Multimed Tools Appl (2016) 75:14617–14639
DOI 10.1007/s11042-016-3316-3

Abnormal event detection in crowded scenes based on deep learning

Zhijun Fang[1,2] · **Fengchang Fei**[1,3] · **Yuming Fang**[1] · **Changhoon Lee**[4] · **Naixue Xiong**[1] · **Lei Shu**[1] · **Sheng Chen**[1]

Received: 26 February 2015 / Revised: 8 July 2015 / Accepted: 27 January 2016 /
Published online: 13 February 2016

Abstract In this paper, we propose to use the deep learning technique for abnormal event detection by extracting spatiotemporal features from video sequences. Human eyes are often attracted to abnormal events in video sequences, thus we firstly extract saliency information (SI) of video frames as the feature representation in the spatial domain. Optical flow (OF) is estimated as an important feature of video sequences in the temporal domain. To extract the accurate motion information, multi-scale histogram optical flow (MHOF) can be obtained through OF. We combine MHOF and SI into the spatiotemporal features of video frames. Finally a deep learning network, PCANet, is adopted to extract high-level features for abnormal event detection. Experimental results show that the proposed abnormal event detection method can obtain much better performance than the existing ones on the public video database.

Keywords Abnormal event detection · Crowd analysis · Saliency information · Optical flow · Deep learning

✉ Yuming Fang
fa0001ng@e.ntu.edu.sg

Naixue Xiong
nxiong@coloradotech.edu

[1] School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330032, China

[2] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

[3] Modern Economics & Management College, Jiangxi University of Finance and Economics, Nanchang 330013, China

[4] Department of Computational Science and Engineering, Seoul National University of Science and Technology, Seoul, South Korea

1 Introduction

Today, more and more video surveillance systems emerge in public places, such as traffic junctions, airports, railway stations, etc. With low hardware costs, communities and families are equipped with surveillance cameras. However, most of these monitoring facilities are just used as a record equipment. In most monitoring centers, only several people are in charge of the monitoring equipment by focusing on the monitor screens and analyze humans' behaviors to find abnormal events in time. This type of manual monitoring systems is not only likely to cost a lot of manpower, but also easy to miss abnormal events. Therefore, intelligent monitoring systems are much desired. The intelligent monitoring system can monitor visual field. It can detect and analyze abnormal events automatically. Compared to traditional video surveillance systems, intelligent monitoring systems can greatly reduce labor costs with 24-h monitoring, and help prevent some unexpected dangerous events in advance, or prevent further expansion of abnormal events.

Abnormal event detection in crowded scenes is an important and challenging task in intelligent surveillance video systems. Abnormal event detection refers to detecting and responding to the abnormal changes or behaviors of humans or objects in videos. Currently, there are various abnormal detection algorithms proposed in crowded scenes, such as abnormal crowd behavior detection [9, 46, 53], human abnormal action detection [38], traffic incident detection [50], etc.

The study [27] proposes the social force model, which uses interaction force between pedestrians as feature for abnormal event detection, and the result is good. But [5] proposes the multi-scale histogram of optical flow (MHOF) feature based on sparse representation is better than [27]. MHOF is composed of optical flow (OF) [30] and improved. But these features are extracted from the pixel value of video, without considering the characters of human eyes and human visual system. This paper proposes using human attention as the feature of video for abnormal event detection. Deep learning algorithm is a learning network which simulates human brain, and it is widely used in the recognition of complex objects. In the field of abnormal event detection, deep learning was not considered to use to extract high-level features. But in order to link up the human visual system, this paper uses deep learning to abstract the high-level of human attention. It simulates the process of human visual perception and human brain learning.

In this study, we propose a new abnormal event detection model based on the features from saliency information and MHOF. A visual attention model is adopted to extract the saliency features for dynamic scenes. Furthermore, MHOF is used to extract motion features for video sequences. After the saliency and motion features are computed, a deep learning technique is used for abnormal event detection. Experimental results on a public database have demonstrated the promising performance of the proposed method.

The reminder of this paper is organized as follows. Section 2 introduces the related work in the literature. The proposed method is described in Section 3. Section 4 provides the comparison experiments to demonstrate the advantages of the proposed method. The final section gives a conclusion of the study.

2 Related work

2.1 Abnormal event detection

The data mining and machine learning techniques are widely used in intelligent video surveillance systems [28]. Now many researchers are trying to explore how to design an effective algorithm or model for real-time video data in order to accurately detect abnormal events for intelligent monitoring.

Abnormal event detection refers to detecting and responding to the abnormal changes in videos, as is shown in Fig. 1. Currently, there are various abnormal detection algorithms proposed in crowded scenes.

One type of abnormal event detection methods is designed based on object detection and tracking [18, 23, 29]. There are three steps in these methods: (1) Detection and tracking: detecting objects in video frames and tracking them. (2) Motion extraction [9, 46, 50, 53]: recording the trajectory of the moving objects. (3) Activity analysis [38]. Usually, the tracking algorithms include object detection. The study [30] presents an object detection method based on sparse representation and Histograms of Sparse Codes (HSC), which can obtain better performance in object detection than histograms of oriented gradients (HOG). There are a lot of studies trying to propose effective tracking algorithms. In the study [52], the authors proposed a model-free tracker for object tracking. In order to track people and their baggage, the authors in the study [5] built up 3D shape models to obtain the stereo depth information. For long-term tracking, self-paced learning algorithm was proposed in the study [37]. The authors in the study [22] designed a Robust linear regression algorithm for object tracking.

The other type of object detection methods is implemented by extracting features from video in the spatiotemporal domain [6, 9, 39]. In spatiotemporal domain, the global and local features are combined for the final abnormal event detection. The local detection is implemented by the differences of the target and its surrounding area. The global feature is extracted by analyzing the visual scene globally to determine whether there is abnormal event occurring. Various features and models are used in this type of method, such as mixtures of dynamic textures [24], global cues [25], and social force model [27]. In the study [11], the authors proposed an Interaction Energy Potentials function which is the representation between the current behavior state of a subject and its actions. And other features have been proposed, such

Fig. 1 Examples of the normal and abnormal frames. **a** A normal frame. **b** An abnormal frame

as: Streak line representation of flow [26], optical flow [47], histogram of optical flow (HOF) [2], multi-scale histogram of optical flow (MHOF) [10], global optical flow orientation histogram [48], social attribute-aware force model [51], and so on. Based on the features described above, some abnormal event detection algorithms were proposed based on machine learning techniques. In the study [4], the authors used optical flow to design an event detection model by adopting the optimal number of models to represent normal crowd behavior. The study [27] uses Latent Dirichlet Allocation (LDA) to detect abnormal events. In the study [45], the authors used probabilistic Latent Semantic Analysis (pLSA) to detect abnormality. In the study [34], the authors applied a Superpixel-based Bag-of-Words (BoW) model to build an event detector.

2.2 Saliency information extraction

Human visual system (HVS) is a mechanism to realize the external world projection in the brain. In recent years, the simulation applications based on HVS are more and more [31–33]. When people watch video, the attention of human eyes can be easily attracted by the abnormal events and behavior appearing in the video frames. Thus, in this paper, we extract the saliency information as one feature for abnormal event detection. Visual attention is an important mechanism of the human visual system (HVS). There are two categories of human visual attention mechanism: top-down [3, 49] and bottom-up [8, 12, 14, 15, 40, 42]. Top-down is influenced by prior knowledge, such as tasks with the purpose, the distribution of target characteristics, the context of visual scene, etc. On the contrary, bottom-up is a spontaneous choice of saliency area by an image and it is the main research direction of visual attention. In this paper, we study visual attention based on bottom-up too.

Treisman developed the Feature-Integration Theory (FIT) in 1980 [40]. When an observer looks at a scene, his visual attention can be attracted easily by some low-level features such as color, intensity and contrast, and he would pay attention to the salient areas in the scene. These areas are considered as salient areas in video frames and they can be computed by the differences between center and surroundings. In the past decades, various visual attention models were designed based on FIT theory [1, 14, 15, 42]. The salient area detection methods calculate the saliency map by computing the differences between the center area and surrounding ones. Achanta et al. designed a saliency detection model by using frequency domain information of images [1]. Guo et al. proposed a saliency detection model of visual scenes based on Fourier transform [15, 16, 20]. In the study [12], the authors proposed a saliency detection algorithm by using quaternion Fourier Transform (QFT) instead of Fourier transform (FT). It uses Amplitude Spectrum of QFT to represent the color, contrast and brightness, and obtain the final saliency map through the weight. As is showed in experimental results, it outperforms the state-of-the-art detection models. In this paper, we extract the saliency map of video frames by using the saliency detection method in [12] to simulate the visual attention and treats it as a spatial feature.

2.3 Multi-scale histogram of optical flow

Generally, the abnormal events occur in consecutive frames. Abnormal detection algorithm would consider the saliency value of every frame and its continuity as well. Optical flow is represented by 2D instantaneous velocity of all pixels in an image. The 2D velocity vector is the point of the 3D velocity vector in the projection imaging surface. Thus, optical flow not

only includes the motion information of the observed object, but also contains the information of the 3D structure in the scene. Each pixel (i, j) has a 2D velocity vector $(d^x_{i,j}, d^y_{i,j})$. If the abnormal event detection task takes optical flow of each pixel as a feature, the computational complexity will be high, and the pixel noise also influences the results. To obtain better performance, multi-scale histogram of optical flow (MHOF) is proposed in [10], which first divides each video frame into small image patches, and then classifies each patch into 16 classes. By using the histogram of 16 classes as the patch features instead of optical flow, this method greatly reduces the computation complexity, and achieves the effect of suppressing the noise in optical flow.

MHOF preserves more precise motion information than the traditional histogram of optical flow (HOF). In the study [10], MHOF can better describe the current frame scene changes, and thus detects abnormal events in video sequences accurately.

The proposed MHOF framework for each block is shown in Fig. 2. First of all, every video frame is divided into image patches with the same size M. Then, the optical flow matrix $D_i(D^x, D^y)$ of each patch is computed, and so is the MHOF of each patch. The following Eqs. 1 and 2 are used to calculate the class-label of each pixel $class_{i,j}$.

$$C_{i,j} \in \begin{cases} 0 & d^x_{i,j}, d^y_{i,j} \le Th \\ 1 & d^x_{i,j}, d^y_{i,j} > Th \end{cases} \tag{1}$$

$$class_{i,j} = round\left(\theta\left(d^x_{i,j}, d^y_{i,j}\right) \Big/ \left(\pi \Big/ 4\right)\right) + 8 \times C_{i,j} \tag{2}$$

where $(d^x_{i,j}, d^y_{i,j})$ is the optical flow of each pixel; Th is the magnitude threshold; θ is the angle of $d^x_{i,j}, d^y_{i,j}$.

2.4 Deep learning

The features in spatial and temporal domains are manually selected, such as color, intensity, contrast and so on. These are low-level features. The human brain often abstracts high-level features from low-level ones, such as shape, depth, movement, etc., and these high-level features can be better perceived by human brain. Deep Learning is highly correlated to AI (Artificial Intelligence) in the field of machine learning. There are three properties in Deep Learning: (1) Learning is unsupervised in each layer; (2) In each layer, training data is unsupervised learning, and the results are used as the input of higher layer; (3) supervised learning is employed to adjust all layers. Deep learning is a widely used in the research area of machine learning. The motivation is simulating the human brain to establish neural network for analysis and learning. In the application of image processing, deep learning is used to discover multiple-level and high-level features for image representation. Thus, the classification tasks no longer depend only on low level features manually. An early study of deep learning was conducted by Hinton et al. [19]. Recently, deep learning has been widely used in the research area of computer vision [13, 21, 36], but generally this model is mainly applied to the detection of complex objects, such as face detection [35]. Because the amount of computation of deep learning is large, existing studies on abnormal event detection based on deep learning are rare. In this paper, we use a very simple deep learning method-PCANet to simulate human brain to abstract high-level feature from low-level feature of video.

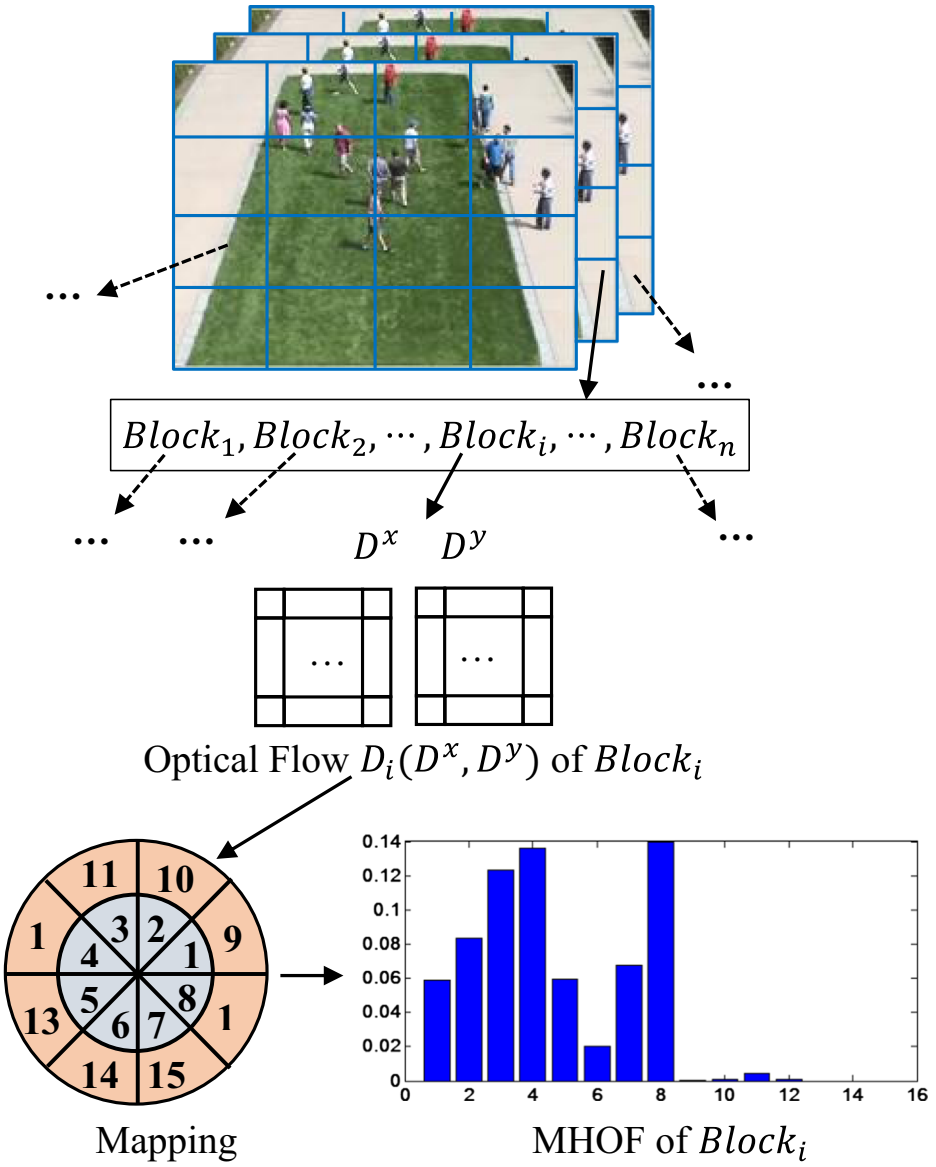


Fig. 2 The framework of MHOF

2.5 SVM

Now the SVM + HOG algorithm has become the mainstream architecture for pedestrian tracking. This paper also uses the SVM as the classifier for abnormal event detection. Support vector machine (SVM) is widely used for statistical classification and regression analysis in various applications. The SVM includes a support vector classifier and support vector regressor. SVM was firstly introduced in 1996 by Vapnik [43, 44], and is a kind of analysis method based on statistical learning theory. It requires very few samples for training,

not sensitive to the number of attributes. SVM not only brings good training data classification effect, but also has good test accuracy on the test data with the same characteristics. In the past 20 years, SVM theory and application have been developed quickly.

For a two-class separable learning task, the samples are mapped into a high dimensional space, and in this space, a hyper-plane will classify the samples into two classes. To obtain promising classification performance on the data, the optimal hyper-plane should be selected. Thus, two-side planes are set up and they are parallel to the hyper-plane and have the same distances to the hyper-plane. The distance between the two-side planes is called margin. The hyper-plane is in the middle of these two-side planes. The larger the margin, the higher the accuracy of the classifier is. Thus the purpose of the SVM algorism is to obtain the maximum marginal hyper-plane.

There is an example of two-class linearly separable learning task in Fig. 3. There are many sample points in a 2-dimension Descartes coordinates and there are two coordinate values for each sample point x and y. In 2-D space, the representation of the hyper-plane is a straight line. These sample points have two classes: positive and negative. There is a hyper-plane in Fig. 3a, b. The aim of SVM is to find a maximum marginal hyper-plane (MMH), and thus the hyper-plane of Fig. 3a is the final result of SVM. For calculation of the MMH, the classification function can be described as follows:

$$f(x) = \boldsymbol{\omega}^T \boldsymbol{x} + b \tag{3}$$

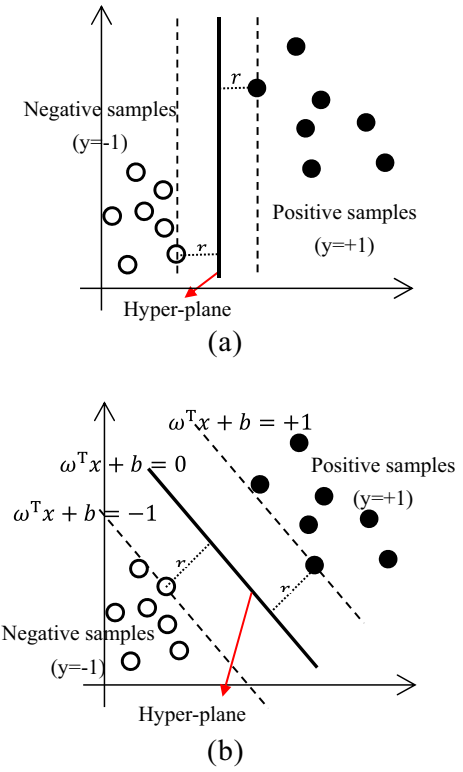


Fig. 3 The samples of the hyper-plane for a linearly separable case

where ω is the normal vector of MMH, $\{(x, y)\}$ is the sample set. The hyper-plane can be defined as

$$\omega^T x + b = 0 \tag{4}$$

According to the point to plane distance formula, the following Eq. can be obtained:

$$r = \frac{\omega^T x + b}{\omega} = \frac{f(x)}{\omega} \tag{5}$$

Suppose r is the distance between a side-plant and the hyper-plane, then the expressions of the two side-plane is

$$\begin{cases} \omega^T x + b = -k \\ \omega^T x + b = +k \end{cases} \tag{6}$$

Normalize k, then the above expression is

$$\begin{cases} \omega^T x + b = -1 \\ \omega^T x + b = +1 \end{cases} \tag{7}$$

Figure 3a, b illustrate the three planes.Then the sample point sequence $\{(x, y)\}$ should follow the following formula:

$$\begin{cases} \omega^T x_i + b \geq 1 \\ \omega^T x_i + b \leq 1 \end{cases} \tag{8}$$

where $(x_i, y_i) \in \{(x, y)\}$, and the sample points (x^*, y^*) for which the equalities of Eq. 8 are satisfied, these points are called support vectors, which are corresponding to r^*

$$r^* = \frac{\omega^T x^* + b}{\omega} = \frac{f(x^*)}{\omega} = \begin{cases} \frac{1}{\omega} \text{ if } y^* = +1 \\ -\frac{1}{\omega} \text{ if } y^* = -1 \end{cases} \tag{9}$$

The distance between these two side-planes d is

$$d = 2r^* = \frac{2}{\omega} \tag{10}$$

When d is with the maximum value, the hyper-plane is the optimal hyper-plane (MMH). Thus, d should be maximized with respect to ω and b.

$$\begin{aligned} &\max(d) = \max\left(\frac{2}{\omega}\right) \\ &s.t.\ y_i\left(\omega^T x_i + b\right) \geq 1, i = 1, 2 \ldots, n \end{aligned} \tag{11}$$

The final solution of ω and b construct the MMH classifier, namely SVM classifier. Now there are a lot of improved algorithms of SVM. The study [47] proposes a one –class classifier based on SVM–online least squares one-class SVM (online LS-OC-SVM), and then detect abnormal events using this method based on MHOF feature.

3 The proposed method

3.1 Saliency information extraction in video

According to selective attention mechanism, human eyes can quickly and effectively focus on important events in complex scenes. Human vision can always pay attention to the salient areas, which are different from their neighboring areas. Generally, the abnormal events existing in video can be represented by the sudden change in spatial or temporal dimension in video frames. In this paper, we present an abnormal event detection algorithm based on saliency information of video frames. Here the saliency detection model in [12] is used to extract the salient areas in each video frame. The frame is firstly divided into image patches, and then the saliency value of each patch is determined according to the difference between one patch and all the other patches from the features of color, intensity, and orientation, which is described as follows:

$$S_i = \sum_{i \neq j} w_{i,j} D_{i,j} \tag{12}$$

where S_i is the saliency value of image patch i, $D_{i,j}$ is the difference between patches i and j, $w_{i,j}$ is the corresponding weight, determined by human visual sensitivity. $D_{i,j}$ is represented by the differences between amplitude spectrum of QFT of patches i and j. The framework is showed in Fig. 4. Figure 5 shows a normal and an abnormal frames and their corresponding saliency information.

3.2 Feature representation of video frames based on PCANet

Deep learning can be used to abstract high-level features from low-level features. We use a simple deep learning technique to dig out MHOF and SI features for further feature

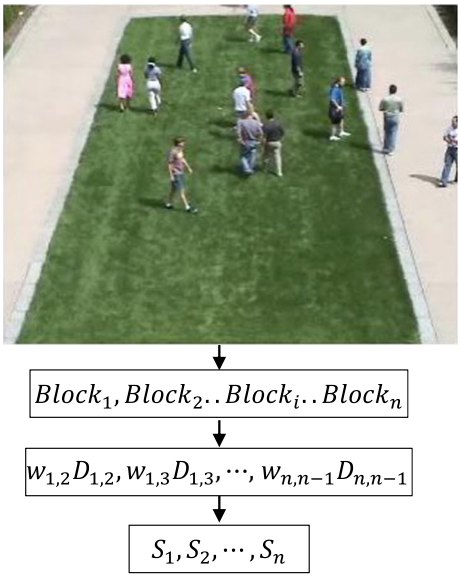


Fig. 4 Framework of the saliency information extraction of video sequences

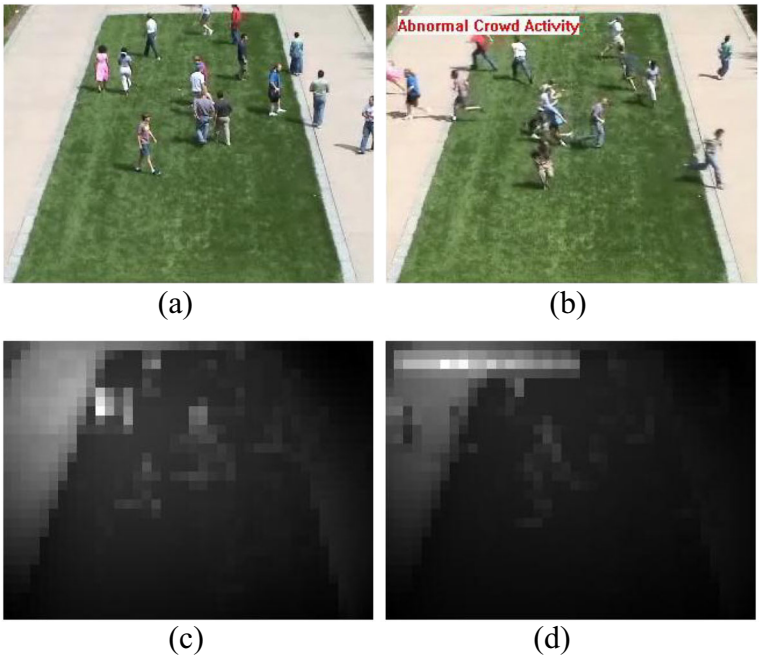


Fig. 5 **a** Normal frame. **b** Abnormal frame. **c** Saliency map of frame (a). **d** Saliency map of frame (b)

representation. In [7], the deep learning algorithm of PCANet is proposed by constructing a simple and robust model. In this model, the representation of multi-level features can be discovered by using simple PCA to carry on the extraction of high-level feature.

Generally, there are three steps in PCANet: the high-level features of training video can be extracted by filters in steps 1 and 2. These filters are obtained by multilayer principal component analysis. The framework of PCANet is shown in Fig. 6.

Suppose there is an video sequence $S(I_1, I_2, \cdots, I_t, \cdots, I_N)$, and the resolution of S is nr × nc, N is the number of frames in video S. The size of filters in step1 and step2 is $k1 \times k2$. The algorithm can be summarized as follows:

Step 1 We extract $k1 \times k2$ pixels around each pixel in each frame to obtain $(nr - k1 + 1) \times (nc - k2 + 1)$ image patches.

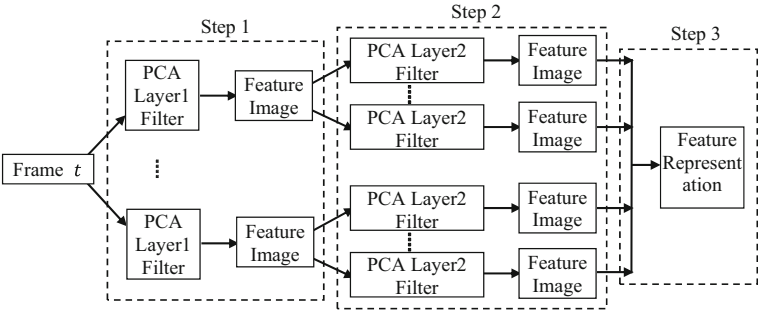


Fig. 6 The framework of PCANet

For example, assume $P_i(p_1, p_2, \cdots, p_j, \cdots, p_{(nr-k1+1)\times(nc-k2+1)})$ is the block set of frame I_i, and j is the number of patch in I_i. Then each patch is transformed into a vector $v \in \mathbb{R}^{(k1\times k2,\ 1)}$. A huge matrix $M \in \mathbb{R}^{(k1\times k2,\ (nr-k1+1)\times(nc-k2+1)\times N)}$ can be obtained and it is the reconstruction of the video S. The eigenvalue $\lambda \in \mathbb{R}^{(k1\times k2,\ 1)}$ and eigenvector $V \in \mathbb{R}^{(k1\times k2,\ k1\times k2)}$ of M are computed, in which K eigenvectors from large to small in V are selected to construct $F1 \in \mathbb{R}^{(k1\times k2,\ K)}$. Then we transform $F1$ to $F_1(f_1, f_2, \cdots f_h, \cdots, f_K)$, where $f_h \in \mathbb{R}^{(k1,\ k2)}$.

Step 2 K feature images of each frame can be obtained by filter-set $F1$. So there are $N \times K$ feature images through the convolution of layer1 filters. And then the filter-set $F2$ can be computed by repeating Step 1.

Step 3 Based on $F1$ and $F2$, K^2 redundant feature images $Feature_t \in \mathbb{R}^{(nr,nc,K^2)}$ can be obtained and then convert $Feature_t$ to binary images. The representation image of T_t can be computed as follows:

$$T_t^l = \sum_{h=1}^{K} 2^{i-1} H\left(I_t^{l-1} * f_h^l\right) \tag{13}$$

where l is the level of PCANet; f_h^l is the filter of layer l filter-set F_l

At last, by dividing T_t to patches, we compute the histogram of each patch, which is the final representation of image features in multiple levels. We can change the filter size to obtain different PCANet. It is not to say that the smaller filter size is better, since the local information affects the accuracy of feature extraction, and we will demonstrate thesein the following experiments.

3.3 Spatiotemporal abnormal event detection model based on PCANet

Saliency information (SI) is extracted based on the characteristics of human perception, and represents the important information in visual scenes. Optical flow is the velocity vector of pixels, and MHOF can be used to extract temporal features of video frames and PCANet simulate the human brain. In this paper, SI and MHOF are combined to build a spatiotemporal model of abnormal detection (SI + MHOF model) by using the PCANet deep learning model and the framework is given in Fig. 7.

The proposed method for abnormal event detection can be described as follows.

Step 1 Dividing each frame of training video into m image patches.

Step 2 Based on Eq. 12, the saliency value of training video $S_{train}(S_1, S_2, \cdots, S_i, \cdots, S_n)$ can be obtained; where n is the number of frames in training video, $S_i(s_1, s_2, \cdots, s_j, \cdots s_m)$ is the saliency matrix of frame i, and s_j is the saliency value of patch j

Step 3 According to Fig. 2 and Eqs. 1 and 2, the MHOF, $\boldsymbol{H}_{train}(H_1, H_2, \cdots, H_i, \cdots, H_n)$, of training video can be obtained. $H_i = (h_1, h_2, \cdots, h_j, \cdots h_m)$ is the MHOF of frame i and the MHOF of patch j is h_j

Step 4 Taking (S_i, H_i) as the features of frame i, we can get the training data sequence $Data_{train}((S_1, H_1), (S_2, H_2), \cdots, (S_i, H_i), \cdots, (S_n, H_n))$.

Step 5 Use PCANet to transform $Data_{train}$ to $\_Data_{train}$

Step 6 Training the $Sparse\_Data_{train}$ with SVM to obtain the corresponding SVM model.

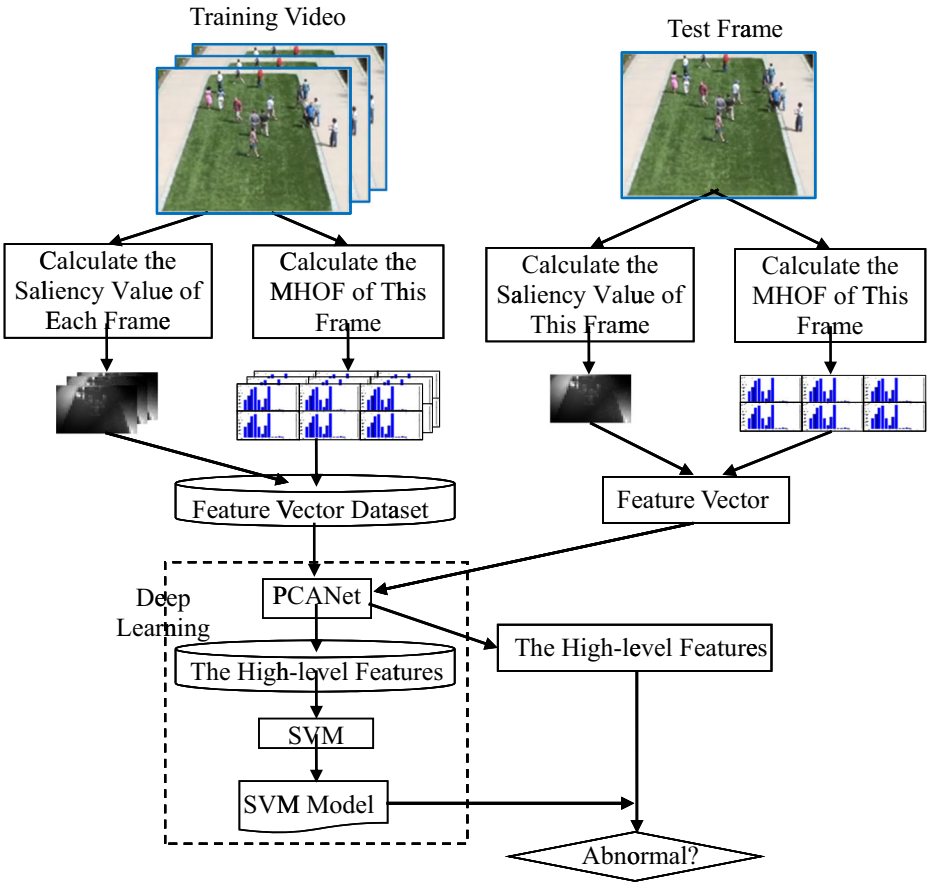


Fig. 7 The framework of the proposed SI + MHOF PCANet model

Step 7 According to step1 ~ step4, $Data_{test}(S_k, H_k)$, the feature vector of each test frame can be computed, and we can obtain $Sparse\_Data_{test}$ by PCANet.

Step 8 Detecting $Sparse\_Data_{train}$ by using the trained SVM model to determine whether the test frame is abnormal.

4 Experiments

We use the UMN dataset [41] to conduct the comparison experiment to demonstrate the performance of the proposed method.

In the experiment, we compare the optical flow (OF) based, multi-scale histogram of optical flow (MHOF) based, saliency information (SI) based, SI and MHOF (SI + MHOF) based algorithms in abnormal event detection. In addition, the results by these algorithms in use of PCANet and without using PCANet are also provided. In [10], each frame is divided into 20 image patches and 320 features are extracted in total. In order to ensure the consistent feature dimension, in OF and SI based algorithms, each frame is divided into 320 image patches and thus 320 values can be obtained from each frame. In

the learning process, we randomly select training frames from the video footage of each scene with a certain proportion, and the remaining frames are used as the test.

4.1 Evaluation criterion

In this paper, F -measure is used as the evaluation method. F -measure is computed by TP (True Positive is that the positive sample is correctly classified by the classifier), TN (True Negative is that the negative sample is correctly classified by the classifier), FP (True Positive is that the negative sample is incorrectly classified by the classifier), and FN (False Negative is that the positive sample is incorrectly classified by the classifier). Precision is the proportion of true positive in positive which is predicted by classifier. Recall is the proportion of true positive in real positive, as is shown in Eq. 14.

$$\begin{aligned} Precision &= \frac{True\ positive}{True\ positive + False\ positive} \\ Recall &= \frac{True\ positive}{True\ positive + False\ negative} = \frac{True\ positive}{positive} \end{aligned} \tag{14}$$

At large values of recall classifier, the number of false negative is low and thus the performance is better, and for the value of the precision, the larger the better. However, it is difficult to simultaneously guarantee that both these two values are high. Thus, it is challenging to build a classifier where the precision and recall are both largest [17].

In order to simultaneously ensure the precision and recall values, Precision and Recall can be combined into a single measure, F -measure as follows.

$$F_\beta = \frac{\left(1+\beta^2\right) \times precision \times recall}{\beta^2 \times precision + recall} \tag{15}$$

F_β -measure is the harmony of precision and recall. $\beta = 1$ denotes that the weights of precision and recall are the same. In the experiment, we use F_1 as the evaluation criteria.

4.2 Lawn scene

There are 1453 frames and 2 abnormal events in this video footage. The frames with large-change pedestrian motion are labeled as abnormal frames. The experimental results of this scene are shown in Table 1 and Fig. 8.

In order to demonstrate the advantages of the deep learning technique, we compare the results from the algorithms with and without deep learning based on MHOF, SI and SI + MHOF. In Table 1, the second to fourth columns show the results of PCANet with filter size 3×3; the fifth to seventh columns show the results of PCANet with filter size 5×5; the last four columns show the results from the algorithm without PCANet. Figure 8 shows the F_1 -measure values of different algorithms with different proportion of training samples.

From Table 1 and Fig. 8, we can see:

(1) Without PCANet, the results of MHOF are better than those of OF when the training sample percentage is less than 0.5. When the training sample percentage is larger than 0.6, the performance by MHOF will decrease. But the results of SI

Table 1 Experimental results of Lawn Scene

The training samples percentage	F_1 -measure									
	3×3PCAnet I + MHOF	3×3PCAnet SI	3×3PCAnet MHOF	5×5PCAnet SI + MHOF	5×5PCAnet SI	5×5PCAnet MHOF	SI + MHOF	SI	MHOF	OF
0.1	0.95841	0.96332	0.95733	0.97222	0.96644	0.96351	0.97666	0.97926	0.96586	0.96393
0.2	0.97710	0.97904	0.95766	0.98468	0.98134	0.98299	0.98604	0.98703	0.97773	0.97363
0.3	0.98767	0.99147	0.96415	0.99262	0.98966	0.98594	0.99010	0.99081	0.98118	0.97870
0.4	0.99613	0.99518	0.97482	0.99681	0.99495	0.99264	0.99533	0.99365	0.98819	0.98250
0.5	0.99699	0.99692	0.97905	0.99744	0.99624	0.99259	0.99703	0.99534	0.99004	0.98740
0.6	0.99757	0.99822	0.98459	0.99765	0.99803	0.99411	0.99736	0.99601	0.99108	0.99112
0.7	0.99738	0.99825	0.98564	0.99862	0.99812	0.99726	0.99825	0.99819	0.99237	0.99291
0.8	0.99868	0.99887	0.98680	0.99870	0.99794	0.99739	0.99850	0.99766	0.99475	0.99524
0.9	0.99629	0.99770	0.99224	0.99853	0.99926	0.99925	0.99831	0.99850	0.99512	0.99681

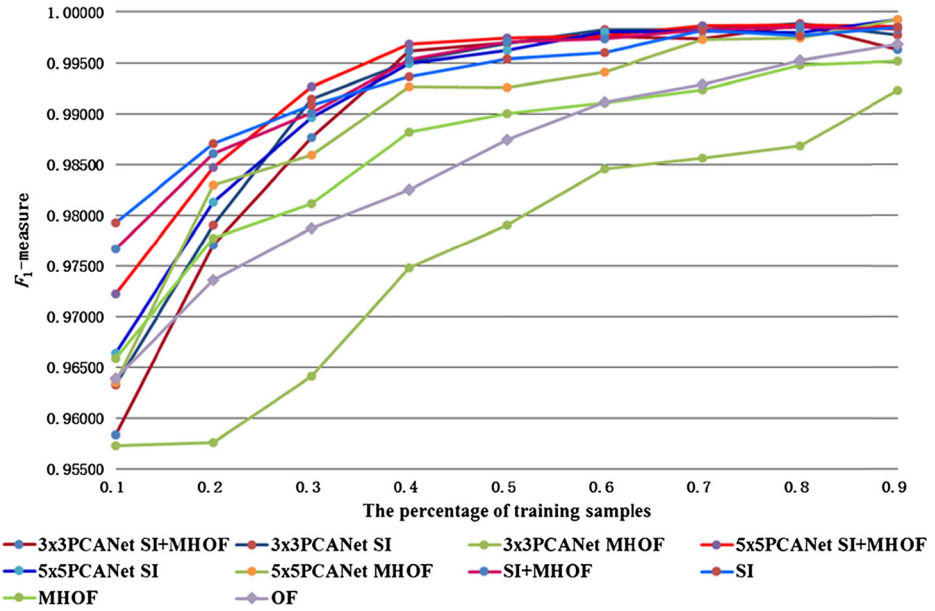


Fig. 8 Experimental results of Lawn Scene

and SI + MHOF are better than those of OF and MHOF, and when the training sample percentage is larger than 0.4, the results of SI + MHOF are better than those of SI. It is proved that saliency information is useful in detecting abnormal evens and spatiotemporal features (SI + MHOF) can be used to better detect abnormal evens in video sequences.

(2) The results of MHOF with PCANet of 5×5 filter size are better than those without using PCANet. However, the results of MHOF with 3×3 filter size PCANet are poor. The reason might be that the noise would influence the results with small filter sizes. Thus, for different features, the PCANet with corresponding sizes should be used.

(3) By using PCANet, the results of SI and SI + MHOF are improved, especially the results of SI + MHOF.

(4) When there are a few training samples, using PCANet cannot obtain good results. This is because deep learning net is a no feedback neural network. It can decompose the complex function relationship by the multi-layer simple function, thus it needs a lot of samples in training. When the training sample is small, the relationship between the multi layers cannot be accurately determined, and thus the experimental results would be not good enough.

4.3 Plaza scene

There are 2142 frames and 3 abnormal events in this video footage. The experimental results of this scene are shown in Table 2 and Fig. 9. From Table 2 and Fig. 9, we can have the following observations.

Table 2 Experimental results of Plaza Scene

The training samples percentage	F_1 -measure									
	3×3PCAnet SI + MHOF	3×3PCAnet SI	3×3PCAnet MHOF	5×5PCAnet SI + MHOF	5×5PCAnet SI	5×5PCAnet MHOF	SI + MHOF	SI	MHOF	OF
0.1	0.92911	0.93743	0.92729	0.94872	0.93929	0.93997	0.95591	0.95614	0.96696	0.96298
0.2	0.95553	0.95275	0.94711	0.96835	0.96444	0.96689	0.97913	0.97982	0.98626	0.96581
0.3	0.97614	0.96554	0.96665	0.98667	0.98253	0.98555	0.98771	0.98758	0.99022	0.97457
0.4	0.98394	0.98533	0.98237	0.99638	0.98582	0.99033	0.99332	0.99126	0.99284	0.97886
0.5	0.99375	0.99122	0.98807	0.99584	0.99534	0.99462	0.99525	0.99554	0.99520	0.98393
0.6	0.99672	0.99342	0.99128	0.99745	0.99671	0.99520	0.99565	0.99734	0.99469	0.98901
0.7	0.99645	0.99629	0.99513	0.99797	0.99731	0.99713	0.99755	0.99687	0.99637	0.99148
0.8	0.99847	0.99592	0.99671	0.99836	0.99670	0.99644	0.99826	0.99815	0.99797	0.99277
0.9	0.99950	0.99898	0.99847	0.99908	0.99949	0.99700	0.99904	0.99823	0.99823	0.99543

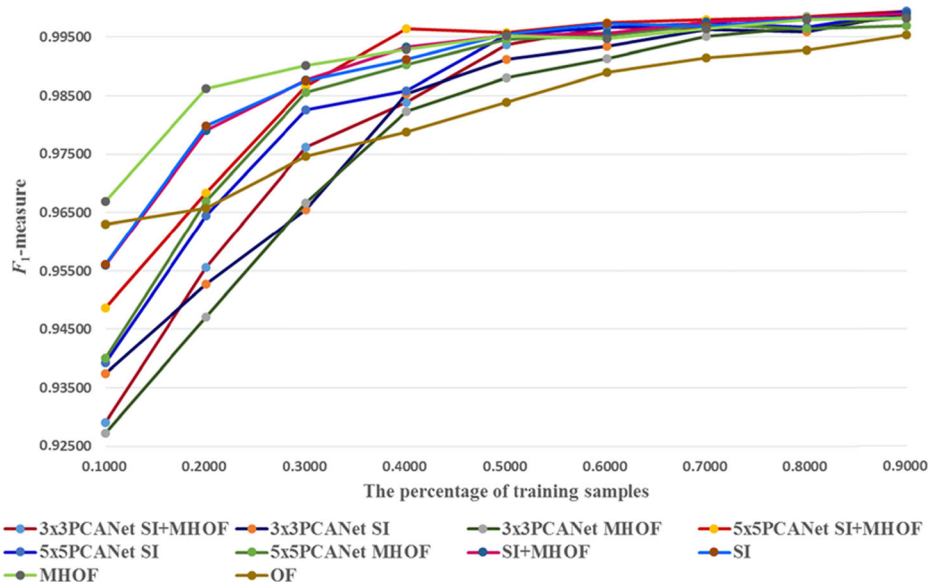


Fig. 9 Experimental results of Plaza Scene

(1) Without using PCANet, the performance of MHOF are better than that of OF, and when the training sample percentage is less than 0.4, the results of MHOF are the best among the compared algorithms. However, when the training sample percentage is larger than 0.5, the results of SI and SI + MHOF are better than those of MHOF.
(2) By using PCANet, the performance of MHOF and SI + MHOF improves, especially for the results of SI + MHOF.
(3) When the training sample percentage is less than 0.4, the results of MHOF are the best, as shown in the tenth column of Table 2; when the training sample percentage is greater than 0.4, the results of SI + MHOF based on PCANet are better than others.

From the above experimental results, we can conclude that:

(1) In abnormal event detection, without using PCANet, SI, MHOF, SI + MHOF are better than OF. For different scenes, both the features of MHOF and SI contribute to the abnormal event detection. With increasing training sample, the algorithm by SI + MHOF can obtain better performance than those by only MHOF or SI.
(2) For different video sequences, the suitable PCANet should be selected for abnormal event detection. With different sizes of filter in PCANet, the accuracy of abnormal event detection might be different.
(3) PCANet is able to extract better features from complex scenes. This also conforms to the original intention of deep learning.
(4) Because PCANet is a neural network with no feedback and unsupervised learning, PCANet model is more sensitive to the number of training samples during abnormal event detection.

5 Conclusion

In this paper, we propose to use the saliency information and MHOF to represent the features of the spatial domain and temporal domain in video sequences, respectively. The PCANet is adopted to simulate human brain to extract the high-level features from SI and MHOF for abnormal event detection. Experimental results demonstrate that the feature of SI + MHOF is better than only MHOF or SI in abnormal event detection, and the results of the proposed algorithm by using PCANet are better than that without using it. In the future, we will try to investigate how the deep learning techniques could further improve the performance of abnormal event detection.

Acknowledgments This research was supported partially by the National Natural Science Foundation of China (No. 61461021, 61571212), the Key Academic Leader Plan in Jiangxi Province (No. 20133BCB22005), the Key Project in Science and Technology from the Education Department of Jiangxi Province (No. GJJ14318) and the Foreign Cooperation Foundation from the Science and Technology Department of Jiangxi Province (No. 20151BDH80003, 20141BDH80003).

References

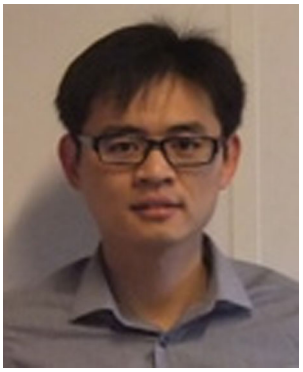
1. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned Salient Region Detection. 2009 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), 1597–1604
2. Adam A, Rivlin E, Shimshoni I (2008) Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans Pattern Anal Mach Intell 30:555–560
3. Alexe B, Deselaers T and Ferrari V (2010) What is an object? 2010 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), 73–80
4. Andrade EL, Blunsden S, Fisher RB (2006) Modelling crowd scenes for event detection. Pattern Recognition (CVPR), 175–178
5. Baumgartner T, Mitzel D, Leibe B (2013) Tracking people and their objects. IEEE Conf Comput Vis Patt Recog (CVPR) Oregon, USA 2013:3658–3665
6. Benezeth Y, Jodoin PM, Saligrama V (2009) Abnormal events detection based on spatio-temporal co-occurences. 2009 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), 2458–2465
7. Chan TH, Jia K, Gao S PCANet: A Simple Deep Learning Baseline for Image Classification? http://arxiv.org/abs/1404.3606, Acceptecd.
8. Cheng M, Warrell J, Lin W, Zheng S, Vineet V, Crook N (2013) Efficient salient region detection with soft image abstraction. 2013 I.E. International Conference on Computer Vision (ICCV) 1529–1536
9. Cho SH, Kang HB (2012) Integrated multiple behavior models for abnormal crowd behavior detection. IEEE Southwest Symposium on Image Analysis and Interpretation, 113–116
10. Cong Y, Yuan JS, Tang YD (2013) Video anomaly search in crowded scenes via spatio-temporal motion context. IEEE Trans Inf Forensics Secur 8:1590–1599
11. Cui X, Liu Q, Gao M (2011) Abnormal detection using interaction energy potentials. 2011 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), 3161–3167
12. Fang YM, Lin WS, Lee BS (2012) Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum. IEEE Trans Multimedia 14:187–198
13. Farabet C, Couprie C, Najman L (2013) Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell 35:1915–1929
14. Gopalakrishnan V, Hu Y, Rajan D (2009) Salient region detection by modeling distributions of color and orientation. IEEE Trans Multimedia 11:892–905
15. Guo C, Ma Q, Zhang L (2008) Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR) Oregon, USA 2008:1–8
16. Guo C, Zhang L (2010) A novel multi-resolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans Image Process 19:185–198
17. Han J, Kamber M, Pei J (2011) Data mining: concepts and techniques. Elsevier, USA
18. Hassner T, Itcher Y, Orit KG (2012) Violent flows: real-time detection of violent crowd behavior. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 1–6

19. Hinton GE, Osindero S, Teh Y (2006) A fast learning algorithm for deep belief nets. Neural Comput 18: 1527–1554
20. Hou X, Zhang L (2007) Saliency Detection: A Spectral Residual Approach. 2007 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), 1–8
21. Keyvanrad MA, Pezeshki M, Homayounpour MA (2014) Deep belief networks for image denoising. International Conference on Learning Representations, Accepted
22. Kwon J, Lee KM (2013) Minimum uncertainty gap for robust visual tracking. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR) Oregon, USA 2013:2355–2362
23. Lee YS, Chung WY (2012) Visual sensor based abnormal event detection with moving shadow removal in home healthcare applications. Sensors 12:573–584
24. Liu Y, Li Y, Ji X (2014) Abnormal event detection in nature settings. Int J Sign Proc Image Proc Patt Recog 7:115–126
25. Ma R, Li L, Huang W(2004) On pixel count based crowd density estimation for visual surveillance. IEEE Conference on Cybernetics and Intelligent Systems, 170–173
26. Mehran R, Moore EB, Shah M (2010) A streakline representation of flow in crowded scenes. 11th European Conference on Computer Vision, 439–452
27. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 20–25
28. Popoola OP, Wang K (2012) Video-based abnormal human behavior recognition-a review. IEEE Trans Sys Man Cybernet Soc 42:865–878
29. Rasheed N, Khan SA, Khalid A(2014) Tracking and abnormal behavior detection in video surveillance using optical. 28th International Conference on Advanced Information Networking and Applications Workshops, 61–66
30. Ren X, Ramanan D (2013) Histograms of sparse codes for object detection. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR) Oregon, USA 2013:3246–3253
31. Shao F, Lin W, Wang S, Jiang G, Yu M Blind image quality assessment for stereoscopic images using binocular guided quality lookup and visual codebook. IEEE Transactions on Broadcasting, accepted.
32. Shao F, Jiang G, Yu M, Li F, Peng Z, Fu R (2014) Binocular energy response based quality assessment of stereoscopic images. Digital Signal Process 29:45–53
33. Shao F, Lin W, Gu S et al (2013) Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics. IEEE Trans Image Process 22:1940–1953
34. Shu G, Dehghan A, Shah M (2013) Improving an object detector and extracting regions using superpixels. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR) Oregon, USA 2013:3721–3727
35. Sun Y, Wang X, Tang X (2013) Deep convolutional network cascade for facial point detection. 2013 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), 3476–3483.
36. Sun Y, Wang XG, Tang XO (2014) Deep learning face representation from predicting 10,000 classes. 2014 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), 1891–1898.
37. Supancic JS III, Ramanan D (2013) Self-paced learning for long-term tracking. Proc IEEE Conf Comput Vis Pattern Recognit Oregon, USA 2013:2379–2386
38. Suriani NS, Hussain A, Zulkifley MA (2013) Sudden event recognition: a survey. Sensors 13:9966–9998
39. Thida M, Eng HL, Remagnino P (2013) Laplacian eigenmap with temporal constraints for local abnormality detection in crowded scenes. IEEE Trans Cybernet 43:2147–2156
40. Treisman AM, Gelade A (1980) A feature-integration theory of attention. Cogn Psychol 12:97–136
41. University of Minnesota, Department of Computer Science and Engineering. http://mha.cs.umn.edu/proj_events.shtml, Accepted.
42. Valenti R, Sebe N, Gevers T (2009) Image saliency by isocentric curvedness and color. 2009 I.E. International Conference on Computer Vision (ICCV), 2185–2192
43. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
44. Vapnik VN (1998) Statistical learning theory. Wiley-Interscience Publication, USA
45. Varadarajan J, Odobez JM (2009) Topic models for scene analysis and abnormality detection. 2009 I.E. 12th International Conference on Computer Vision Workshops (ICCV Workshops), 1338–1345.
46. Wang (2012) Real-time detection of abnormal crowd behavior using a matrix approximation-based approach. IEEE International Conference on Image Processing, 2701–2704
47. Wang T, Chen J, Zhou Y (2013) Online least squares one-class support vector machines-based abnormal visual event detection. Sensors 13:17139–17155
48. Wang T, Snoussi HC (2014) Detection of abnormal visual events via global optical flow orientation histogram. IEEE Trans Inf Forensics Secur 9:988–998
49. Yang J, Yang M (2012) Top-down visual saliency via joint CRF and dictionary learning. 2012 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), 2296–2303

50. Yuan J (2011) Discriminative video pattern search for efficient action detection. IEEE Trans Pattern Anal Mach Intell 33:1728–1743
51. Zhang YH, Qin L, Ji RR Social attribute-aware force model: exploiting richness of interaction for abnormal crowd detection. IEEE Transactions on Circuits and Systems for Video Technology, Accepted.
52. Zhang L, Maaten L (2013) Structure preserving object tracking. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR) Oregon, USA 2013:1838–1845
53. Zhang YH, Qin L, Yao HX (2012) Abnormal crowd behavior detection based on social attribute-aware force model. 19th IEEE International Conference on Image Processing, 2689–2692

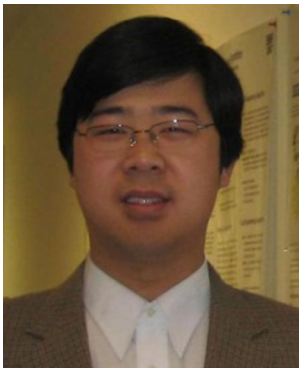
Zhijun Fang received the Ph.D. degree from Shanghai Jiaotong University, Shanghai, China. He is currently a professor and dean in the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, Jiangxi, China. His current research interests include image processing, video coding, and pattern recognition. Dr. Fang was a general chair of HHME2013 (the 9th Joint Conference on Harmonious Human Machine Environment). He received the awards of Jiangxi Provincial 'the GanPo Elite 555 Plan' and 'One-hundred, One-thousand, Ten-thousand Talent Project'

Fengchang Feiis a PhD candidate in the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include video processing and machine vision.

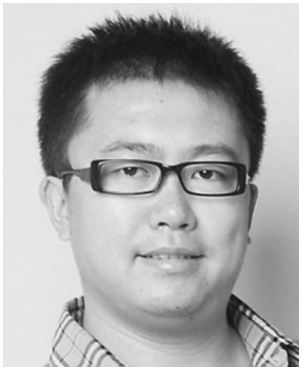


Yuming Fang is currently a faculty member in the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. He received the Ph.D. degree in Computer Engineering from Nanyang Technological University, Singapore in 2013. Previously, he obtained B.E. and M.S. from Sichuan University and Beijing University of Technology, China, respectively. From October 2011 to January 2012, he was a visiting Ph.D. student in National Tsinghua University, Taiwan. From September 2012 to December 2012, he was a visiting scholar in University of Waterloo, Canada. He was also a (visiting) Postdoc Research Fellow in IRCCyN lab, PolyTech' Nantes & Univ. Nantes, Nantes, France, University of Waterloo, Waterloo, Canada and Nanyang Technological University, Singapore. His research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, 3D image/video processing, etc. He was a secretary of HHME2013 (the 9th Joint Conference on Harmonious Human Machine Environment). He was also a special session organizer in VCIP 2013.

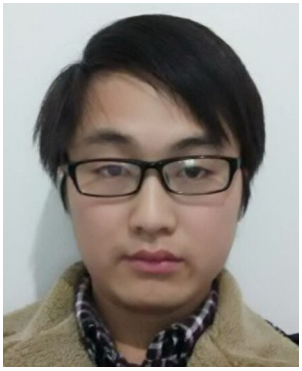
Dr Changhoon Lee received his Ph.D. degree in Graduate School of Information Management and Security (GSIMS) from Korea University, Korea. In 2008, he was a research professor at the Center for Information Security Technologies in Korea University. In 2009–2011, he was a professor in the School of Computer Engineeringin Hanshin University. He is now a professor at the Department of Computer Science and Engineering, Seoul National University of Science and Technology (SeoulTech), Korea. He has been serving not only as chairs, program committee, or organizing committee chair for many international conferences and workshops but also as a (guest) editor for international journals by some publishers. His research interests include information security, cryptography, digital forensics, smart grid security, computer theory etc. He is currently a member of the IEEE, IEEE Computer Society, IEEE Communications, IACR, KIISC, KDFS, KIPS, KITCS, KMMS, KONI, and KIIT societies.



Neal N. Xiong is current a Professor at School of Computer Science, Colorado Technical University, Colorado Spring, CO, USA. He received his both PhD degrees in Wuhan University (about software engineering), and Japan Advanced Institute of Science and Technology (about dependable networks), respectively. Before he attends Colorado Technical University, he worked in Wentworth Technology Institution, Georgia State University for many years. His research interests include Cloud Computing, Security and Dependability, Parallel and Distributed Computing, Networks, and Optimization Theory. Dr./Prof. Xiong published over 100 international journal papers and over 100 international conference papers. Some of his works were published in IEEE JSAC, IEEE or ACM transactions, ACM Sigcomm workshop, IEEE INFOCOM, ICDCS, and IPDPS. He has been a General Chair, Program Chair, Publicity Chair, PC member and OC member of over 100 international conferences, and as a reviewer of about 100 international journals, including IEEE JSAC, IEEE SMC (Park: A/B/C), IEEE Transactions on Communications, IEEE Transactions on Mobile Computing, IEEE Trans. on Parallel and Distributed Systems. He is serving as an Editor-in-Chief, Associate editor or Editor member for over 10 international journals (including Associate Editor for IEEE Tran. on Systems, Man & Cybernetics: Systems, and Editor-in-Chief for Journal of Parallel & Cloud Computing (PCC)), and a guest editor for over 10 international journals, including Sensor Journal, WINET and MONET. He has received the Best Paper Award in the 10th IEEE International Conference on High Performance Computing and Communications (HPCC-08) and the Best student Paper Award in the 28th North American Fuzzy Information Processing Society Annual Conference (NAFIPS2009). Dr./Prof. Xiong is the Chair of "Trusted Cloud Computing" Task Force, IEEE Computational Intelligence Society (CIS), http://www.cs.gsu.edu/~cscnxx/index-TF.html, and the Industry System Applications Technical Committee, http://ieee-cis.org/technical/isatc/ He is a Senior member of IEEE Computer Society



Lei Shu is a PhD candidate the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include image processing and machine vision.



Sheng Chen is a master graduate student in computer science, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include video processing and machine vision.