CrossMark

Abnormal event detection and localization in crowded scenes based on PCANet

Tianlong Bao[1] · **Saleem Karmoshi**[1] · **Chunhui Ding**[1] ·
Ming Zhu[1]

Abstract Detecting and localizing abnormal events in crowded scenes still remains a challenging task among computer vision community. An unsupervised framework is proposed in this paper to address the problem. Low-level features and optical flows (OF) of video sequences are extracted to represent motion information in the temporal domain. Moreover, abnormal events usually occur in local regions and are closely linked to their surrounding areas in the spatial domain. To extract high-level information from local regions and model the relationship in spatial domain, the first step is to calculate optical flow maps and divide them into a set of non-overlapping sub-maps. Next, corresponding PCANet models are trained using the sub-maps at same spatial location in the optical flow maps. Based on the block-wise histograms extracted by PCANet models, a set of one-class classifiers are trained to predict the anomaly scores of test frames. The framework is completely unsupervised because it utilizes only normal videos. Experiments were carried out on UCSD Ped2 and UMN datasets, and the results show competitive performance of this framework when compared with other state-of-the-art methods.

✉ Tianlong Bao
   btl1991@mail.ustc.edu.cn

   Saleem Karmoshi
   Saleem@mail.ustc.edu.cn

   Chunhui Ding
   dchui@mail.ustc.edu.cn

[1] School of Information Science and Technology, University of Science and Technology of China,
   Hefei 230026, China

⏷ Springer

1 Introduction

In recent decades, the use of surveillance cameras has grown considerably due to a heightened security demand. This has significantly increased the number of videos produced, making the ability to automatically process these videos an urgent issue. A detailed survey [16] in this research field shows the growth in video publications in the recent years. The sheer number of videos being produced has turned the detection of abnormal events into a practical and challenging task. Abnormal event detection cannot be treated as a typical classification problem because the infinite types of abnormal events makes defining them impossible. In addition, negative samples are not sufficient for building models. In this case, building models using normal videos in an unsupervised way has drawn substantial attention. Models of normal events are first trained from positive samples, which means all the events that do not fit the normal models are abnormal events. The process of running an unsupervised abnormal event detection model can be summarized as, first, extracting features to represent normal video frames; second, building normal models using features of positive samples; and third, checking test frames against normal models.

Much research on unsupervised abnormal event detection has been published in recent years. In [10], object-of-interest trajectories are extracted and designated as the representation of normal events. Trajectories that do not follow the normal patterns are treated as anomalies. How well this method performs depends on the accuracy of the object tracking. In [13], spatial-temporal changes are used to learn a set of sparse basis combinations, and test samples are checked by reconstruction error.

Lately, the use of deep learning algorithms has taken off in the computer vision community, and the performance of image classification has reached impressive new levels. In the context of abnormal event detection, deep learning tools extract high-level features from videos in order to detect these abnormal events. In [19], a deep learning model auto-encoder is learned and serves to extract features, while Gaussian models are trained to detect anomalies. Method in [8] utilizes the PCANet to extract global features of video frames from saliency information (SI) and multi-scale histogram of optical flow (MHOF), so it has limitation in detecting and localizing anomalies in local areas. In order to detect and localize abnormal events simultaneously and make the best of PCANet, a new method is proposed in this paper. The PCANet is revised to extract local information from raw motion information of video sequences rather than from higher level features, i.e., SI and MHOF. Then corresponding one-class classifiers are used to detect anomalies.

The remainder of the paper is organized as follows: In Section 2, related work is covered in order to give background on the proposal. The proposed method is discussed in Section 3. Followed by experimental results and their analysis in Section 4. Finally, conclusions are presented in Section 5.

2 Related work

With the wide use of surveillance systems around the world, computer vision researchers have taken a keen interest in algorithms for intelligent video processing. One practical application of this research is the ability to detect abnormal events in surveillance videos, such that users can be alerted to suspicious activity as it occurs.

The detection of abnormal events in videos involves identifying events with abnormal patterns. The lack of negative samples has led to the use of unsupervised learning methods

as a sensible solution. In order to build normal models from training videos, extracting discriminative features becomes crucial to the success of the methods. Algorithms for abnormal event detection can be categorized into one of two types based on their methods of feature extraction: trajectory-based algorithms and spatio-temporal domain-based algorithms.

In [10, 18, 21], extracted trajectories of normal events represent normal patterns, and trajectories that do not follow the normal patterns are treated as anomalies. However, there are some limitations to applying the trajectory-based methods to crowded scenes because of the occlusion among multiple moving objects severely influences the efficiency and accuracy of object tracking and the extraction of trajectories of multiple objects. Social attribute-aware force model (SAFM) are proposed in [26] to incorporate social characteristics of crowd behaviors to improve the description of interactive behaviors, which effectively reduces the influence of occlusion between multiple moving objects. And in [24], 3-D discrete cosine transform (DCT) is used to design a robust multi-object tracker to associate the targets in different frames.

For spatio-temporal domain-based algorithms, features extracted based on spatio-temporal methods represent normal videos, and then normal models are learned in order to detect abnormal events. For example, extracted HOG [7, 9], optical flow [11, 15], or multi-scale histogram of optical flow [6] represents videos, and then normal models are learned in order to detect abnormal events. Various models are used to depict the features based on spatio-temporal changes. In [1], a location-based approach is proposed for illustrating behavior. In the study [2], a learned nonparametric model analyzes statistics of a multi-scale local descriptor. The study [15] proposes a social force model that estimates interaction forces between moving particles. The [12] studies the use of a joint detector that works with temporal and spatial anomalies, as well as the use of mixture of dynamic textures (MDT) models. In [1, 25], normal patterns are fitted into a Markov random field. Latent Dirichlet allocation models are utilized to detect abnormal events in [15, 20]. In [5, 6, 12, 13], sparse representations of normal events are proposed to reconstruct test frames.

3 Proposed method

Figure 1 shows the overall framework of the proposed method, which can be regarded as a kind of spatio-temporal domain-based method. Abnormal events differ from normal events in their motion patterns, so the first step is to extract optical flows from the video frames. Next, the calculation of the magnitude of optical flow at each pixel forms an optical flow map which can be regarded as the spatio-temporal representation of video frames. Because abnormal events generally occur in local regions, optical flow maps are uniformly partitioned to non-overlapping patches to extract local and discriminative information. Patches at same spatial coordinates in optical flow maps are then used to learn models. A set of PCANet models are learned separately based on patches at corresponding spatial coordinates. For each patch, removed block-wise histograms represent each block in the patch, as opposed to linking them all to a single vector. After obtaining training features, one-class classifiers are trained. This method also utilizes simple and efficient classifiers based on k-medoids, as discussed in Section 3.2.

The extraction of block-wise features is the first step for a test frame. Then, one-class classifiers calculate anomaly scores of each block to form a map. However, in crowded scenes, high anomaly scores of some normal isolated blocks may occur due to the occlusion and interference among multiple moving objects, as this leads to a high false positive rate.
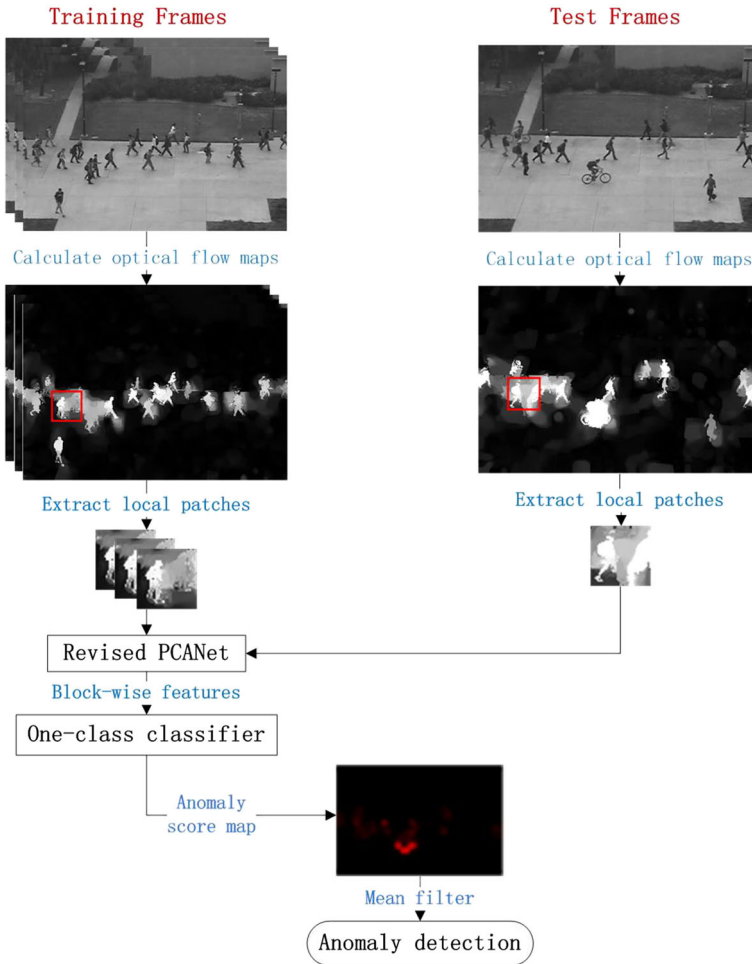
Fig. 1 Framework of the proposed method. Optical flow maps are firstly calculated from video frames. Then block-wise features are extracted by revised PCANet from local patches. Finally, detection results are obtained by the operation of one-class classifier and mean filter

The influence by misclassified isolated points is efficiently eliminated in this method by processing anomaly score maps through a mean filter. Experimental results demonstrate the overall effectiveness of the method.

3.1 Feature extraction based on PCANet

As a part of the trending deep learning field, PCANet was a new deep learning network proposed in [4]. Compared with other deep learning networks, such as convolutional deep neural network (ConvNet) that requires obscure expertise and large amounts of labeled training data, PCANet trains more easily and better suits the issue at hand. The structure of a two-stage PCANet is illustrated in Fig. 2. Similar to ConvNet, two stages of PCANet is a model with two layers. PCA filters in different stages is linearly cascaded to extract
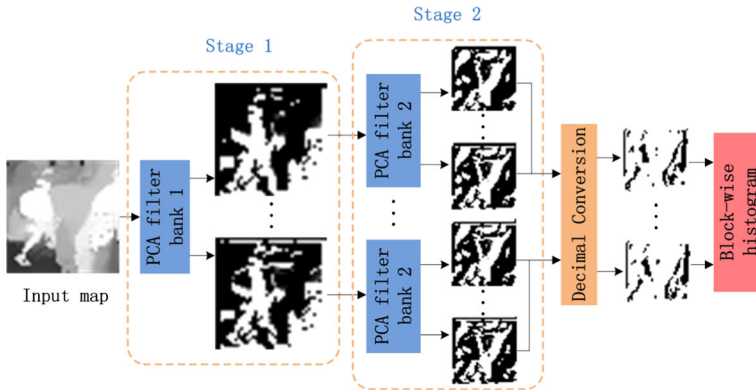
Fig. 2 Structure of a Two-Stage PCANet. Input is the optical flow map and output is the extracted feature. Features of input maps are extracted by three operations: PCA filters, binary hashing, and histogram

high-level features. And two-stage PCANet is more effective than one-stage PCANet in extracting higher level features as shown in [4].

Optical flow maps are uniformly partitioned to non-overlapping patches referred to as sub-maps with size $h \times w$. And sub-maps at same spatial coordinates are gathered temporally for training. Suppose there are N sub-maps for each spatial location, denoted as $X = [X_1, X_2, ..., X_N] \in \mathbb{R}^{hw \times N}$. For a given sub-map X_i, a patch with size $k_1 \times k_2$ is taken around each point, and all overlapping patches of X_i are collected as $X_{i,1}, X_{i,2}, ..., X_{i,(h-k_1+1) \times (w-k_2+1)}$, where each $X_{i,j}$ denotes the jth vectorized patch in X_i. By subtracting patch mean from each patch, we obtain $\bar{X}_i = [\bar{X}_{i,1}, \bar{X}_{i,2}, ..., \bar{X}_{i,(h-k_1+1) \times (w-k_2+1)}]$. Similar matrices for all sub-maps can be obtained using the same approach. Putting all the matrices together forms:

$$\bar{X} = [\bar{X}_1, \bar{X}_2, ..., \bar{X}_N] \in \mathbb{R}^{k_1 k_2 \times Nq} \tag{1}$$

where $q = (h - k_1 + 1) \times (w - k_2 + 1)$. The PCA filters in the first stage are expressed as:

$$W_l^1 = vec2mat_{k_1,k_2}(q_l(\bar{X}\bar{X}^T)) \in \mathbb{R}^{k_1 \times k_2}, l = 1, 2, ..., L_1 \tag{2}$$

where L_1 is the number of PCA filters in the first stage, $q_l(\bar{X}\bar{X}^T)$ denotes the lth principal eigenvector of $\bar{X}\bar{X}^T$, $vec2mat_{k_1,k_2}(v)$ is a function that reshapes the vector $v \in \mathbb{R}^{k_1 k_2}$ to a matirx $W \in \mathbb{R}^{k_1 \times k_2}$.

The process of the second stage is similar to the first stage. Given the lth filter W_l^1 and X_i from the first stage, output is $X_i^l = X_i * W_l^1$. By collecting all patches in X_i^l and subtracting patch mean from each patch, $Y^l = [Y_1^l, Y_2^l, ..., Y_N^l] \in \mathbb{R}^{k_1 k_2 \times Nq}$ is obtained for all the filter outputs and is concatenated as:

$$Y = [Y^1, Y^2, ..., Y^{L_1}] \in \mathbb{R}^{k_1 k_2 \times L_1 Nq} \tag{3}$$

The PCA filters of the second stage are obtained as:

$$W_\ell^2 = vec2mat_{k_1,k_2}(q_l(YY^T)) \in \mathbb{R}^{k_1 \times k_2}, \ell = 1, 2, ..., L_2 \tag{4}$$

For each input X_i^l in the second stage, a single integer-valued map is obtained with:

$$T_i^l = \sum_{\ell=1}^{L_2} 2^{\ell-1} H(X_i^l * W_\ell^2) \tag{5}$$

where $H(\cdot)$ is a function whose value is one for positive entries and zero for all others. Each map of T_i^l is partitioned into B blocks. Histograms of the decimal values in each block are computed and concatenated into one vector in traditional PCANet. In order to extract local information from optical flow maps and localize abnormal motion patterns, histograms of each block are not concatenated into one vector. Histograms of all the partitioned blocks are obtained as $BlockHist_b(T_i^l), b = 1, 2, ..., B$. Given a input sub-map X_i, the features are defined as:

$$f_{i,b} = [BlockHist_b(T_i^1), ..., BlockHist_b(T_i^{L_1})] \in \mathbb{R}^{2^{L_2} \times L_1}, b = 1, 2, ...B \qquad (6)$$

The local blocks can be overlapping or non-overlapping. In this method, the blocks are kept as non-overlapping. Assume the size of optical flow maps is 240×360, and the size of each sub-map is 40×40, so a set of PCANet models can be obtained as $PCANets = \{PCANet_{i,j}\}, i = 1 \sim 6, j = 1 \sim 9$. Additionally, for each sub-map X_k, the PCANet model, extracts a corresponding set of block-wise features, denoted as $\{f_{k,b}\}, b = 1, 2, ...16$.

3.2 Abnormal events detection by one-class classifier

The training dataset contains only normal events, so one-class classifier is more suitable in this case than other classifiers. Suppose there are N normal training optical flow maps, and each training sample is partitioned to M non-overlapping sub-maps. Now, N sub-maps at same spatial coordinates are gathered temporally for training. Then, N feature vectors are extracted by PCANet for each block location, denoted as $F = \{f_1, f_2, ..., f_N\}$. The K-mediods clustering algorithm is then used to divide F into k clusters. The k cluster centers can be obtained as $C = \{c_1, c_2, ..., c_k\}$ for each block location. Given a test sample, PCANet extracts feature vectors for each block, and anomaly scores are calculated as:

$$g(x) = min(x - c_i)^T (x - c_i) \qquad (7)$$

The scores are arranged according to the spatial location of each block in order to construct an anomaly score map. To eliminate the noise of isolated points, anomaly score maps are processed by a mean filter, and $g'(x)$ is obtained. Then, the classifier is defined as:

$$h(x) = \begin{cases} normality & g'(x) < \theta \\ anomaly & g'(x) \geq \theta \end{cases} \qquad (8)$$

4 Experiments

Experiments were carried on UCSD Ped2 [14] and UMN [17] datasets. Comparisons with other state-of-the-art methods were also made, and experimental results demonstrate the superiority of the proposed method.

4.1 Experiments on UCSD Ped2 dataset

The UCSD Ped2 dataset [14] was acquired with a stationary camera at $10\ fps$, with resolution of 240×360. The video footage was split into various clips with approximate lengths of 200 frames. In the normal setting, the video contains only pedestrians. Events are considered abnormal when abnormal pedestrian motion patterns are present or when non pedestrian
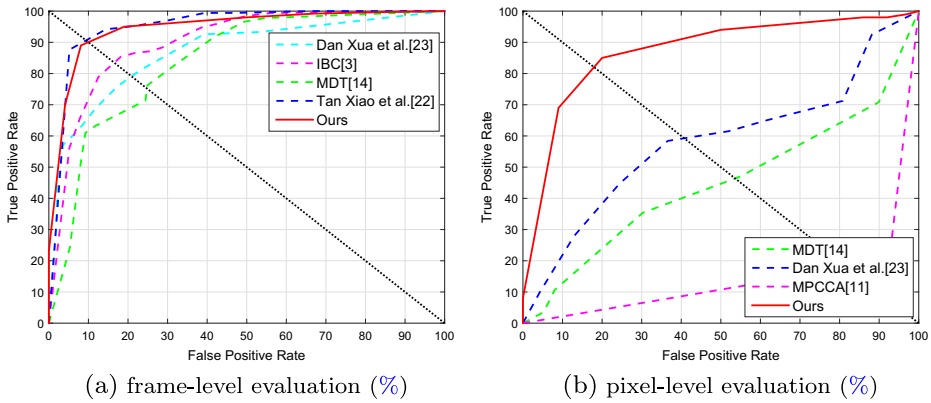
(a) frame-level evaluation (%)   (b) pixel-level evaluation (%)

Fig. 3 ROC curves of UCSD Ped2. **a** ROC curves based on frame-level evaluation. **b** ROC curves based on pixel-level evaluation

objects are in the walkways. The training set contains 16 normal short clips, and the testing set contains 12 clips in which some of the frames include one or more abnormal events. Similar to [14], receiver operating curve (ROC) and equal error rate (EER) were analyzed at both the frame and pixel levels, and comparisons with other state-of-the-art methods were examined.

In frame-level evaluation, a frame is considered correctly detected if at least one pixel is judged as an anomaly, and an exact location of the anomaly is not necessary. Therefore, it is possible for some true positive detections to identify abnormal events in the wrong location. In pixel-level evaluation, a frame is considered detected correctly if at least 40 percent of the truly anomalous pixels have been identified. In order to create an ROC curve, a set of false positive and true positive rates are extracted from the detection procedure repeated on multiple thresholds.

For UCSD Ped2, where the frame size is 240×360, optical flow maps with the same size are calculated. The size of sub-maps was set as 40×40, the size of mean filters was set as 3×3, and the PCANet parameters were set as $flitersize = 5 \times 5$, $blocksize = 10 \times 10$, L_1 is set to 8 to reserve 90 % of the variance, K-mediods parameter k is empirically set equal to L_1, L_2 is set to 6 to restrict the dimension of extracted features and reserve enough variance. The original optical flow map was partitioned into 54 sub-maps, and the corresponding 54 PCANet models were learned. Each sub-map was divided into 16 blocks and features with

Table 1 Comparison of EER on UCSD Ped2 (%)

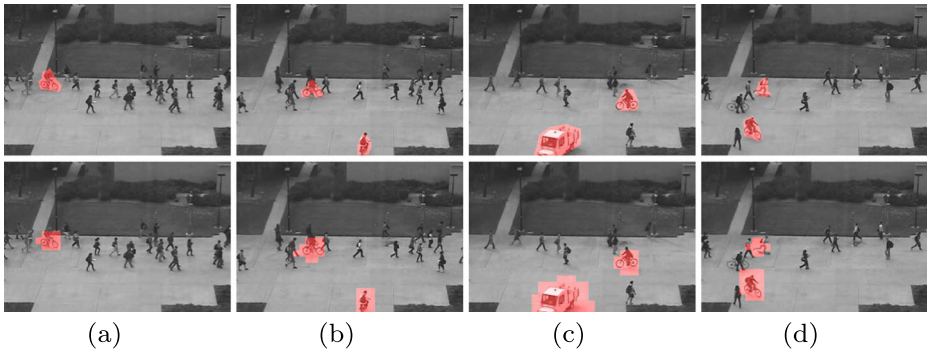| Method	Frame-level	Pixel-level
MDT [14]	24	54
IBC [3]	13	26
MPCCA [11]	30	71
Li et al. [12]	18.5	29.9
Tan Xiao et al. [22]	10	17
Dan Xua et al. [23]	20	42
Ours	10	18

Fig. 4 Examples of anomaly detection and localization. Red regions are abnormal pixels. Each column of the frame is the same. The top row of frames shows the ground truth, and the bottom row of frames shows the abnormal events detected by our method

dimension of $2^{L_1}L_2 = 512$ were extracted to represent each block. Finally, one $24 * 36$ anomaly score map was obtained for each frame.

Threshold θ was altered to produce ROC curves for frame-level and pixel-level evaluations. And the threshold can be set according to the application's requirements to get reasonable false positive rate and true positive rate. ROC curves and comparisons with other methods are illustrated in Fig. 3.

The comparisons of EER with other methods are shown in Table 1. Smaller EER indicates better performance.

Experimental results show that this method achieves respectable performance in both detecting and localizing abnormal events in surveillance videos. Figure 4 provides a portion of detection results of this method.
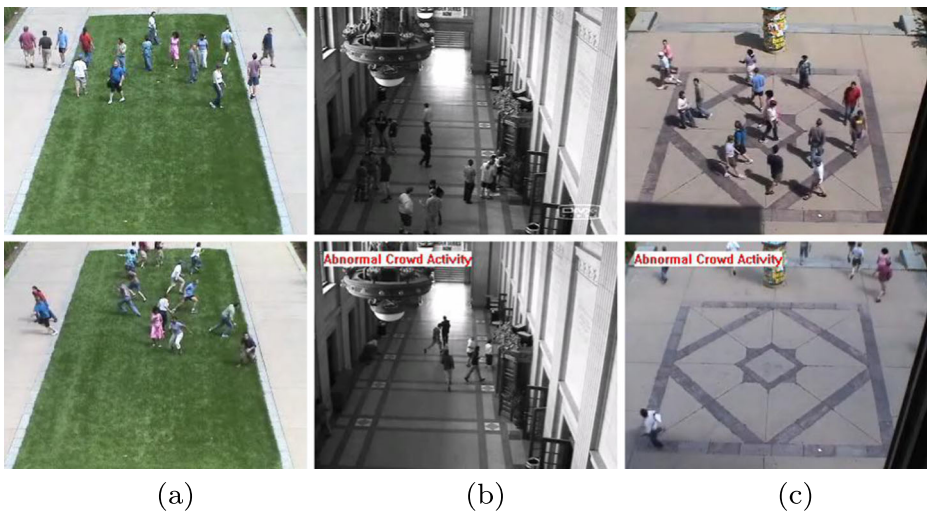

Fig. 5 Examples of Each Scene in UMN Dataset. The frames in the top row show normal crowd behavior, and the frames in the bottom row show abnormal behavior
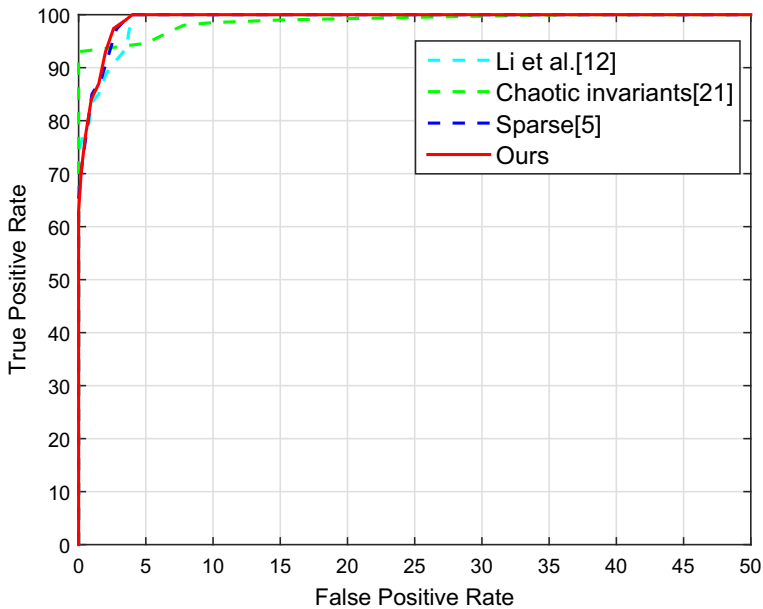
Fig. 6 ROC curves of UMN Dataset (%)

4.2 Experiments on UMN dataset

The UMN dataset [17] consists of three scenes containing crowds executing rapid escape behaviors, which are considered as abnormal events. Scene (a) contains 1453 frames, scene (b) contains 4144 frames and scene (c) contains 2142 frames, all of which have a resolution of 240×320. Figure 5 shows snapshots from each of the three scenes.

PCANet models and one-class classifiers are trained by using the first 600 normal frames of each scene, and other frames are used for testing. The UMN dataset contains no pixel-level ground truth, so the EER in frame level is used to evaluate the method. The parameter settings of experiments on UMN and UCSD Ped2 dataset are the same. In UMN dataset, rapid escape behaviors of crowds are considered as abnormal events. And the escape behaviors of each individual can be regarded as local anomalies. Our method is able to extract high-level information from local regions, so abnormal events in local areas can be detected to generate reasonable results in UMN dataset. ROC curves and comparisons with other methods are illustrated in Fig. 6. The comparisons of EER with other methods are shown in Table 2.

Table 2 Comparison of EER on UMN Dataset (%)

| Method	EER
Li et al. [12]	3.7
Chaotic invariants [21]	5.3
Sparse [5]	2.8
Ours	2.6

5 Conclusion

The proposed method involves an unsupervised learning framework for abnormal event detection and localization in surveillance videos. Optical flows are calculated to represent the motion information in video sequences. The deep learning model, PCANet, is used to extract high-level features from low-level optical flows. To localize abnormal events, the PCANet is revised to extract local information from video frames. Experimental results show that the method can effectively detect and localize abnormal events in surveillance videos.

References

1. Benezeth Y, Jodoin PM, Saligrama V, Rosenberger C (2009) Abnormal events detection based on spatio-temporal co-occurences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2458–2465
2. Bertini M, Del Bimbo A, Seidenari L (2012) Multi-scale and real-time non-parametric approach for anomaly detection and localization. Comput Vis Image Understand 116(3):320–329
3. Boiman O, Irani M (2007) Detecting irregularities in images and in video. Int J Comput Vis 74(1):17–31
4. Chan TH, Jia K, Gao S, Lu J, Zeng Z, Ma Y (2015) Pcanet: A simple deep learning baseline for image classification? IEEE Trans Image Process 24(12):5017–5032
5. Cong Y, Yuan J, Liu J (2011) Sparse reconstruction cost for abnormal event detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3449–3456
6. Cong Y, Yuan J, Tang Y (2013) Video anomaly search in crowded scenes via spatio-temporal motion context. IEEE Trans Inf Forensic Secur 8(10):1590–1599
7. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol 1, pp 886–893
8. Fang Z, Fei F, Fang Y, Lee C, Xiong N, Shu L, Chen S (2016) Abnormal event detection in crowded scenes based on deep learning. Multimedia Tools and Applications pp 1–23
9. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models, vol 32, pp 1627–1645
10. Khalid S (2010) Activity classification and anomaly detection using m-mediods based modelling of motion patterns. Pattern Recogn 43(10):3636–3647
11. Kim J, Grauman K (2009) Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2921–2928
12. Li W, Mahadevan V, Vasconcelos N (2014) Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Anal Mach Intell 36(1):18–32
13. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. In: IEEE International Conference on Computer Vision (ICCV), pp 2720–2727
14. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1975–1981
15. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 935–942
16. Popoola OP, Wang K (2012) Video-based abnormal human behavior recognition review. IEEE Trans Syst, Man, Cybern, Part C: Appl Rev 42(6):865–878
17. Raghavendra R, Bue AD, Cristani M (2006) Unusual crowd activity dataset of University of Minnesota. http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi
18. Rasheed N, Khan SA, Khalid A (2014) Tracking and abnormal behavior detection in video surveillance using optical flow and neural networks. In: 28th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp 61–66
19. Sabokrou M, Fathy M, Hosseini M (2015) Real-time anomalous behavior detection and localization in crowded scenes. CoRR 1511.07425

20. Wang X, Ma X, Grimson E (2007) Unsupervised activity perception by hierarchical bayesian models. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 1–8
21. Wu S, Moore BE, Shah M (2010) Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: IEEE Computer on Computer Vision and Pattern Recognition (CVPR), pp 2054–2060
22. Xiao T, Zhang C, Zha H (2015) Learning to detect anoMalies in surveillance video. IEEE Signal Process Lett 22(9):1477–1481
23. Xu D, Song R, Wu X, Li N, Feng W, Qian H (2014) Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. Neurocomputing 143:144–152
24. Yuan Y, Fang J, Wang Q (2015) Online anomaly detection in crowd scenes via structure analysis. IEEE Trans Cybern 45(3):548–561
25. Zhang D, Gatica-Perez D, Bengio S, McCowan I (2005) Semi-supervised adapted hmms for unusual event detection. In: IEEE Conference on Computer Vision and Pattern Recognition, vol 1. IEEE, pp 611-618
26. Zhang Y, Qin L, Ji R, Yao H, Huang Q (2015) Social attribute-aware force model: exploiting richness of interaction for abnormal crowd detection. IEEE Trans Circ Syst Video Technol 25(7):1231–1245
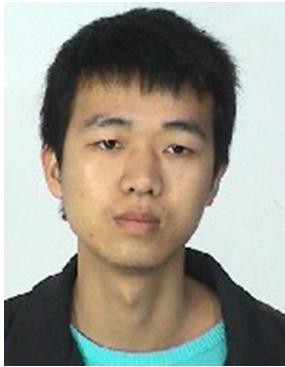
Tianlong Bao acquired his B.S. degree in the University of Science and Technology of China in 2013. Currently, he is a Ph.D. candidate in the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include computer vision, video analysis and image processing.
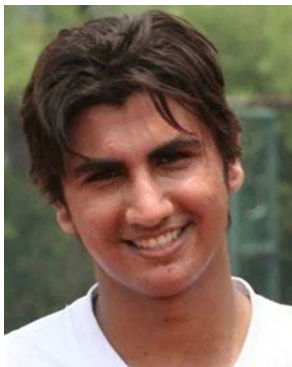
Saleem Karmoshi received his B.S. degree from the University of Science and Technology of China in 2012. And now, he is a Ph.D. candidate in the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include cloud computing, deep learning and sensor analysis.

Chunhui Ding acquired his B.S. degree in the University of Science and Technology of China in 2012. Currently, he is a Ph.D. candidate in the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include face recognition and video analysis.

Ming Zhu is a professor in the department of automation, University of Science and Tech- nology of China. He received the B.S. degree, M.S. degree and Ph.D. degree from the School of Computer Science and Tech- nology in the University of Science and Technology of China, in 1981, 1989 and 2002 respectively. From 1997 to 1998, he was a visiting scholar in the Hong Kong Polytechnic University. Prof. Zhu published more than 80 SCI/EI papers in last five years. His research interests include video analysis, video retrieval and distributed computation.