in \{C_1^{P_e}\}$ of consecutive RoIs is classified in the online detection stage.

$$f(C_i) = \text{sgn}\left(\sum_{j=1}^{P'} (\alpha_j \kappa(C_j, C_i) - \rho)\right)$$

$$= \begin{cases} 1 \rightarrow \text{normal}, & f(C_i) \geq 0 \\ -1 \rightarrow \text{abnormal}, & f(C_i) < 0. \end{cases}$$

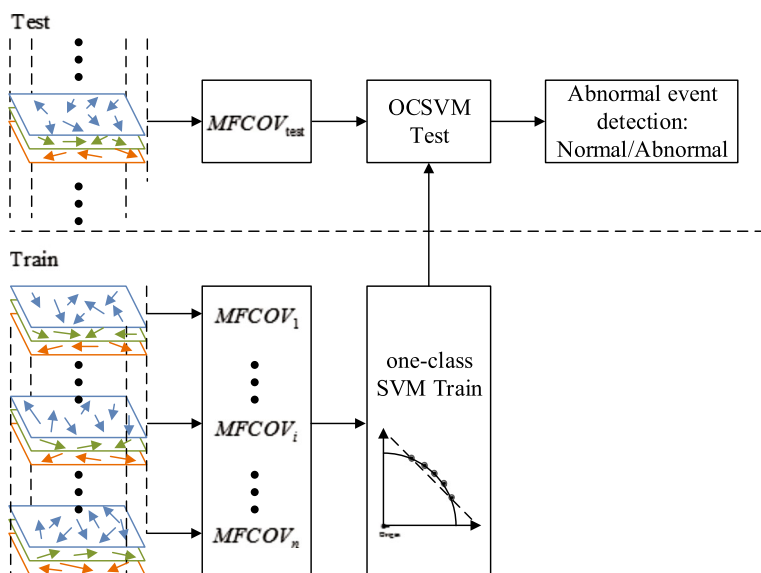


Fig. 2 The flowchart of the multi-ROI covariance feature based abnormal event detection method

Assuming that the training set S_a of images from 1-st to N_a -th, and testing set S_e of images from 1-st to N_e -th are obtained, the procedure of the algorithm is described as the following.

- Step 1:** The optical flow of training frames is calculated via Horn and Schunck (HS) optical flow method as introduced in Section 3. The optical flow is expressed as O , and for the training samples, the optical flow set is expressed as $O_1^{N_a}$.
- Step 2:** The training set is split into several consecutive segments. Each segment has n consecutive RoIs. The training set $\{O_1^{N_a}\}$ is transformed into set $\{O_1^{P_a}\}$, where $P_a = \lfloor N_a/n \rfloor$. Since the number of videos is much smaller than the number of video frames, this design ensures good scalability of our solution to deal with large-scale video sets [31].
- Step 3:** The multi-ROI covariance matrix feature descriptor as proposed in Section 3 is calculated for each consecutive segment. Thus, the training sample set $\{C_1^{P_a}\}$ is obtained.
- Step 4:** The training sample set $\{C_1^{P_a}\}$ is learnt by one-class SVM with the proposed kernel function in Section 3, and the support vector and the decision hyperplane are obtained. The support vectors which are small part of the training samples are recoded as $\{C_1^{P'_a}\}$.
- Step 5:** Similarly, the optical flow $\{O_1^{N_e}\}$ of testing frame set $S_e = \{I_1^{N_e}\}$ is calculated, and then the set is transformed into $\{O_1^{P_e}\}$, where $P_e = \lfloor N_e/n \rfloor$. Thus the testing samples $\{C_1^{P_e}\}$ is calculated.
- Step 6:** The testing samples are classified based on the support vectors and the hyperplane function of one-class SVM as shown in (8). Thus, the abnormal event is detected.

5 Event analysis results

This section presents the experiments of benchmark datasets PETS [35], UMN [50] and UCSD [49] to evaluate the performance of the proposed method for abnormal event

detection. The RoI of consecutive 4 frames is used to construct the multi-RoI covariance feature. For the global abnormal event detection in PETS and UMN datasets, the whole frame is considered as the RoI. While for the local abnormal event detection in UCSD dataset, each 16×12 block is selected as the RoI.

5.1 PETS dataset

In Section 5.1, the PETS dataset is adopted for global abnormal behavior detection. The crowd count, detection of separate flows and specific crowd events are taken into account in the PETS dataset. The specific crowd event detection problems which consist of people moving in the same direction and splitting queue are studied in our work. Two experiments which distinguish the people moving toward the same direction from the loitering, and distinguish the splitting queues from the one queue are undergone. In the abnormal event detection, the anomalous event is denoted as positive and the absence of anomalous events is negative. For global abnormal event detection (GAE), TP (true positive) is the frames which are abnormal in practice, are detected as abnormal by our algorithm. TN (true negative) is the frames which are normal in practice, are detected as normal by our algorithm.

The detection results of *Time 14-17* are shown in Fig. 3. The normal training samples are selected from *Time 14-55* where people are loitering in different directions. The abnormal samples are the events where people are walking and then running in the same direction. The scene simulates that the junction is for loitering, thus people are walking in different directions. Contrastively, the abnormal event is that the people are marching in the same direction. The parameters of the SVM are changed for different classification performance, then the intermediate points of the ROC are obtained. The detection performance is shown in Table 2. While the features are chosen from $F_1 (4 \times 4)$ to $F_2 (8 \times 8)$ in Table 1, the detection performance is increased correspondingly. But if more higher partial derivatives are selected as feature $F_3 (14 \times 14)$, the performance is not better than F_2 . The higher partial derivatives dilute the significance of the difference between the abnormal and normal events.

The second experiment detects the separate flows which are defined as the abnormal events, as shown in Fig. 4. This experiment simulates the situation that some members leave the queue and form new queues. Training frames are selected from *Time 14-16* where people are walking in the same direction. In *Time 14-31*, a crowd of people initially walk in the same queue and then split into several queues. Compared with experiment in *Time 14-17*, the task of *Time 14-31* is more difficult. The detection performance is shown in Table 2. In *Time 14-17* we need to distinguish the one direction movement from mixed and disorderly movement. In *Time 14-31*, one direction movement is differentiated from three direction movement. The difference between the events is milder. Nevertheless, the proposed algorithm detects the abnormal event with the high performance.

5.2 UMN dataset

The panic situation is detected as the abnormal event of UMN dataset [50] which includes the lawn, indoor and plaza scenes. In the UMN dataset, the panic situations in the lawn, plaza and indoor scenes are detected as the abnormal events. There are eleven video sequences in the dataset, and each dataset is split into normal and abnormal samples. The frames where people are walking are taken as training samples and normal testing samples. Frames simulating panic scenes where people running are regarded as abnormal samples. Figures 6, 7 and 8 show the ROC (receiver operating characteristic) curve of UMN dataset in lawn, indoor and plaza scenes, respectively. The detection performance is shown in Table 3. The

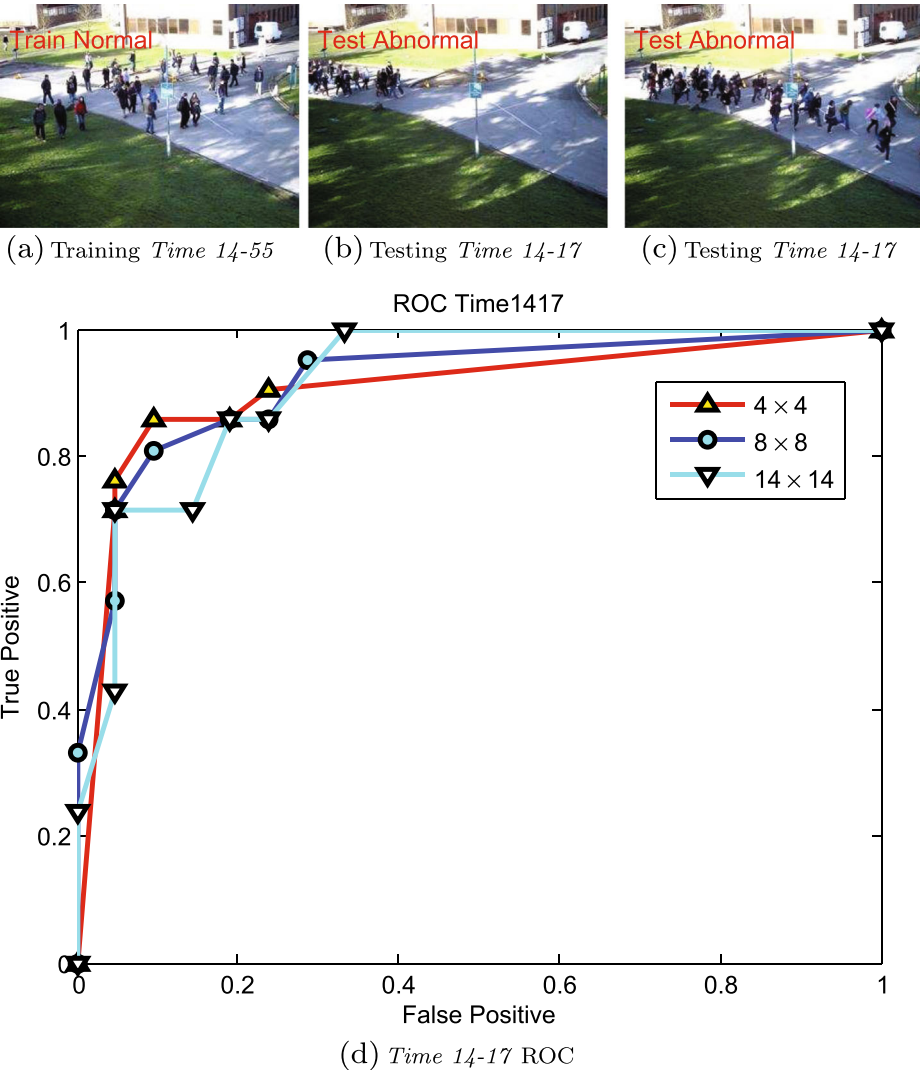


Fig. 3 Time 14-17 scene detection results. **a** A training normal sample from sequence Time 14-55 in which the individuals are walking disorderly. **b** An abnormal sample from sequence Time 14-17 in which crowd are walking in the same direction. **c** An abnormal sample from sequence Time 14-17 in which crowd are running in one direction. **d** The ROC curve of the detection result via different feature

Table 2 The AUC of proposed algorithms based on different features for Time1417 and Time1431 experiments

Feature	Area under ROC	
	Time 1417	Time 1431
$F_1(4 \times 4)$	0.9048	0.8070
$F_2(8 \times 8)$	0.9184	0.9101
$F_3(14 \times 14)$	0.9172	0.8684

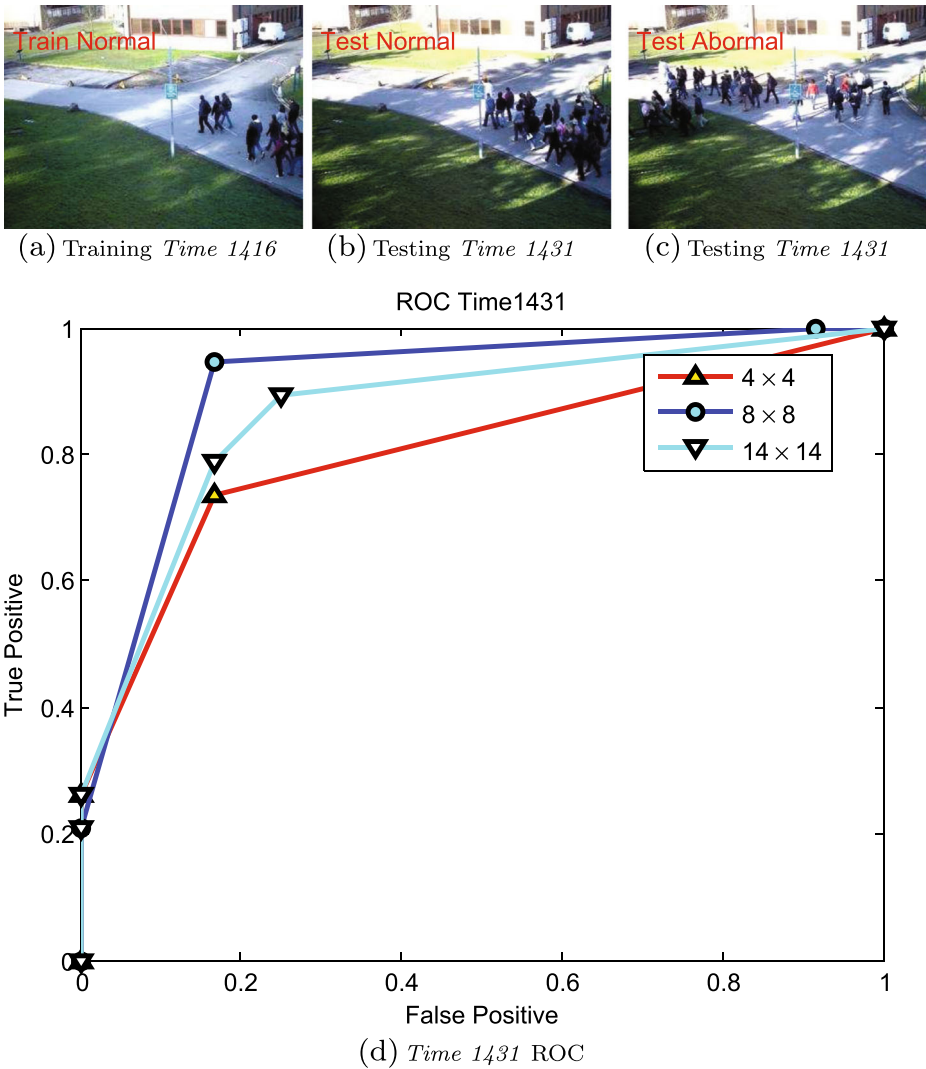


Fig. 4 Time14-31 scene detection results. **a** A training normal sample chosen from Time 14-16 in which people walk along the same direction. **b** A testing normal sample from Time 14-31. **c** A testing abnormal sample from Time 14-31 in which one queue split into several flows. **d** The ROC curve of the detection result via different feature

Table 3 The AUC of proposed algorithms based on different features for lawn, indoor and plaza experiments

Feature	Area under ROC		
	Lawn	Indoor	Plaza
$F_1(4 \times 4)$	0.9935	0.6732	0.9422
$F_2(8 \times 8)$	0.9997	0.9521	0.9770
$F_3(14 \times 14)$	0.9996	0.9448	0.9759

Table 4 Performance evaluation of the multi-RoI covariance matrix of optical flow based feature based method and the state-of-the-art methods of the UMN dataset

Method	Area under ROC		
	Lawn	Indoor	Plaza
Social Force [32]	0.96		
Optical Flow [32]	0.84		
NN [10]	0.93		
SRC [10]	0.995	0.975	0.964
STCOG [41]	0.9362	0.7759	0.9661
HOFO [53]	0.9845	0.9037	0.9815
MCOV (proposed)	0.9997	0.9521	0.9770

performance evaluation compared with other state-of-the-art methods is shown in Table 4. The covariance descriptor based method demonstrates competitive performances. Similar to the PETS dataset, the 8×8 feature type in Table 1 obtain the best performance. Therefore, the proper feature leads to better performance, and the feature contains maximum components does not bring about a better result necessarily.

From the experiment of the PETS dataset and UMN dataset, we prove that the proposed abnormal detection method can distinguish the global abnormal event from the normal training frames. And also, as shown from Figs. 3, 4, 5, 6, 7 and 8, the F_1 (4×4) feature descriptor leads to results with lower accuracy, and the high dimensional feature vector gets more accurate results.

Generally, the more features the covariance matrix fuses, the higher accuracy the classification based detection method can obtain. But more partial derivatives may also dilute the discrimination. Thus, the proper feature components are important for abnormal event detection.

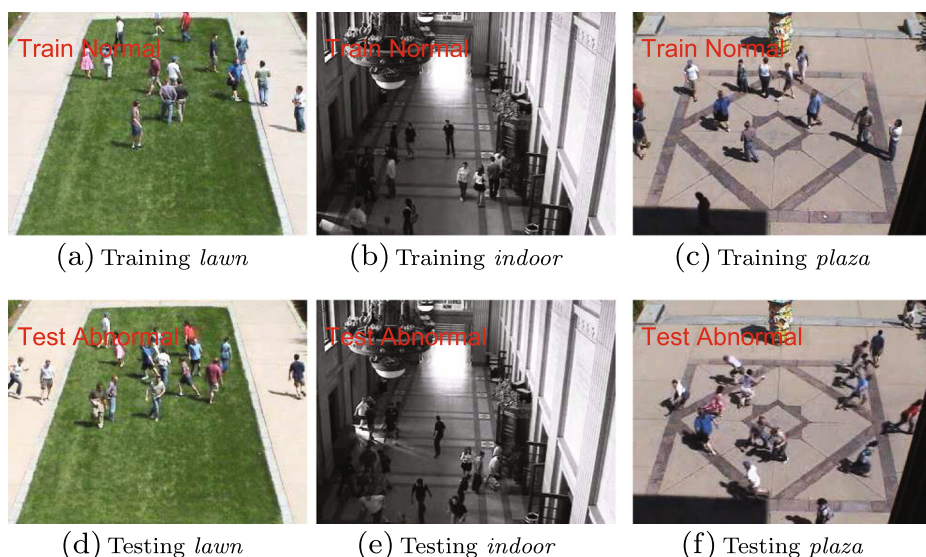


Fig. 5 UMN dataset results. **a, b, c** Training normal samples from *lawn*, *indoor* and *plaza* scenes. People are walking in random directions. **d, e, f** Testing abnormal samples. The individuals are running

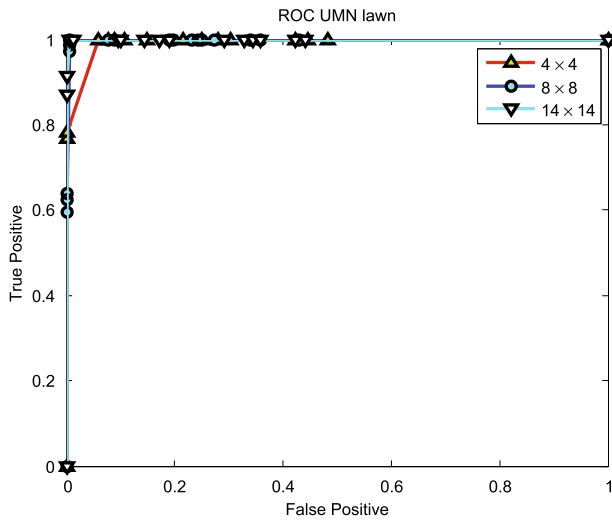


Fig. 6 The ROC of *lawn* scene with different feature

5.3 UCSD dataset

Besides global abnormal events which have been detected with the proposed method in PETS and UMN datasets, UCSD dataset is applied to detect local abnormal events. Normal events which contain only pedestrians are used for training, while the abnormal events correspond to anomalous individual motion patterns or the circulation of non-pedestrian entities such as bike or van in the road. In Ped1 scene, 34 normal video clips are employed for training, and 36 video clips which have one or more abnormal events are applied for testing. Each clip consists of 200 frames with the resolution 158×238 . The frame-level and pixel-level

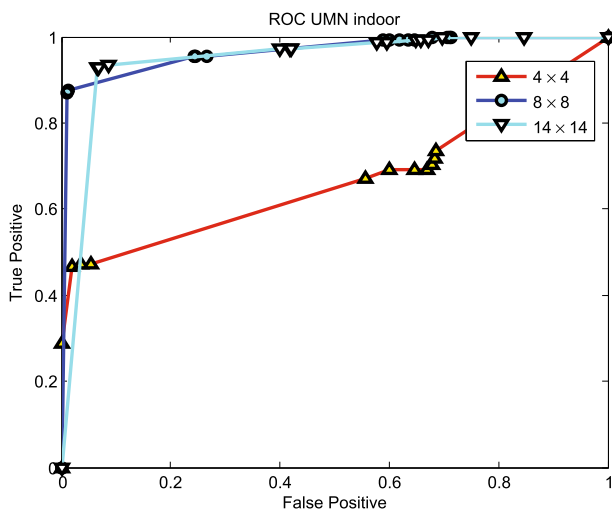


Fig. 7 The ROC of *indoor* scene with different feature

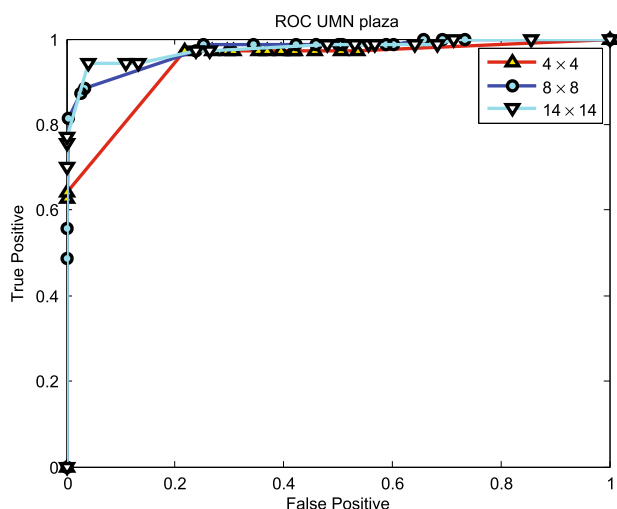


Fig. 8 The ROC of *plaza* scene with different feature

criteria are employed for performance evaluation. In the frame-level criterion, a method detects which frame includes abnormal events. The detection is compared with the frame-level ground truth annotations. The pixel-level criterion is stricter than the frame-level one. In the pixel-level criterion, the method predicts which pixels are related to abnormal events. Only if 40% or more pixels are detected correctly, the frame is considered as a true abnormal one. The EER, RD and AUC are used for performance evaluation. EER (equal error rate) is the ratio of misclassified frames when the false positive rate is equal to the miss rate for the frame-level criterion. RD (rate of detection) is the detection rate at equal error. AUC (area under curve) is the area under ROC (receiver operating characteristic) curve [26, 30].

In our proposed algorithm, we do not extract accurate objects, and do not obtain the precise silhouettes of the moving objects either. Each image is split into local patches of size

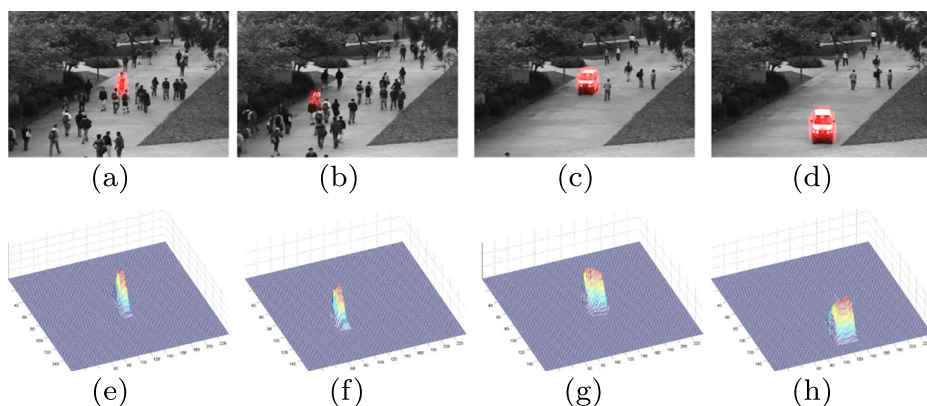


Fig. 9 Examples of local abnormal detection via our proposed algorithm. **a, b, e, f** Local abnormal detection results of *Test003* scene. The abnormal event referring to a bike is detected. **c, d, g, h** Local abnormal detection results of *Test019* scene. The abnormal event referring to a car is detected

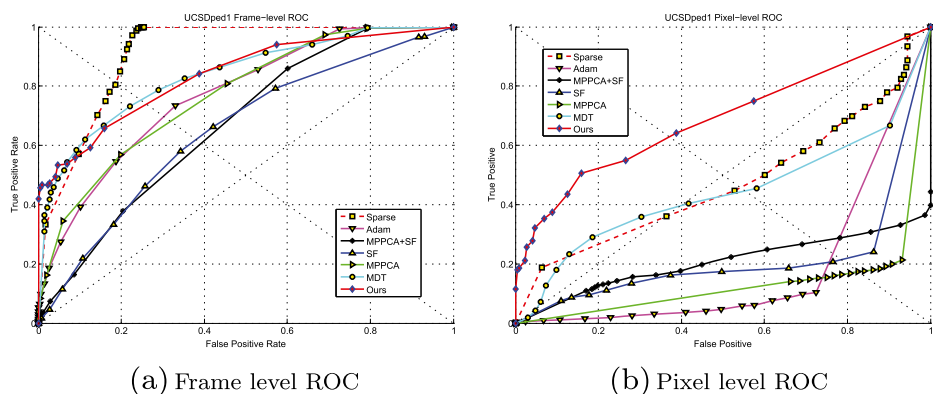


Fig. 10 Evaluation results of Ped1 scene. **a** Frame-level ROC for Ped1 scene. **b** Pixel-level ROC for Ped1 scene

16×12 . The training sample with $F_2(8 \times 8)$ feature for local abnormal event detection is the extracted descriptor in the training dataset. The feature descriptor at the corresponding position of the images in the testing dataset is classified. As the patches are shifted with 2 pixel strides both in height and width dimension, each 2×2 block is classified by one-class SVM several times. The local abnormal event and the performance are shown in Figs. 9 and 10. The frame-level criterion only focuses on the detection of the frames which contain abnormal event, while the pixel-level criterion focuses on the location of the abnormal pixels. The result shows that the sparse reconstruction method can detect the frames where the abnormal event happens, while doesn't detect the real abnormal pixels in the ground truth. The accuracy of the location of the proposed method is higher than the sparse reconstruction method. Our proposed algorithm is compared with the state-of-the-art methods, such as Sparse reconstruction cost for abnormal event detection, work of Adam, MPPCA with social force model, social force model and MDT. The comparison results are presented in Table 5 with different evaluation criteria. Our proposed RoI covariance matrix feature and the abnormal detection algorithm outperforms the state-of-the-art algorithms.

Table 5 Quantitative comparison of our method with the state-of-the-art methods

Method	Evaluation criteria		
	EER	RD	AUC
Sparse [10]	19 %	46 %	46.1 %
Adam [1]	38 %	24 %	13.3 %
MPPCA+SF [30]	32 %	27 %	21.3 %
SF [30]	31 %	21 %	17.9 %
MPPCA [30]	40 %	18 %	20.5 %
MDT [26]	25 %	45 %	44.1 %
ours	26 %	63 %	65.8 %

EER is equal error rate of frame level criterion. RD is rate of detection of pixel level criterion. AUC is the area under ROC of pixel level criterion

6 Conclusions

In this paper, we propose the multi-RoI covariance (MCOV) descriptor fusing optical flow based feature of multiple regions of interest. As the proposed feature takes several regions as an integer unit, this feature can tolerate more frame fluctuation. Because the feature is in a Lie Group, a suitable kernel function is proposed to handle the feature descriptor with the nonlinear one-class SVM to detect abnormal events by merely training the normal samples. The experiments have undergone the benchmark datasets, i.e. PETS, UMN and UCSD, proving the competitive property of our proposed method. We are going to test on more live environments in the future. The action recognition problem and the multimedia event detection problem based on our current research are also in the prospective research scopes.

Acknowledgements This work is partially supported by the National Natural Science Foundation of China (61503017, U1435220, 61365003), the Aeronautical Science Foundation of China (2016ZC51022), Gansu Province Basic Research Innovation Group Project (1506RJIA031), the Fundamental Research Funds for the Central Universities (YWF-14-RSC-102), the ANR AutoFerm project and the Platform CAPSEC funded by Région Champagne-Ardenne and FEDER.

References

- Adam A, Rivlin E, Shimshoni I, Reinitz D (2008) Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans Pattern Anal Mach Intell* 30(3):555–560
- Benezeth Y, Jodoin PM, Saligrama V (2011) Abnormality detection using low-level co-occurring events. *Pattern Recogn Lett* 32(3):423–431
- Bhatnagar G, Wu QJ, Raman B (2013) Discrete fractional wavelet transform and its application to multiple encryption. *Inf Sci* 223:297–316
- Bianco S, Ciocca G, Schettini R (2015) How far can you get by combining change detection algorithms? [arXiv:1505.02921](https://arxiv.org/abs/1505.02921)
- Bojanowski P, Bach F, Laptev I, Ponce J, Schmid C, Sivic J et al (2013) Finding actors and actions in movies. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*
- Burton A, Radford J (1978) *Thinking in perspective: critical essays in the study of thought processes*. Methuen
- Canu S, Grandvalet Y, Guigue V, Rakotomamonjy A (2005) *Svm and kernel methods matlab toolbox*. Perception Systèmes et Information. INSA de Rouen, Rouen
- Chen C, Ren Y, Kuo CCJ (2014) Large-scale indoor/outdoor image classification via expert decision fusion (edf). In: *Asian Conference on Computer Vision (ACCV)*. Springer, Berlin, pp 426–442
- Cheng MM, Zhang Z, Lin WY, Torr P (2014) Bing: Binarized normed gradients for objectness estimation at 300fps. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 3286–3293
- Cong Y, Yuan J, Liu J (2011) Sparse reconstruction cost for abnormal event detection. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 3449–3456
- Cricri F, Roininen MJ, Leppanen J, Mate S, Curcio ID, Uhlmann S, Gabbouj M (2014) Sport type classification of mobile videos. *IEEE Trans Multimed* 16(4):917–932
- Cui P, Wang F, Sun LF, Zhang JW, Yang S (2012) A matrix-based approach to unsupervised human action categorization. *IEEE Trans Multimed* 14(1):102–110
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 886–893, vol 1. IEEE, USA
- Ergezer H, Leblebicioğlu K (2016) Anomaly detection and activity perception using covariance descriptor for trajectories. In: *European Conference on Computer Vision (ECCV)*. Springer, Berlin, pp 728–742
- Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. *IEEE Trans Pattern Anal Mach Intell* 29(12):2247–2253
- Hall B (2003) *Lie groups, Lie algebras and representations: an elementary introduction*, vol 222. Springer, Berlin

17. Harandi M, Salzmann M, Porikli F (2014) Bregman divergences for infinite dimensional covariance matrices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1003–1010
18. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 733–742
19. Henriques JF, Caseiro R, Martins P, Batista J (2012) Exploiting the circulant structure of tracking-by-detection with kernels. In: European Conference on Computer Vision. Springer, Berlin, pp 702–715
20. Horn BK, Schunck BG (1981) Determining optical flow. *Artif Intell* 17(1):185–203
21. Hussein ME, Torki M, Gawayyed MA, El-Saban M (2013) Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), vol 13, pp 2466–2472
22. Jiménez-Hernández H, González-Barbosa JJ, García-Ramírez T (2010) Detecting abnormal vehicular dynamics at intersections based on an unsupervised learning approach and a stochastic model. *Sensors* 10(8):7576–7601
23. Kalal Z, Matas J, Mikolajczyk K (2010) Pn learning: Bootstrapping binary classifiers by structural constraints. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 49–56
24. Kosmopoulos D, Chatzis SP (2010) Robust visual behavior recognition. *IEEE Signal Process Mag* 27(5):34–45
25. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1–8
26. Li W, Mahadevan V, Vasconcelos N (2014) Anomaly detection and localization in crowded scenes. *IEEE Trans Pattern Anal Mach Intell* 36(1):18–32
27. Li H, Lin Z, Shen X, Brandt J, Hua G (2015) A convolutional neural network cascade for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5325–5334
28. Liu Y, Pados DA (2016) Compressed-sensed-domain l 1-pca video surveillance. *IEEE Trans Multimed* 18(3):351–363
29. Lucas BD, Kanade T et al (1981) An iterative image registration technique with an application to stereo vision
30. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1975–1981
31. Mazloom M, Li X, Snoek C (2016) Tagbook: A semantic video representation without supervision for event detection. *IEEE Trans Multimed* 18(7):1378–1388
32. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, pp 935–942
33. Orozco J, Martinez B, Pantic M (2015) Empirical analysis of cascade deformable models for multi-view face detection. *Image Vis Comput* 42:47–61
34. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: British Machine Vision Conference, vol 1, p 6
35. PETS (2009) Performance evaluation of tracking and surveillance (pets) 2009 benchmark data. multi-sensor sequences containing different crowd activities. <http://www.cvg.rdg.ac.uk/pets2009/a.html>
36. Piciarelli C, Micheloni C, Foresti GL (2008) Trajectory-based anomalous event detection. *IEEE Trans Circ Syst Video Technol* 18(11):1544–1554
37. Porikli F, Tuzel O (2005) Bayesian background modeling for foreground detection. In: Proceedings of the third ACM international workshop on Video surveillance & sensor networks (VSSN), pp 55–58
38. Rosani A, Boato G, De Natale FG (2015) Eventmask: A game-based framework for event-saliency identification in images. *IEEE Trans Multimed* 17(8):1359–1371
39. Rowley HA, Baluja S, Kanade T (1998) Neural network-based face detection. *IEEE Trans Pattern Anal Mach Intell* 20(1):23–38
40. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the support of a high-dimensional distribution. *Neurals Comput* 13(7):1443–1471
41. Shi Y, Gao Y, Wang R (2010) Real-time abnormal event detection in complicated scenes. In: Proceedings of International Conference on Pattern Recognition (ICPR), Istanbul, pp 3653–3656
42. Stauffer C, Grimson WEL (1999) Adaptive background mixture models for real-time tracking. In: 1999. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 2. IEEE, Berlin
43. Sun C, Nevatia R (2013) Active: Activity concept transitions in video event classification. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp 913–920
44. Sun D, Roth S, Black MJ (2010) Secrets of optical flow estimation and their principles. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2432–2439

45. Sun Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification. pp 1988–1996
46. Tang K, Fei-Fei L, Koller D (2012) Learning latent temporal structure for complex event detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1250–1257
47. Turk MA, Pentland AP (1991) Face recognition using eigenfaces. In: 1991. Proceedings CVPR'91., IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, USA, pp 586–591
48. Tuzel O, Porikli F, Meer P (2006) Region covariance: A fast descriptor for detection and classification. In: Proceeding of European Conference on Computer Vision (ECCV). Springer, Berlin, pp 589–600
49. UCSD (2010) UCSD anomaly detection dataset, available from <http://www.svcl.ucsd.edu/projects/anomaly/dataset.html>
50. UMN (2006) Unusual crowd activity dataset of university of minnesota, department of computer science and engineering, <http://mha.cs.umn.edu/movies/crowd-activity-all.avi>
51. Utasi Á, Czúni L (2010) Detection of unusual optical flow patterns by multilevel hidden markov models. *Opt Eng* 49(1):017,201–017,201
52. Varadarajan J, Odobez JM (2009) Topic models for scene analysis and abnormality detection. In: Proceedings of the 12th International Conference on Computer Vision Workshops (ICCV Workshops), pp 1338–1345
53. Wang T, Snoussi H (2014) Detection of abnormal visual events via global optical flow orientation histogram. *IEEE Trans Inf Forensic Secur* 9(6):988–998
54. Wang T, Chen J, Zhou Y, Snoussi H (2013) Online least squares one-class support vector machines based abnormal visual event detection. *Sensors* 13(12):17130–17155
55. Wang F, Sun Z, Jiang YG, Ngo CW (2014) Video event detection using motion relativity and feature selection. *IEEE Trans Multimed* 16(5):1303–1315
56. Warren DH, Strelow ER (2013) Electronic spatial sensing for the blind: contributions from perception, rehabilitation, and computer vision, vol 99. Springer Science & Business Media, Berlin
57. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. *IEEE Trans Pattern Anal Mach Intell* 31(2):210–227
58. You X, Du L, Cheung Ym, Chen Q (2010) A blind watermarking scheme using new nontensor product wavelet filter banks. *IEEE Trans Image Process* 19(12):3271–3284
59. Zhang K, Zhang L, Yang MH (2014) Fast compressive tracking. *IEEE Trans Pattern Anal Mach Intell* 36(10):2002–2015
60. Zhang X, Yang S, Tang YY, Zhang W (2016) A thermodynamics-inspired feature for anomaly detection on crowd motions in surveillance videos. *Multimed Tools Appl* 75(14):8799–8826
61. Zhou S, Shen W, Zeng D, Fang M, Wei Y, Zhang Z (2016) Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Process Image Commun* 47:358–368



Tian Wang received the M.S. degree and Ph.D. degree from Xi'an Jiaotong University, China and University of Technology of Troyes, France in 2010 and 2014, respectively. He is an assistant professor at the School of Automation of Science and Electrical Engineering, Beihang University. His research interests include computer vision and pattern recognition.



Meina Qiao is a Master in School of Automation Science and Electrical Engineering in Beihang University. She is involved in abnormal events detection and video surveillance. Her academic interests are computer vision and machine learning.



Aichun Zhu received the M.S. degree in engineering from China University of Mining and Technology in 2012, and received the Ph.D. degree from the University of Technology of Troyes in 2016. He is currently an Assistant Professor in the School of Computer Science and Technology of Nanjing Tech University. He is involved in human pose estimation and motion recognition. His academic interests span computer vision and machine learning.



Yida Niu is currently studying in automation in Beihang University. He is now involved in electromechanical. His academic interests span video surveillance, automation and mechanical.



Ce Li received his Ph.D. degree in pattern recognition and intelligence system from Xi'an Jiaotong University, China in 2013. He is a professor at the College of Electrical and Information Engineering, Lanzhou University of Technology. His research interests include computer vision and pattern recognition.



Hichem Snoussi received his diploma from Ecole Supérieure d'Electricité (Supelec) in 2000, and his DEA and PhD degrees from the University of Paris-Sud in 2000 and 2003, respectively. From 2003 to 2004, he was a postdoctoral researcher with the Institut de Recherche en Communications et Cybernétiques de Nantes. Since 2010, he has been a full professor at the University of Technology of Troyes. His research interests include signal processing, computer vision and machine learning.