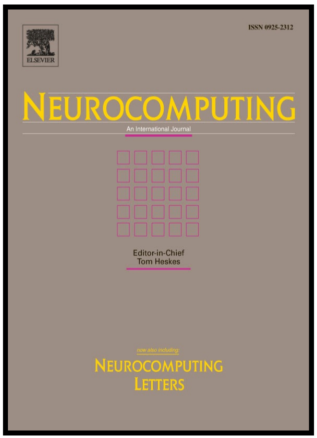
Learning Deep Event Models for Crowd Anomaly Detection

Yachuang Feng, Yuan Yuan, Xiaoqiang Lu



www.elsevier.com/locate/neucom

PII: S0925-2312(16)31098-0
DOI: http://dx.doi.org/10.1016/j.neucom.2016.09.063
Reference: NEUCOM17592

To appear in: *Neurocomputing*

Received date: 28 July 2016
Revised date: 22 September 2016
Accepted date: 26 September 2016

Learning Deep Event Models for Crowd Anomaly Detection

Yachuang Feng[a,b], Yuan Yuan[a], Xiaoqiang Lu[a,*]

[a] *Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China.*
[b] *University of Chinese Academy of Sciences, Beijing 100049, P. R. China.*

Abstract

Abnormal event detection in video surveillance is extremely important, especially for crowded scenes. In recent years, many algorithms have been proposed based on handcrafted features. However, it still remains challenging to decide which kind of feature is suitable for a specific situation. In addition, it is hard and time-consuming to design an effective descriptor. In this paper, video events are automatically represented and modeled in unsupervised fashions. Specifically, appearance and motion features are simultaneously extracted using a PCANet from 3D gradients. In order to model event patterns, a deep *Gaussian mixture model* (GMM) is constructed with observed normal events. The deep GMM is a scalable deep generative model which stacks multiple GMM-layers on top of each other. As a result, the proposed method acquires competitive performance with relatively few parameters. In the testing phase, the likelihood is calculated to judge whether a video event is abnormal or not. In this paper, the proposed method is verified on two publicly available datasets and compared with state-of-the-art algorithms. Experimental results show that the deep model is effective for abnormal event detection in video surveillance.

Keywords: Deep neural network, PCANet, deep GMM, crowded scene, abnormal event detection, video surveillance.

[*] Corresponding author
Email address: `luxq666666@gmail.com` (Xiaoqiang Lu)

1. Introduction

Abnormal event detection aims at automatically identifying abnormal events from surveillance videos. In recent years, both academia and industry [1, 2, 3, 4] have shown great interests in anomaly detection. However, it is still quite difficult to design a uniform framework for abnormal event detection. The reason is that the definition of an anomaly changes in different applications. One common solution is to learn normal event regularities from training (normal) videos, and treat abnormal events as ones which are distinct from these regularities.

Along with the development of computer vision techniques, remarkable improvements have been acquired by designing a lot of effective video event features. For example, tracking [5, 6, 7] is usually employed to analyze moving objects. Trajectory-based features are simple and have high-level semantics. However, these methods are limited by the factors of shadows and occlusions in crowded scenes. In recent years, many works describe video events at pixel-level, 2D patches or 3D blocks. For example, Reddy *et al.* [8] extract features of motion, size and texture from non-overlapping cells of video sequences. In [9], the *histogram of gradient* (HoG) and the *histogram of optical flow* (HOF) are calculated at spatio-temporal interest points. Besides, there are many other widely used video event features for anomaly detection. One commonality is that these features are hand-crafted. Generally, designing an effective descriptor is difficult and time-consuming. Meanwhile, it is hard to decide which kind of descriptor is appropriate for a specific situation.

Nowadays, deep learning [10] has been widely studied, since it learns features automatically from raw data. In many computer vision applications, deep learning has shown impressive performance, such as image segmentation [11], object detection [12], and activity recognition [13]. These works mainly focus on supervised scenarios. However, in the field of anomaly detection, labeled abnormal events are seldom available for training. Fortunately, unsupervised deep learning approaches have also been studied in recent years to address important tasks, such as image classification [14] and object tracking [15]. The reason why deep learning conducts inspiring performance is that multi-layer non-linear transformations can adaptively extract meaningful and discriminative features. Nevertheless, deep learning is seldom studied for abnormal event detection, expect for [16] and [17].

In [16], appearance, motion, and their joint representations are learned with three stacked denoising autoencoders [18]. Then anomaly scores are predicted by three one-class *support vector machines* (SVMs) on these learned features, respectively. Finally, detection results are fused with an automatically learned weight vector. In [17], Fang *et al.* represent appearance and motion features with salience maps and multi-scale HOFs [19]. After that, a deep learning framework named *PCA network* (PCANet) [14] is used to extract high-level features. As [16], a one-class SVM is adopted in [17] to detect abnormal events.

In this paper, a different deep model is designed for abnormal event detection. Firstly, 3D gradient features are computed to describe video events, due to their simplicity and effectiveness [20]. Specifically, the gradients at horizontal and vertical directions represent the appearance features. Meanwhile, motion characters are described by gradients at the temporal direction. Subsequently, a PCANet is trained to generate high-level descriptors from these 3D gradient features. In order to model normal video events, the *Gaussian mixture model* (GMM) is generally used to learn normal event patterns [21, 22, 23]. Since video events are quite complicated, it always requires massive Gaussian components to model these event patterns. Unfortunately, this brings about plenty of parameters and increases the complexity. In order to deal with this problem, this paper develops a deep GMM [24] method to explore video event patterns. The deep GMM method is provided with high representation power while having relatively few parameters.

Compared with existing algorithms, contributions of this paper are listed as follows:

- A deep model is proposed for abnormal event detection, with regard to both feature learning and model building.

- In term of feature learning, event representations are automatically extracted with a PCANet, which is a simple and effective unsupervised deep learning framework.

- With respect to model building, we develop a deep GMM method, which explores video event patterns with limited number of parameters. In this case, the proposed method possesses high representation power with relatively few parameters.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 is the detailed presentation of the proposed algorithm. Experimental results are given in Section 4. In the end, Section 5 concludes this paper.

2. Related Works

Generally, labeled abnormal events are unavailable for training, therefore existing algorithms [25, 26, 27, 28, 29] try to learn an event model from normal videos. Subsequently, abnormal events are detected as samples which disagree with this normal event model. According to the clue types, existing algorithms for abnormal event detection can be divided into two categories: 1) Trajectory based techniques. Abnormal trajectories are the ones occurring much more rarely compared with normal ones. 2) Local block based algorithms. Blocks which contain dramatic event patterns are treated as anomalies.

For the first category, trajectories are extracted in advance for moving objects. By exploring potential rules among normal trajectories, abnormal events are identified as ones which disobey these rules. For example, Anjum and Cavallaro [30] extract multiple features based on trajectories, such as the trajectory mean, speed and acceleration. In their method, each feature is applied with a clustering algorithm, and the final clustering result is obtained by taking clusters from all features into account. The clusters with few members and samples far away from these cluster centers are treated as anomalies. In order to deal with the occlusion and segmentation problems, Cheng and Hwang [31] make use of the adaptive particle sampling and Kalman filtering to obtain reliable trajectories. Besides, tracking at particle and feature point level have also been taken into consideration. For example, Wu *et al.* [25] propose a Lagrangian particle dynamics approach, and extract chaotic invariant features from representative trajectories. Cui *et al.* [27] track interest points and represent the crowd dynamic by calculating interaction energy potentials.

As for the second category, event features are extracted from 2D image patches or 3D video blocks, such as spatio-temporal gradient and HOF. Cong *et al.* [19, 32, 26] represent the motion patterns with multi-scale HOFs. By calculating sparse representation coefficients with the trained dictionary, abnormal events are detected as samples owning large *sparse reconstruction costs* (SRCs). Adam *et al.* [33] adopt histograms to model the probability of optical flow at a group of spatial locations. Kim and Grauman [34] use a *mixture of probability principle component analyzers* (MPPCA) to model local optical flow, and enforce the consistency by a spatio-temporal Markov random field. Mehran *et al.* [28] analyze crowd behaviors based on a *social force* (SF) model, where interaction forces are calculated with optical flow. Kratz and Nishino [35] fit spatio-temporal gradient features with a Gaussian model, and detect abnormal events with a hidden Markov model. Thida *et al.* [36] learn variations of motion in an embedded space, and describe crowd activities with a spatio-temporal Laplacian eigenmap. Mahadevan *et al.* [29] jointly model dynamics and appearances of crowded scenes using a *mixture of dynamic textures* (MDT). In [37], Li *et al.* develop the MDT model to a *hierarchical MDT* (H-MDT) model, which adopts a background subtraction technique for detecting temporal anomalies and utilizes a discriminant saliency method for identifying spatial anomalies.

3. Deep Model for Abnormal Event Detection

In this section, the proposed method is described in detail. Firstly, 3D gradients are calculated for each video frame. Secondly, high-level features for video events are automatically extracted with the PCANet. Finally, normal event patters are modeled by the deep GMM.

3.1. Feature Learning based on the PCANet

For most existing methods, spatial and temporal features are manually selected, such as intensity, color, gradient, and optical flow. In this paper, 3D gradient features are computed to represent video events. In [20], experiments validate the effectiveness and efficiency of 3D gradient features for abnormal event detection. Meanwhile, the

3D gradient models both appearance and motion clues. Based on these 3D gradients, this paper utilizes a deep neural network to abstract high-level features. In recent years, deep learning has achieved impressive performance in many computer vision applications [11, 12, 13, 14, 38]. This is benefited from the multi-layer non-linear transformations which can adaptively extract meaningful and discriminative features. In the field of anomaly detection, there is no labeled abnormal events available for training. Actually, only normal videos are provided in the training dataset. For these reasons, this paper learns video event features by the PCANet [14], which is a simple and effective unsupervised deep learning approach. In what follows, details about feature learning are discussed.

Given a video sequence, 3D gradients are computed from grayscale images, obtaining a set of features $\{\cdots, \mathcal{I}_i, \cdots\}$, where $\mathcal{I}_i \in \mathbb{R}^{r_1 \times r_2 \times 3}$ is the 3D gradients for the ith frame with spatial resolution of $r_1 \times r_2$. The 3 channels in $\mathcal{I}_i$ are gradients at horizontal, vertical and temporal directions. Through a frame difference method, moving pixels are detected. Based on this, patches ($k_1 \times k_2 \times 3$) containing moving pixels are extracted and vectorized. By subtracting mean values from each sample, we can get a set of samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$, where $\mathbf{x} \in \mathbb{R}^{3k_1k_2}$ and N is the total number of samples. The aim of PCA is to minimize the reconstruction error using a set of orthogonal vectors $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{L_1}] \in \mathbb{R}^{3k_1k_2 \times L_1}$, *i.e.*,

$$\begin{aligned} &\min_{\mathbf{V}} \frac{1}{2}\|\mathbf{X} - \mathbf{V}\mathbf{V}^T\mathbf{X}\|_F^2, \\ &s.t. \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_{L_1}, \end{aligned} \tag{1}$$

where $\mathbf{I}_{L_1}$ is an identity matrix with size $L_1 \times L_1$, and L_1 is the number of filters in the first layer. The solution for Eq. (1) is the the first L_1 eigenvectors of $\mathbf{X}\mathbf{X}^T$, and the PCA filters are represented as

$$\mathbf{W}_l^1 = mat_{k_1,k_2,3}(\mathbf{v}_l) \in \mathbb{R}^{k_1 \times k_2 \times 3}, l = 1, 2, \cdots, L_1, \tag{2}$$

where mat is a function that maps a vector to a matrix, and $\mathbf{v}_l$ is the lth principal eigenvector.

With respect to the second layer, each input is a filter output of the first layer. The lth output of the first layer is

$$\mathcal{I}_i^l = \mathcal{I}_i * \mathbf{W}_l^1, \tag{3}$$

where $*$ indicates the convolution operator. Subsequently, patches containing moving pixels are extracted and vectorized from $\mathcal{I}_i^l$. As in the first layer, mean values are subtracted from each sample and PCA filters in the second layer are computed similarly with Eq. (2). By repeating the above process, one can simply build more PCA layers.

Suppose there are two layers in the deep model, one input will produce $L_1 \times L_2$ outputs, where L_2 is the number of filters in the second layer. As in [14], the L_2 outputs for each input in the second layer are binarized and each pixel is viewed as a decimal number by

$$\mathcal{T}_i^l = \sum_{\ell=1}^{L_2} 2^{\ell-1} H(\mathcal{I}_i^l * \mathbf{W}_\ell^2), \tag{4}$$

where $\mathbf{W}_\ell^2$ is the ℓth filter of the second layer. $H(z)$ is a binarization function

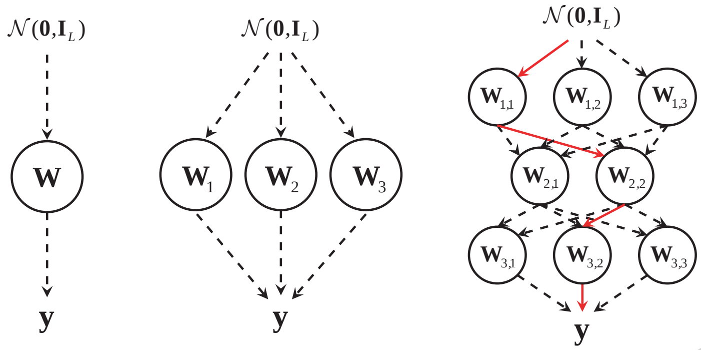
$$H(z) = \begin{cases} 1, & z > 0 \\ 0, & otherwise, \end{cases} \tag{5}$$

which is beneficial to decide whether a video sample contains a particular principle component. As a result, the outputs of the deep model are L_1 images $\mathcal{T}_i^l, l = 1, 2, \cdots, L_1$.

In order to detect local abnormal events, each image $\mathcal{T}_i$ is split into patches. The histograms (with 2^{L_2} bins) are computed for patches containing moving pixels. By concatenating all L_1 histograms, we can obtain a deep representation for each sample.

3.2. Representation of Normal Event Patterns with the Deep GMM

In many works [21, 22, 23], GMMs are used to learn normal event patterns. However, massive Gaussian components are required to model the complex video events. As a result, complexities of these methods increase drastically. In this paper, a deep GMM [24] is developed to model normal video events. The motivation of the deep GMM is shown in Fig. 1.



(a) Single Gaussian (b) GMM (c) Deep GMM

Figure 1: Visualizations of single Gaussian, GMM and deep GMM distribution.

For a variable $\mathbf{y} \in \mathbb{R}^L$ which follows the distribution $p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T\right)$, it can be represented as a linear transformation $\mathbf{y} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu}$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L)$. Similarly, a GMM can be represented by a set of transformations $(\mathbf{W}_k, \boldsymbol{\mu}_k)$, $k = 1, 2, \cdots, K$, with probabilities $\pi_k, k = 1, 2, \cdots, K$. The resulting distribution is

$$p(\mathbf{y}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu}_k, \mathbf{W}_k\mathbf{W}_k^T\right). \tag{6}$$

In this way, it is easy to generalize GMMs in a multi-layered fashion, as is shown in Fig. 1(c). Each Gaussian component is a path of transformations in a network. Suppose there are d layers and each layer has $N_j, j = 1, 2, \cdots, d$ units, the network contains $K = \prod_j N_j$ Gaussian components. Let Φ be the set of all possible paths in the network, the prior probability of each path $p = (p_1, p_2, \cdots, p_d) \in \Phi$ is π_p, satisfying

$$\sum_{p \in \Phi} \pi_p = \sum_{p_1, p_2, \cdots, p_d} \pi_{(p_1, p_2, \cdots, p_d)} = 1, \tag{7}$$

where p_j is the node of path p in the jth layer. As a result, the density function of $\mathbf{y}$ can be represented as

$$p(\mathbf{y}) = \sum_{p \in \Phi} \pi_p \mathcal{N}\left(\mathbf{y}|\boldsymbol{\beta}_p, \boldsymbol{\Lambda}_p\boldsymbol{\Lambda}_p^T\right), \tag{8}$$

where

$$
\begin{aligned}
\boldsymbol{\beta}_p &= \boldsymbol{\mu}_{d,p_d} + \mathbf{W}_{d,p_d}(\cdots, (\boldsymbol{\mu}_{2,p_2} + \mathbf{W}_{2,p_2}\boldsymbol{\mu}_{1,p_1})) \\
&= \boldsymbol{\mu}_{d,p_d} + \sum_{j=d}^{2}\prod_{l=d}^{j}\mathbf{W}_{l,p_l}\boldsymbol{\mu}_{j-1,p_{j-1}}.
\end{aligned} \tag{9}
$$

and

$$
\boldsymbol{\Lambda}_p = \prod_{j=d}^{1}\mathbf{W}_{j,p_j}. \tag{10}
$$

$\mathbf{W}_{s,t}$ and $\boldsymbol{\mu}_{s,t}$ are the tth transformation matrix and bias of the sth layer.

The optimization of deep GMMs is similar to that of GMMs, which is based on the *expectation maximization* (EM) algorithm. In the **E step**, the posterior probability γ_{np} that a sample $\mathbf{y}_n$ is generated by path p is calculated by

$$
\gamma_{np} = \frac{\pi_p \mathcal{N}\left(\mathbf{y}_n | \boldsymbol{\beta}_p, \boldsymbol{\Lambda}_p\boldsymbol{\Lambda}_p^T\right)}{\sum_{q\in\Phi}\pi_p \mathcal{N}\left(\mathbf{y}_n | \boldsymbol{\beta}_q, \boldsymbol{\Lambda}_q\boldsymbol{\Lambda}_q^T\right)}. \tag{11}
$$

The lower bound of the log-likelihood function in Eq. (8) is

$$
\begin{aligned}
\ell = \sum_n \sum_{p\in\Phi}\gamma_{np}\big[\ln \pi_p - \frac{L}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Lambda}_p\boldsymbol{\Lambda}_p^T| \\
- \frac{1}{2}(\mathbf{y}_n - \boldsymbol{\beta}_p)^T(\boldsymbol{\Lambda}_p\boldsymbol{\Lambda}_p^T)^{-1}(\mathbf{y}_n - \boldsymbol{\beta}_p) - \ln\gamma_{np}\big].
\end{aligned} \tag{12}
$$

In the **M step**, parameters are optimized by maximizing Eq. (12). By making the derivative equal to zero, the closed-form updates for π_p is

$$
\pi_p = \frac{\sum_n \gamma_{np}}{N}. \tag{13}
$$

With respect to $\boldsymbol{\mu}$ and $\mathbf{W}$, the deep GMM is optimized layer by layer.

Update $\boldsymbol{\mu}_{i,p_i}$: Denote $\mathbf{D}_{1,p} = \prod_{j=d}^{i+1}\mathbf{W}_{j,p_j}$ and $\mathbf{D}_{2,p} = \prod_{j=i-1}^{1}\mathbf{W}_{j,p_j}$, we can get $\boldsymbol{\Lambda}_p = \mathbf{D}_{1,p}\mathbf{W}_{i,p_i}\mathbf{D}_{2,p}$, and $\boldsymbol{\beta}_p$ can be rewritten as

$$
\begin{aligned}
\boldsymbol{\beta}_p &= \mathbf{c}_1 + \prod_{l=d}^{i+1}\mathbf{W}_{l,p_l}\boldsymbol{\mu}_{i,p_i} \\
&= \mathbf{c}_1 + \mathbf{D}_{1,p}\boldsymbol{\mu}_{i,p_i},
\end{aligned} \tag{14}
$$

where $\mathbf{c}_1 = \boldsymbol{\mu}_{d,p_d} + \sum_{j \neq i+1} \prod_{l=d}^{j} \mathbf{W}_{l,p_l} \boldsymbol{\mu}_{j-1,p_{j-1}}$ is a constant matrix. According to Eq. (12), the objective function related to $\boldsymbol{\mu}_{i,p_i}$ is

$$\begin{aligned} &\arg\max_{\boldsymbol{\mu}_{i,p_i}} \sum_n \sum_{p \in \Phi} \gamma_{np} \Big[-\frac{1}{2} (\mathbf{y}_n - \boldsymbol{\beta}_p)^T (\boldsymbol{\Lambda}_p \boldsymbol{\Lambda}_p^T)^{-1} (\mathbf{y}_n - \boldsymbol{\beta}_p) \Big] \\ =& \arg\max_{\boldsymbol{\mu}_{i,p_i}} \sum_n \sum_{p \in \Phi} \gamma_{np} \Big[-\frac{1}{2} (\mathbf{y}_n - \mathbf{c}_1 - \mathbf{D}_{1,p} \boldsymbol{\mu}_{i,p_i})^T (\boldsymbol{\Lambda}_p \boldsymbol{\Lambda}_p^T)^{-1} (\mathbf{y}_n - \mathbf{c}_1 - \mathbf{D}_{1,p} \boldsymbol{\mu}_{i,p_i}) \Big]. \end{aligned} \tag{15}$$

$\sum_{p \in \Phi}$ means summing over all possible paths containing the current node. The derivative for the above function is

$$\sum_n \sum_{p \in \Phi} \gamma_{np} \mathbf{D}_{1,p}^T (\boldsymbol{\Lambda}_p \boldsymbol{\Lambda}_p^T)^{-1} (\mathbf{y}_n - \mathbf{c}_1 - \mathbf{D}_{1,p} \boldsymbol{\mu}_{i,p_i}). \tag{16}$$

Taking the derivative equal to zero, $\boldsymbol{\mu}_{i,p_i}$ is updated as

$$\boldsymbol{\mu}_{i,p_i} = \big(\sum_n \sum_{p \in \Phi} \gamma_{np} \mathbf{D}_{1,p}^T (\boldsymbol{\Lambda}_p \boldsymbol{\Lambda}_p^T)^{-1} \mathbf{D}_{1,p} \big) \setminus \sum_n \sum_{p \in \Phi} \gamma_{np} \mathbf{D}_{1,p}^T (\boldsymbol{\Lambda}_p \boldsymbol{\Lambda}_p^T)^{-1} (\mathbf{y}_n - \mathbf{c}_1). \tag{17}$$

Update $\mathbf{W}_{i,p_i}$: Rewrite $\boldsymbol{\beta}_p$ as

$$\begin{aligned} \boldsymbol{\beta}_p &= \mathbf{c}_2 + \sum_{j=i}^{2} \prod_{l=d}^{j} \mathbf{W}_{l,p_l} \boldsymbol{\mu}_{j-1,p_{j-1}} \\ &= \mathbf{c}_2 + \prod_{j=d}^{i+1} \mathbf{W}_{j,p_j} \mathbf{W}_{i,p_i} (\boldsymbol{\mu}_{i-1,p_{i-1}} + \sum_{j=i-1}^{2} \prod_{l=i-1}^{j} \mathbf{W}_{l,p_l} \boldsymbol{\mu}_{j-1,p_{j-1}}) \\ &= \mathbf{c}_2 + \mathbf{D}_{1,p} \mathbf{W}_{i,p_i} \mathbf{c}_3, \end{aligned} \tag{18}$$

where $\mathbf{c}_2 = \boldsymbol{\mu}_{d,p_d} + \sum_{j=d}^{i+1} \prod_{l=d}^{j} \mathbf{W}_{l,p_l} \boldsymbol{\mu}_{j-1,p_{j-1}}$ and $\mathbf{c}_3 = \boldsymbol{\mu}_{i-1,p_{i-1}} + \sum_{j=i-1}^{2} \prod_{l=i-1}^{j} \mathbf{W}_{l,p_l} \boldsymbol{\mu}_{j-1,p_{j-1}}$ are constant vectors. According to Eq. (12), the objective function related to $\mathbf{W}_{i,p_i}$ is

$$\begin{aligned}
&\arg\max_{\mathbf{W}_{i,p_i}} \sum_n \sum_{p\in\Phi} \gamma_{np}\Big[-\frac{1}{2}\ln|\boldsymbol{\Lambda}_p\boldsymbol{\Lambda}_p^T| - \frac{1}{2}(\mathbf{y}_n-\boldsymbol{\beta}_p)^T(\boldsymbol{\Lambda}_p\boldsymbol{\Lambda}_p^T)^{-1}(\mathbf{y}_n-\boldsymbol{\beta}_p)\Big]\\
=&\arg\max_{\mathbf{W}_{i,p_i}} \sum_n \sum_{p\in\Phi} \gamma_{np}\Big[-\frac{1}{2}\ln|\mathbf{D}_{1,p}\mathbf{W}_{i,p_i}\mathbf{D}_{2,p}\mathbf{D}_{2,p}^T\mathbf{W}_{i,p_i}^T\mathbf{D}_{1,p}^T|\\
&-\frac{1}{2}(\mathbf{y}_n-\mathbf{c}_2-\mathbf{D}_{1,p}\mathbf{W}_{i,p_i}\mathbf{c}_3)^T(\mathbf{D}_{1,p}\mathbf{W}_{i,p_i}\mathbf{D}_{2,p}\mathbf{D}_{2,p}^T\mathbf{W}_{i,p_i}^T\mathbf{D}_{1,p}^T)^{-1}\\
&(\mathbf{y}_n-\mathbf{c}_2-\mathbf{D}_{1,p}\mathbf{W}_{i,p_i}\mathbf{c}_3)\Big]\\
=&\arg\max_{\mathbf{W}_{i,p_i}} \sum_n \sum_{p\in\Phi} \gamma_{np}\Big[\ln|\mathbf{W}_{i,p_i}^{-1}| - \frac{1}{2}\|\mathbf{D}_{2,p}^{-1}\mathbf{W}_{i,p_i}^{-1}\mathbf{D}_{1,p}^{-1}(\mathbf{y}_n-\mathbf{c}_2)-\mathbf{D}_{2,p}^{-1}\mathbf{c}_3\|_2^2\Big].
\end{aligned} \tag{19}$$

The gradient for the above function with respect to $\mathbf{W}_{i,p_i}^{-1}$ is

$$\sum_n \sum_{p\in\Phi} \gamma_{np}\Big\{\mathbf{W}_{i,p_i}^T - \mathbf{D}_{2,p}^{-T}\left[\mathbf{D}_{2,p}^{-1}\mathbf{W}_{i,p_i}^{-1}\mathbf{D}_{1,p}^{-1}(\mathbf{y}_n-\mathbf{c}_2)-\mathbf{D}_{2,p}^{-1}\mathbf{c}_3\right](\mathbf{y}_n-\mathbf{c}_2)^T\mathbf{D}_{1,p}^{-T}\Big\}. \tag{20}$$

Based on Eq. (20), $\mathbf{W}_{i,p_i}^{-1}$ is updated using the *stochastic gradient descent* (SGD) method. Aside from the SGD, other optimization methods are also recommended, such as the AdaDelta method [39] and the RMSProp method [40].

3.3. Abnormal Event Detection

In the testing phase, patches containing moving pixels are extracted, and features are computed using the trained PCANet from 3D gradients. For each testing sample $\mathbf{y}$, the probability $p(\mathbf{y})$ is calculated based on the deep GMM using Eq. (8). If $p(\mathbf{y})$ is smaller than a predefined threshold θ, *i.e.*, $p(\mathbf{y}) < \theta$, it is classified as an anomaly.

4. Experiments

In this section, two public datasets are utilized to evaluate the proposed algorithm. Based on them, qualitative and quantitative comparisons are provided with state-of-the-art algorithms.

4.1. Evaluation Methodology

In order to evaluate the performance, three commonly used measurements are adopted: frame-level, pixel-level, and object-level. All these measurements compute the matching rate between detection results and the ground-truth. Definitions of the ground-truth are: “positive” – the existence of an abnormal event; “negative” – the absence of abnormal events.

- **Frame-level.** One video frame is treated as abnormal as long as one abnormal pixel is detected. If the corresponding ground-truth is abnormal, it is a true positive. Otherwise, it is a false positive. The meaning of frame-level measurement is simple and clear, but it cannot ensure the correct detection of abnormal pixels. This is because some true positives may be caused by co-occurrences of false detections and true abnormalities.

- **Pixel-level.** In this measurement, a detection is a true positive only when more than 40% truly abnormal pixels are detected. A normal frame will be a false positive as long as one normal pixel is identified as abnormal. Compared with the frame-level measurement, the pixel-level measurement pays more attention to the correct detection of abnormal events.

- **Object-level.** A high pixel-level true positive rate may lead to plenty of false positive pixels. The reason is that when all pixels are detected as anomalies, there must be more than 40% abnormal pixels being detected. The object-level measurement defines true positives as video frames, in which

$$\frac{\text{Detected anomaly} \bigcap \text{True anomaly}}{\text{Detected anomaly} \bigcup \text{True anomaly}} \geq \vartheta, \tag{21}$$

  where ϑ is a predefined threshold. $\bigcap$ and $\bigcup$ represent the intersection and union operators, respectively. The object-level measurement concerns more about the accurate detection of abnormal objects.

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

$$\begin{aligned}\mathrm{TPR} =& \frac{\mathrm{TruePositive}}{\mathrm{TruePositive} + \mathrm{FalseNegative}},\\ \mathrm{FPR} =& \frac{\mathrm{FalsePositive}}{\mathrm{TrueNegative} + \mathrm{FalsePositive}}.\end{aligned}$$

Based on ROC curves, there are three evaluation criteria:

- *Area under curve* (AUC). The area under the ROC curve.
- *Equal error rate* (EER). The ratio of misclassified frames when the false positive rate equals to the miss rate, *i.e.*, the FPR at which $\mathrm{FPR} = 1 - \mathrm{TPR}$.
- *Equal detected rate* (EDR). The detection rate at the point of EER, *i.e.*, $\mathrm{EDR} = 1 - \mathrm{EER}$.

4.2. UCSD Ped1 Dataset

The UCSD[1] dataset is collected on the UCSD campus. In this dataset, the crowd density changes greatly from sparse to extremely crowded. The training dataset contains only normal video events which are composed of pedestrians. Abnormal events are characterized by non-pedestrians or abnormal motion patterns. These abnormal events are not staged or synthesized, but occur naturally. Consequently, this dataset is very challenging.

In this paper, the UCSD Ped1 dataset is utilized for comparison. There are 34 and 36 video clips in the training and testing datasets, respectively. Each short video clip consists of 200 frames, whose spatial resolution is 158×238. In this experiment, the filter size of the PCANet is 5×5, and $L_1 = 4$, $L_2 = 3$. Outputs of the PCANet are divided into 11×11 patches. For each patch, a 32-dimension ($L_1 \cdot 2^{L_2}$) feature is obtained. With regard to the deep GMM, there are three layers, with $N_1 = 4$, $N_2 = 3$, and $N_3 = 3$.

The competitors include MDT [29], MPPCA [34], SF [28], Adam's algorithm [33], SRC [19], and Lu's algorithm [20]. Some visualized comparisons are shown in Fig. 2,

[1] `http://www.svcl.ucsd.edu/projects/anomaly/dataset.html`.

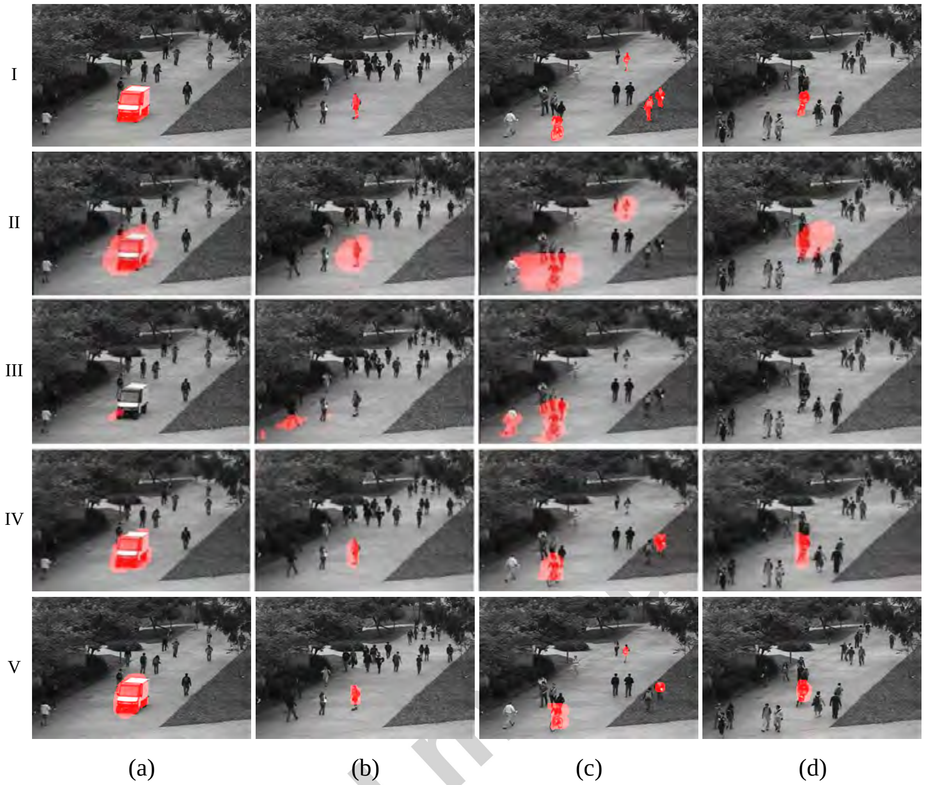
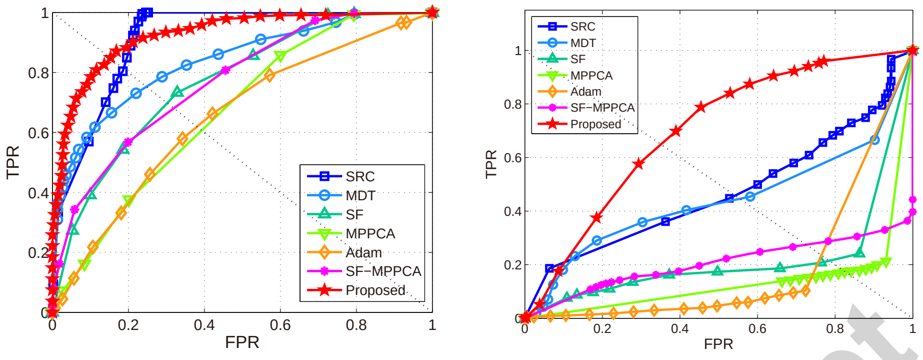


Figure 2: Examples of abnormal event detections on UCSD Ped1 dataset. (I) the ground-truth; (II) the MDT algorithm [29]; (III) the SF-MPPCA algorithm [29]; (IV) the SRC algorithm [19]; and (V) the proposed algorithm.

in which red masks indicate the locations of abnormal events. The anomalies include a car, a skater, two bikers, a runner, and two people walking through the grass. In Fig. 2, the top row is the ground-truth, the second row is given by the MDT algorithm [29], the third and fourth rows are generated by the SF-MPPCA algorithm [29] and the SRC algorithm [19], respectively. The last row presents the results of the proposed algorithm. For the MDT algorithm, it misses the people walking through the grass (column c). As for the SF-MPPCA algorithm, it completely loses the skater (column b), the runner and both people walking through the grass (column c), as well as a biker (column d). With respect to the SRC [19], it misses the runner (column c) and a person on the grass (column c). Although the proposed algorithm loses one person walking on the grass, the overall performance is better than the others.



(a)

(b)

Figure 3: ROC curves for UCSD Ped1 dataset. (a) Frame-level; (b) Pixel-level.

Table 1: Comparison of frame-level results on the UCSD Ped1 dataset.

	EER	AUC
MDT [29]	25%	81.8%
MPPCA [34]	40%	67%
SF-MPPCA [29]	32%	76.9%
SF [28]	31%	76.8%
Adam [33]	38%	64.9%
SRC [19]	19%	86%
Lu [20]	15%	91.8%
AMDN [16]	16%	92.1%
Proposed	15.1%	92.5%

Table 2: Comparison of pixel-level results on the UCSD Ped1 dataset.

	EDR	AUC
MDT [29]	45%	44.1%
MPPCA [34]	18%	13.3%
SF-MPPCA [29]	28%	20.5%
SF [28]	21%	21.3%
Adam [33]	24%	19.7%
SRC [19]	46%	46.1%
Lu [20]	59.1%	63.8%
AMDN [16]	59.9%	67.2%
Proposed	64.9%	69.9%

With regard to the frame-level and pixel-level measurements, Fig. 3 shows ROC curves of the competitors. Fig. 3(a) is the frame-level performance, and Fig. 3(b) is the pixel-level performance. It can be found that for the frame-level ROC, the proposed

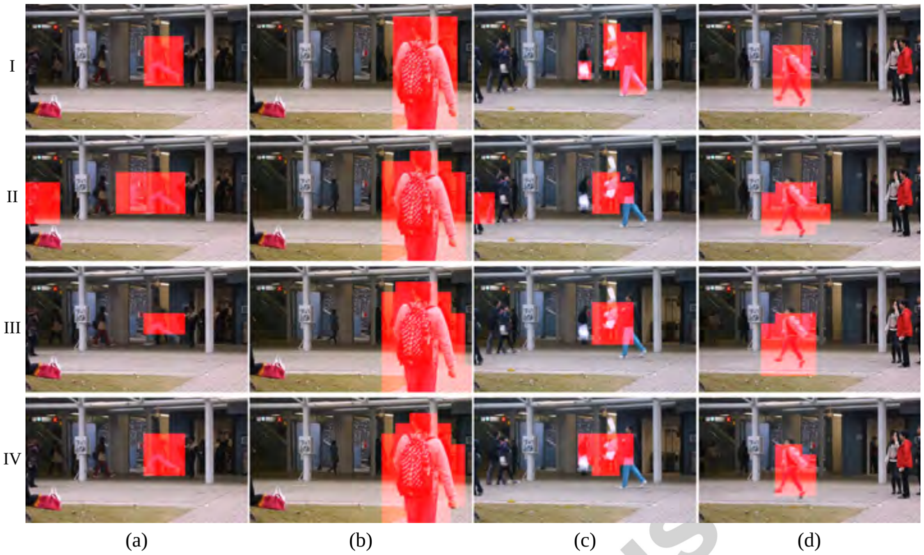


Figure 4: Examples of abnormal event detections on the Avenue dataset. (I) the ground-truth; (II) Lu's algorithm [20]; (III) 3D Gradient; and (IV) the proposed algorithm.

algorithm performs better than others except for the SRC algorithm when the false positive rate is high. For the pixel-level ROC, the proposed algorithm outperforms all these competitors. Based on the ROC curves, EER, EDR, and AUC values are listed in Table 1 and Table 2. From these numbers, it can be easily found that our method achieves high detection accuracies for both frame-level and pixel-level measurements.

4.3. Avenue Dataset

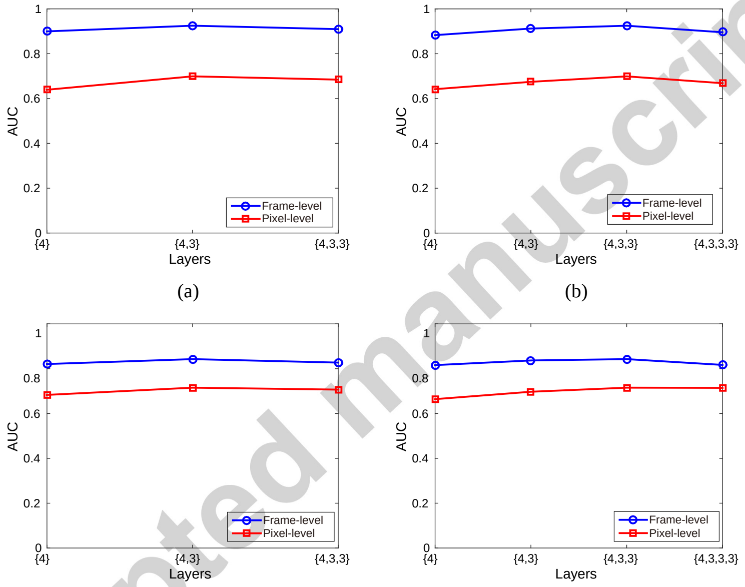
This dataset is provided by Lu *et al.* [20], and contains 16 and 21 video clips for training and testing, respectively. In this dataset, abnormal events include loitering, running, throwing objects, and so on. As claimed by the authors[2], main challenges are caused by slight camera shakes in the testing videos, a few outliers in the training videos, and the short of some normal patterns in the training videos.

Some visualized results are shown in Fig. 4. "3D Gradient" is the result of the proposed algorithm using 3D gradient features, which are used in [20]. In Fig. 4(a) and Fig. 4(c), some false alarms are detected by Lu's algorithm. In the last two rows,

[2] `http://appsrv.cse.cuhk.edu.hk/~cwlu/Anormality_1000_FPS/dataset.html`.

Table 3: Comparisons of detection accuracy on the the Avenue dataset.

ϑ	Lu's algorithm [20]	3D Gradient	Proposed
0.2	70.0%	72.3%	75.4%
0.3	67.3%	71.1%	73.4%
0.4	63.3%	68.8%	70.6%
0.5	59.3%	65.8%	67.5%
0.6	57.5%	64.1%	64.6%
0.7	55.7%	62.9%	63.5%
0.8	54.4%	61.9%	63.0%



(a)

(b)

(c)

(d)

Figure 5: Performance with different numbers of layers. (a) AUC values on the UCSD Ped1 dataset with varying numbers of layers for the PCANet; (b) AUC values on the UCSD Ped1 dataset with varying numbers of layers for the deep GMM; (a) AUC values on the Avenue dataset with varying numbers of layers for the PCANet; (b) AUC values on the Avenue dataset with varying numbers of layers for the deep GMM.

both "3D Gradient" and the proposed algorithm exclude these false detections. Table 3 compares object-level accuracies under various overlapping thresholds, *i.e.*, ϑ in Eq. (21). It can be easily found that the proposed algorithm achieves competitive performance. Compared with Lu's method [20], our method improves the average detection accuracy by 7.21 percent, which verifies the effectiveness. Meanwhile, compared with "3D Gradient", the average detection accuracy is improved by 1.59 percent, which proves that the deep representation outperforms hand-crafted features (3D gradients) for abnormal event detection.

4.4. Discussion

In this work, both the PCANet and the deep GMM are deep architectures, which can possess various numbers of layers. In order to analyze the impact of the number of layers, several experiments are conducted on the UCSD Ped1 and the Avenue datasets. For the PCANet, the number of layers varies from 1 to 3, and the numbers of filters are $\{4\}$, $\{4, 3\}$, and $\{4, 3, 3\}$, respectively. With respect to the deep GMM, the number of layers changes from 1 to 4, and the numbers of nodes are $\{4\}$, $\{4, 3\}$, $\{4, 3, 3\}$, and $\{4, 3, 3, 3\}$, respectively. The experiments are conducted by changing one aspect while fixing the other. For the UCSD Ped1 dataset, frame-level and pixel-level AUC values are shown in Fig. 5(a) and Fig. 5(b). Meanwhile, AUC values for the Avenue dataset are shown in Fig. 5(c) and Fig. 5(d). As can be seen from Fig. 5(a) and Fig. 5(c), PCANets with one layer are insufficient to describe video events. PCANets with three layers have similar performance to that with two layers, but result in high computation complexities. From Fig. 5(b) and Fig. 5(d), we can find that deep GMMs with three layers achieve the best performance. This is because deep GMMs with few layers have low representation ability, and it tends to overfitting when the number of Gaussian components increases.

5. Conclusion

This paper presents a novel abnormal event detection method based on unsupervised deep neural networks. Specifically, effective video event features are automatically extracted from 3D gradients to represent both appearance and motion clues. In order to lean normal event patterns, this paper utilizes a deep Gaussian mixture model, which conducts competitive performance using relatively few parameters. Experiments on two public datasets show distinct improvements when compared with state-of-the-art algorithms. The comparison results not only valid the effectiveness of the proposed method, but also prove the advantages of deep representations compared with hand-crafted features. In future work, more efforts will be made to build deep models on appearance, short-term, and long-term temporal motion clues.

Acknowledgments

This work is supported by the National Basic Research Program of China (Youth 973 Program) (Grant No. 2013CB336500), the State Key Program of National Natural Science of China (Grant No. 61232010), the National Natural Science Foundation of China (Grant No. 61472413), the Key Research Program of the Chinese Academy of Sciences (Grant No. KGZD-EW-T03), the Open Research Fund of the Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences (Grant No. LSIT201408), and the Young Top-notch Talent Program of Chinese Academy of Sciences (Grant No. QYZDB-SSW-JSC015).

References

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

[2] H. Guo, X. Wu, S. Cai, N. Li, J. Cheng, Y.-L. Chen, Quaternion discrete cosine transformation signature analysis in crowd scenes for abnormal event detection, Neurocomputing 204 (2016) 106–115.

[3] N. Li, X. Wu, D. Xu, H. Guo, W. Feng, Spatio-temporal context analysis within video volumes for anomalous-event detection and localization, Neurocomputing 155 (2015) 309–319.

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

[5] C. Li, Z. Han, Q. Ye, J. Jiao, Visual abnormal behavior detection based on trajectory sparse reconstruction analysis, Neurocomputing 119 (2013) 94–100.

[6] J. Fang, Q. Wang, Y. Yuan, Part-based online tracking with geometry constraint and attention selection, IEEE Transactions on Circuits and Systems for Video Technology 24 (5) (2014) 854–864.

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

[8] V. Reddy, C. Sanderson, B. C. Lovell, Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 55–61.

[9] B. Zhao, L. Fei-Fei, E. P. Xing, Online detection of unusual events in videos via dynamic sparse coding, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3313–3320.

[10] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M. S. Lew, Deep learning for visual understanding: A review, Neurocomputing 187 (2016) 27–48.

[11] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1915–1929.

[12] Y. Zhang, X. Li, Z. Zhang, F. Wu, L. Zhao, Deep learning driven blockwise moving object detection with binary scene modeling, Neurocomputing 168 (2015) 454–463.

[13] M. Liu, H. Liu, Depth context: A new descriptor for human activity recognition by using sole depth sequences, Neurocomputing 175 (2016) 747–758.

[14] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, PCANet: A simple deep learning baseline for image classification?, IEEE Transactions on Image Processing 24 (12) (2015) 5017–5032.

[15] N. Wang, D.-Y. Yeung, Learning a deep compact image representation for visual tracking, in: Proc. Advances in Neural Information Processing Systems, 2013, pp. 809–817.

[16] D. Xu, E. Ricci, Y. Yan, J. Song, N. Sebe, Learning deep representations of appearance and motion for anomalous event detection, in: Proc. British Machine Vision Conference, 2015, pp. 1–12.

[17] Z. Fang, F. Fei, Y. Fang, C. Lee, N. Xiong, L. Shu, S. Chen, Abnormal event detection in crowded scenes based on deep learning, Multimedia Tools and Applications (2015) 1–23.

[18] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, The Journal of Machine Learning Research 11 (2010) 3371–3408.

[19] Y. Cong, J. Yuan, J. Liu, Sparse reconstruction cost for abnormal event detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3449–3456.

[20] C. Lu, J. Shi, J. Jia, Abnormal event detection at 150 FPS in MATLAB, in: Proc. IEEE International Conference on Computer Vision, 2013, pp. 2720–2727.

[21] T. Lu, L. Wu, X. Ma, P. Shivakumara, C. L. Tan, Anomaly detection through spatio-temporal context modeling in crowded scenes, in: Proc. International Conference on Pattern Recognition, 2014, pp. 2203–2208.

[22] D. Du, H. Qi, Q. Huang, W. Zeng, C. Zhang, Abnormal event detection in crowded scenes based on structural multi-scale motion interrelated patterns, in: Proc. IEEE International Conference on Multimedia and Expo, 2013, pp. 1–6.

[23] A. Basharat, A. Gritai, M. Shah, Learning object motion patterns for anomaly detection and improved object detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[24] A. van den Oord, B. Schrauwen, Factoring variations in natural images with deep gaussian mixture models, in: Proc. Advances in Neural Information Processing Systems, 2014, pp. 3518–3526.

[25] S. Wu, B. E. Moore, M. Shah, Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2054–2060.

[26] D. Xu, R. Song, X. Wu, N. Li, W. Feng, H. Qian, Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts, Neurocomputing 143 (2014) 144–152.

[27] X. Cui, Q. Liu, M. Gao, D. N. Metaxas, Abnormal detection using interaction energy potentials, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3161–3167.

[28] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 935–942.

[29] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1975–1981.

[30] N. Anjum, A. Cavallaro, Multifeature object trajectory clustering for video analysis, IEEE Transactions on Circuits and Systems for Video Technology 18 (11) (2008) 1555–1564.

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

[32] Y. Cong, J. Yuan, J. Liu, Abnormal event detection in crowded scenes using sparse representation, Pattern Recognition 46 (2013) 1851–1864.

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

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

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

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

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

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proc. IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.

[39] M. D. Zeiler, ADADELTA: An adaptive learning rate method, arXiv preprint arXiv:1212.5701.

[40] T. Tieleman, G. Hinton, Lecture 6.5 - RMSProp, COURSERA: Neural networks for machine learning, Tech. rep., Technical Report (2012).

Yachuang Feng is currently pursuing the Ph.D. degree with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an, Shaanxi, P. R. China.

Yuan Yuan is a full professor with the Chinese Academy of Sciences (CAS), China. Her major research interests include Visual Information Processing and Image/Video Content Analysis. She has published over a hundred papers, including over 70 in reputable journals, like IEEE transactions and Pattern Recognition, as well as conferences papers in CVPR, BMVC, ICIP, ICASSP, etc.

Xiaoqiang Lu is a Full Professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi, P. R. China.