Abnormal event detection via multi-kernel learning for distributed camera networks

Tian Wang[1], Jie Chen[2], Paul Honeine [3], Hichem Snoussi[3],

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

[2] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, U.S

[3] Institut Charles Delaunay-LM2S-UMR STMR 6279 CNRS, University of Technology of Troyes, Troyes, France

E-mail: wangtian@buaa.edu.cn, dr.jie.chen@ieee.org, {paul.honeine,hichem.snoussi}@utt.fr

Abstract

Distributed camera networks play an important role in public security surveillance. Analyzing video sequences from cameras set at different angles will provided enhanced performance for detecting abnormal events. In this paper, an algorithm is proposed to detect the abnormal event under distributed camera networks via multi-kernel learning. The visual event is presented by the histogram of the optical flow orientation descriptor, and then a multi-kernel strategy that takes the multi-view scene into account is given to improve the detection accuracy. The nonlinear one-class SVM algorithm with the constructed kernel is then trained to detect abnormal frames of video sequences. We validate and evaluate the proposed method on the video surveillance dataset PETS, and obtained promising results.

Index Terms

Distributed camera networks, abnormal detection, optical flow, one-class SVM.

I. Introduction

Detecting abnormal events via video sequence analysis is crucial for public security management. In complex scenes, distributed camera networks with overlapping views are capable to obtain additional information to surveil the movement of the crowds compared to the single camera setting. In the PETS dataset [1], camera locations are illustrated in Fig. 1. Several normal and abnormal scenes from different angles of view are shown in Fig. 2. In Figs. 2(a)(c)(e), all people are walking in different directions, which are considered as normal. In Figs. 2(b)(d)(f), people are walking or running towards the same

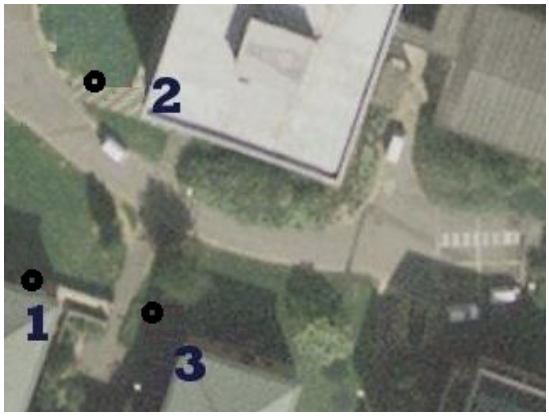

Fig. 1: The plan of the multi-camera localizations in the PETS dataset. Three cameras are set in the campus.

directions, implying that people are attracted by some particular events, consequently, these scenes are considered as abnormal. The scenes are captured by different cameras. Fig. 2(a)(b) are captured by camera 1 which is set at the side of the road, the movement of the people are well captured. Figs. 2(c)(d) are captured by camera 2 which faces the movement direction of the people, and there is occlusion of the individuals in this view. Figs. 2(e)(f) are captured by camera 3 which is also set at the side of the road, with larger distance. The purpose of the distributed camera surveillance is to detect abnormal events by benefitting the multi-view video sequences. Distribute camera networks are now widely used for surveillance application.

In the literature papers [2], [3], the framework for multiple pedestrian tracking by using overlapping cameras was presented. In [4], several major challenges in distributed video processing, including robust and computationally efficient inference, opportunistic and parsimonious sensing were discussed. Large-scale video networks starts to play an important rule for video surveillance, object recognition, abnormal event detection and people tracking in crowded environments.

Modeling the movement feature of pixels is fundamental for detecting the abnormal event. In [5], a method that tracked the local spatio-temporal interest points was proposed, and the abnormal activity was indicated by uncommon energy-velocity of the feature points. In [6], a spatio-temporal descriptor was computed based on computing the histograms of optical flow in the neighborhood of detected points. In [7], dense points were sampled from each frame and were tracked based on displacement information from an optical flow field. But for the crowd event analysis, it is difficult to obtain the pre-detected pixels of the blob due to the occlusion of the individuals.

In order to deal with the uncertainty of observations existing in video events, Bayesian modeling approaches such as hierarchical Dirichlet processes were used in [8], probabilistic latent semantic analysis was used in [9]. A method based on the variable-duration hidden Markov model was proposed in [10], where the durations of states were modeled except for the transitions between states, and the temporal understanding of the structure of complex events was tackled. Latent Dirichlet allocation (LDA) was also a typical standard topic model which has been used to model video clips as being derived from

Fig. 2: Examples of the normal and abnormal scenes captured by distributed cameras of PETS dataset. (a)(c)(e): The people are walking in different directions, the normal scenes of *Time 14-55* sequence; (b)(d)(f): The people are walking in the same direction, the abnormal scenes of *Time 14-17* sequence; (a)(b): scenes captured by camera 1; (c)(d): scenes captured by camera 2; (e)(f): scenes captured by camera 3.

a bag of topics drawn from a fixed set of proportions [11]. In [12], the covariance matrix descriptor fusing the optical flow to encode moving information of a frame was presented. These work focused on the single view scene video analyzing, the feature abstraction method or the event models were heavily researched in the abnormal detection problem. Moreover, the abnormal event detection problems in distributed camera networks have a considerable room.

The rest of the paper is organized as follows. In Section 2, the optical flow-based feature is presented. In Section 3, the abnormal detection framework based on one-class SVM classification method is presented, and then, a multi-kernel strategy is proposed to deal with abnormal event detection problem for distributed camera networks. In Section 4, the experimental

results are illustrated and discussed. Finally, Section 5 concludes the paper and gives a perspective of future work.

II. Feature selection for abnormal detection

Horn-Schunck (HS) [13] is chosen to compute the optical flow which represents the movement information. The HS method formulates the optical flow as a global energy functional for the gray image sequence:

$$E = \int\int [(I_x u + I_y v + I_t)^2 + \alpha(\|\nabla u\|^2 + \|\nabla v\|^2)]dxdy \tag{1}$$

where I_x, I_y and I_t are the derivatives of the image intensity values along the horizontal direction x, vertical direction y and time t dimension, respectively. u, v are the horizontal and vertical optical flow. α is a regularization constant.

Based on the optical flow, the histogram of the optical flow orientation (HOFO) [14] is computed to fuse the movement as a high dimension vector. A 2×2 rectangular cell HOFO descriptor of the image is shown in Fig. 3. The orientation is computed by horizontal and vertical optical flow, and then it is voted into n bins. Each cell contains h_c pixels in height dimension, and w_c in width dimension. A block contains $h_b \times w_b$ cells, it is set as 2×2 in this paper.

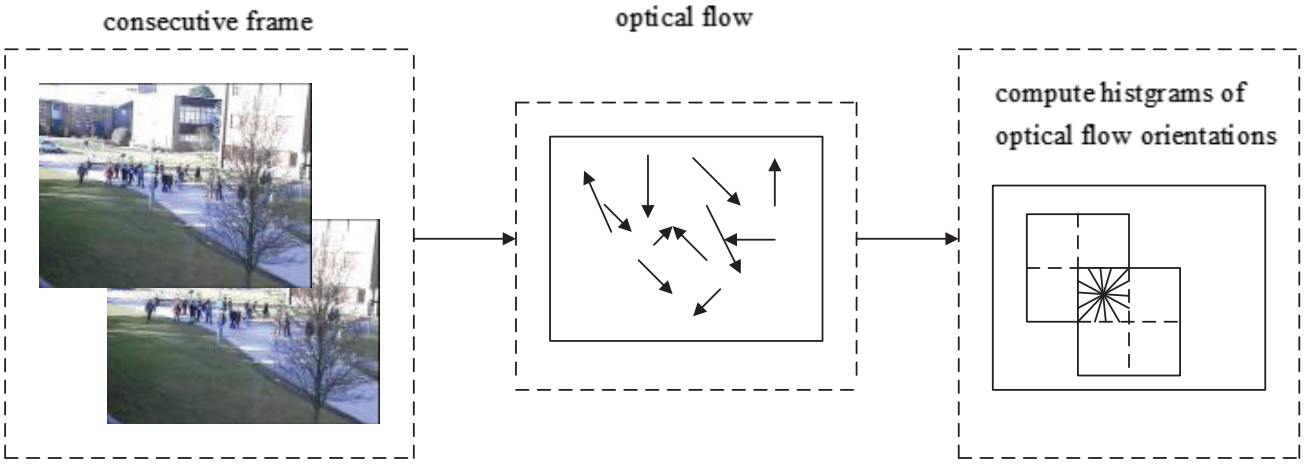

Fig. 3: Histogram of the optical flow orientation (HOFO) feature descriptor of the image.

III. One-class SVM with multiple kernels

The problem of non-linear one-class SVM [15], [16] can be cast as a quadratic programming problem:

$$\begin{aligned} &\min_{\omega,\xi,\rho} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho, \\ &\text{subject to} \quad \langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i)\rangle \geq \rho - \xi_i, \quad \xi_i \geq 0. \end{aligned} \tag{2}$$

where $\boldsymbol{x_i} \in \mathcal{X}$ with $i = 1 \ldots n$ are n training samples in the original data space $\mathcal{X}$. ξ_i is the slack variable for penalizing the outliers. The hyperparameter $\nu \in (0, 1]$ is the weight for controlled slack variable, it tunes the number of acceptable outliers. $\langle \boldsymbol{w}, \Phi(\boldsymbol{x_i})\rangle - b = 0$ is the decision hyperplane. Φ is defined for building the non-linear classification problems, and it is a map

from the set of the original input data $\mathcal{X}$ to a feature space $\mathcal{H}$ where the classification problem has a linear solution. The inner product in space $\mathcal{H}$ is defined by the kernel function $\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \Phi(\boldsymbol{x_i}) \cdot \Phi(\boldsymbol{x_j}) \rangle$. Introducing the Lagrangian multipliers α_i, the decision function in the data space $\mathcal{X}$ is:

$$f(\boldsymbol{x}) = \mathbf{sgn}(\sum_{i=1}^{n} \alpha_i \kappa(\boldsymbol{x_i}, \boldsymbol{x}) - \rho), \tag{3}$$

where $\boldsymbol{x}$ is a vector in the input data space $\mathcal{X}$, κ is the kernel function implicitly mapping the data into a high dimensional feature space where a linear classifier can be designed.

The Gaussian kernel is chosen to handle movement feature in this work. It is a semi-positive definite kernel that satisfies the Mercer condition [17], [18], and defined by

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\Big(- \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\Big), \qquad (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{X} \times \mathcal{X}, \tag{4}$$

where $\boldsymbol{x}_i, \boldsymbol{x}_j$ are the data in the original data space $\mathcal{X}$, the variance σ indicates the scale factor at which the data should be clustered.

For constructing a more representative and discriminative feature descriptor for the distributed camera network, we take the scene captured by each view as a partial feature. The multi-kernel strategy considers the linear combination of candidates kernels:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{s=1}^{m} \mu_s \kappa_s(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{5}$$

where $\kappa_s, s = 1, \cdots, m$ are m candidate kernels that satisfy the Mercer condition, and μ_s are nonnegative factors. Consequently, their combination κ is also a semi-positive definite kernel. In this expression, the Gaussian kernel is adopted with:

$$\kappa_s(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\Big(- \frac{\|\boldsymbol{x}_{i_{[s]}} - \boldsymbol{x}_{j_{[s]}}\|^2}{2\sigma^2}\Big). \tag{6}$$

The kernels $\kappa_s, s = 1, \cdots, m$ are the Gaussian kernels in this paper. Each sample vector $\boldsymbol{x}$ consists of m parts $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$.

For a given scene monitored by multi-camera, supposing that a set of training frames are obtained, based on the one-class SVM hypothesis, the abnormal behavior is the sample deviating from the training set. For example, a plaza is monitored by 3 cameras as show in Fig. 2. If $s = 1$, the frame captured by camera 1 is selected. We preset the weight μ_s according to the characteristics of the image to tune the importance of each view. By considering this combination, the resulting kernel outperform each kernel κ_s used individually. Based on the histogram of the optical flow orientation feature descriptor and the nonlinear one-class SVM, the abnormal event detection method is summarized in Algorithm 1, and explained as following.

Algorithm 1 Abnormal event detection algorithm.

Require:

Image set captured by the cameras.

1: Computing the optical flow of the training frame set $[I_1^{v_i}, \ldots, I_m^{v_i}]$, where v_i means the scene is monitored by camera i via the HS optical flow method:

$$[I_1^{v_i}, \ldots, I_m^{v_i}] \longrightarrow [O_1^{v_i}, O_2^{v_i}, \ldots, O_m^{v_i}]$$

2: Computing the histogram of the optical flow orientation (HOFO) of the image in different views:

$$[O_1^{v_i}, O_2^{v_i}, \ldots, O_m^{v_i}] \longrightarrow [H_1^{v_i}, \ldots, H_m^{v_i}]$$

3: The feature sample of the image k under c distributed camera network is:

$$H_k = [H_k^{v_1}, H_k^{v_2}, \ldots, H_k^{v_c}]$$

4: Training feature sample are learned by the nonlinear one-class SVM method to obtain the support vectors:

$$[H_1, \ldots, H_m] \longrightarrow \text{support vector } [S_1^{v_i}, \ldots, S_{m_1}^{v_i}]$$

5: Each incoming frame $H_{p,\ldots,q}$ is classified by the decision function of one-class SVM.

6: The normal event or abnormal event is detected.

Step 1: The optical flow features are computed. The training frame set $[I_1^{v_i}, \ldots, I_m^{v_i}]$ monitored by multi-camera network describing the normal behavior is available. $I_j^{v_i}$ means that the j-th image is captured by the i-th camera. The Horn-Schunck method is applied to acquire the optical flow feature. This step can be presented as:

$$[I_1^{v_i}, \ldots, I_m^{v_i}] \longrightarrow [O_1^{v_i}, O_2^{v_i}, \ldots, O_m^{v_i}]$$

where $[I_1^{v_i}, \ldots, I_m^{v_i}]$ are m training frames of the i-th view, $[O_1^{v_i}, O_2^{v_i}, \ldots, O_m^{v_i}]$ are the corresponding optical flows.

Step 2: The second step consists of calculating the histogram of optical flow orientation (HOFO) of the training frames. It can be generalized as:

$$[O_1^{v_i}, O_2^{v_i}, \ldots, O_m^{v_i}] \longrightarrow [H_1^{v_i}, \ldots, H_m^{v_i}]$$

where $[O_1^{v_i}, O_2^{v_i}, \ldots, O_m^{v_i}]$ are the optical flow of the training frames captured by camera i. $[H_1^{v_i}, \ldots, H_m^{v_i}]$ are the corresponding histogram of optical flow orientation (HOFO) feature.

Step 3: Fusing the 1-st to c-th HOFO feature to one high dimension feature vector. It is described in the equation below:

$$H_k = [H_k^{v_1}, H_k^{v_2}, \ldots, H_k^{v_c}]$$

where H_k is the feature vector fusing the multi-view movement information. $H_k^{v_1}$ is the HOFO feature of the k-th image in view i.

Step 4: Nonlinear one-class SVM is applied on the training frame HOFO descriptor in multi-view to obtain the support vector. It is described as the following:

$$[H_1, \ldots, H_m] \longrightarrow \text{support vector } [S_1, \ldots, S_{m_1}]$$

where $[H_1, \ldots, H_m]$ are the histogram of optical flow orientation descriptors of the training frames under multi-view environment. $[S_1, \ldots, S_{m_1}]$ are support vectors that are the minority of the training vectors contribute to the decision function.

Step 5: In the online detection phase, based on the support vectors obtained in the taring step, the one-class SVM classifies each incoming frame feature $[H_p, \ldots, H_q]$. Thus, the normal or abnormal event of the frame is classified, thus the abnormal event is detected.

IV. Abnormal events detection results

We then conduct experiments to evaluate the performance of the one-class SVM classification method for abnormal frame event detection with a distributed camera network. The PETS [1] dataset is used for the evaluation purpose. In the experiments of PETS [1] dataset, each event is represented by 3 separately scenes, thus, the event is described by 3 separated HOFO features which is marked as "*3 HOFO*". Otherwise, we mark the feature as "*1 HOFO*". If the multi-kernel strategy is used, we mark it as "*3 kernels*", otherwise, we mark the kernel strategy as "*1 kernel*". The detection accuracy of the detection results are shown for the experiments.

The normal and abnormal event of sequence *Time 14-17* in the PETS dataset are shown in Fig. 4. The training samples and the normal samples for testing are the frames in 3 views chosen from the sequence (*Time 14-55*) where the people are walking in different directions. 400 training frames (Frame0000 to Frame0399), and 90 normal testing frames (Frame0400 to Frame0489) are selected from *Time14-55*. The abnormal testing samples are selected from the sequence (*Time 14-17*) where the people are walking or running in the same direction. 89 abnormal testing frames (Frame0000 to Frame0089) are selected from *Time14-17*.

The AUC (area under the curve) value of the abnormal detection results in different views and different multi-kernel strategies are shown in TABLE **??**. "Single View" means *"1 HOFO 1 kernel"* strategy, and "Multi-view" means *"3 HOFO 3 kernels"*. "View 1" means the scene is monitored by camera 1, the abnormal detection results are shown in Figs. 4(a)(b). "View 2" implies that the scene is monitored by camera 2, as show in Figs. 4(c)(d). "View 3" refers to the scene that is monitored by camera 3, as show in Figs. 4(e)(f). "Multi-view(1)" indicates in the multi-kernel strategy, $\mu_1 = 0.5$, $\mu_2 = 0.2$, $\mu_3 = 0.3$. "Multi-view(2)" means that in the multi-kernel strategy, $\mu_1 = \frac{1}{3}$, $\mu_2 = \frac{1}{3}$, $\mu_3 = \frac{1}{3}$. The ROC curve of *Time 14-17* scene results is shown in Fig. 5.

Fig. 4: Detection results of the normal and abnormal scenes of *Time 14-17* captured by distributed camera network. (a)(c)(e): The detection result of one normal frame. (b)(d)(f): The detection result of one abnormal frame.

View 2 has the lowest area under the ROC (receiver operating characteristic curve), since as previously mentioned, the camera 2 faces the movement direction of the crowd, the occlusion influences the computation of the optical flow. Thus the HOFO feature based on the optical flow cannot represent the accurate movement information. The results show that the abnormal detection algorithm of HOFO feature can obtain satisfactory detection results. Moreover, the multi-kernel strategy can generally improve the performance. The parameters of the multi-kernel influence the performance. Thus, the optimal parameters should be sought.

The experiment detecting the running activity as the abnormal event is shown in Fig. 6 and Fig. 7 . The normal event corresponds to the frames where the people are walking. The training data are selected from the sequence (*Time14-17* and

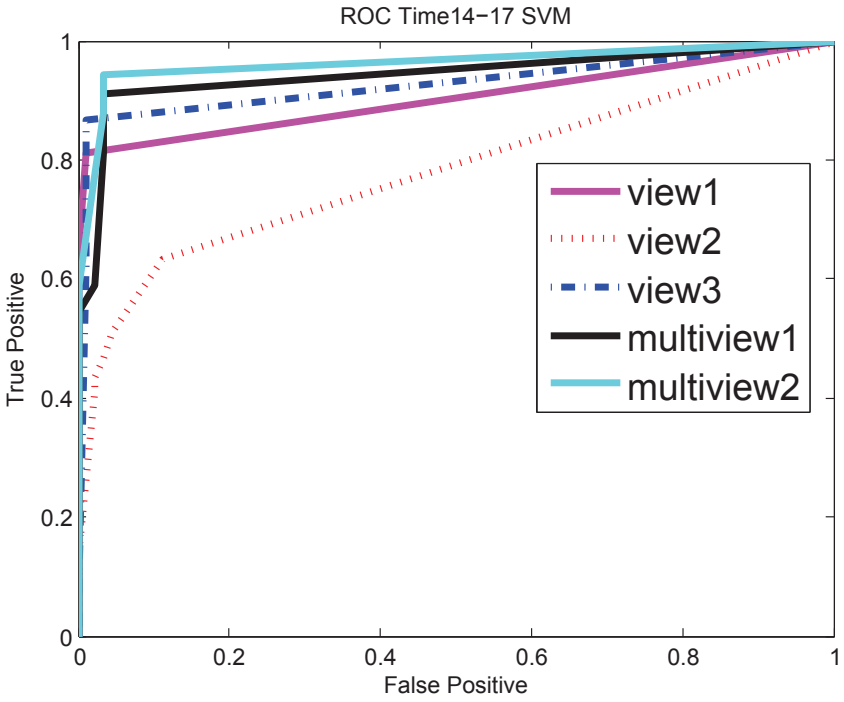

Fig. 5: ROC curve of abnormal detection result of sequence *Time 14-17* under different views or different kernel strategies.

TABLE I: The abnormal detection results of sequence *Time 14-17*. The comparison of the abnormal frame event detection results in single view scene and in distributed camera scenes via multi-kernel strategies.

Strategy	Area under ROC		
	View 1	View 2	View 3
Single View	0.9036	0.7807	0.9270
Multi-view(1)	0.9441		
Multi-view(2)	0.9643		

Time14-31) in PETS dataset where the individuals are walking in one direction. In this experiment, 61 frames (Frame0000 to Frame0060) where people are walking from left to right in sequence *Time14-17*, and 50 frames (Frame0000 to Frame0049) where the individuals are walking from right to left in sequence *Time14-31*. Corresponding, 104 normal samples and 118 abnormal frames in sequence *Time 14-16* are detected. The abnormal detection performance is improved by multi-kernel strategy also. For example, the AUC values of the the sequence where the individuals are moving from left to right are shown in TABLE **??**.

TABLE II: The abnormal detection results of sequence *Time 14-16*. The comparison of the abnormal frame event detection results in single view scene and in distributed camera scenes via multi-kernel strategies.

Strategy	Area under ROC		
	View 1	View 2	View 3
Single View	0.8637	0.8071	0.9205
Multi-view(1)	0.9312		
Multi-view(2)	0.9403		

V. Conclusions

A method for abnormal frame event detection with distributed camera networks is proposed in this paper. The histogram of optical flow is computed as the descriptor to represent the movement of a frame. A multi-kernel strategy is presented to benefit the characteristics of distributed camera network, and the performance of the abnormal detection results via nonlinear one-class SVM is improved. The benchmark dataset PETS has been tested to demonstrate the effectiveness of the proposed algorithm.

In the future work, the optimal coefficients of the multi-kernel strategy should be obtained automatically based on the scene, while these parameters were pre-set in current work. Additional datasets will be considered to show the advantages of distributed camera networks, such as single person action recognition or action tracking in multi-view scenes, etc.

Acknowledgment

This work is partially supported by the SURECAP CPER project (fonction de surveillance dans les réseaux de capteurs sans fil via contrat de plan Etat-Région) and the Platform CAPSEC (capteurs pour la sécurité) funded by Région Champagne-Ardenne and FEDER (fonds européen de développement régional), the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China (Grant No. U1435220).

References

[1] PETS, "Performance evaluation of tracking and surveillance (pets) 2009 benchmark data. multisensor sequences containing different crowd activities. http://www.cvg.rdg.ac.uk/pets2009/a.html," 2009.

[2] J. Wan and L. Liu, "Distributed bayesian inference for consistent labeling of tracked objects in nonoverlapping camera networks," *International Journal of Distributed Sensor Networks*, vol. 2013, 2013.

[3] W. Jiuqing and L. Achuan, "Multiple people tracking using camera networks with overlapping views," *International Journal of Distributed Sensor Networks*, vol. 2015, 2015.

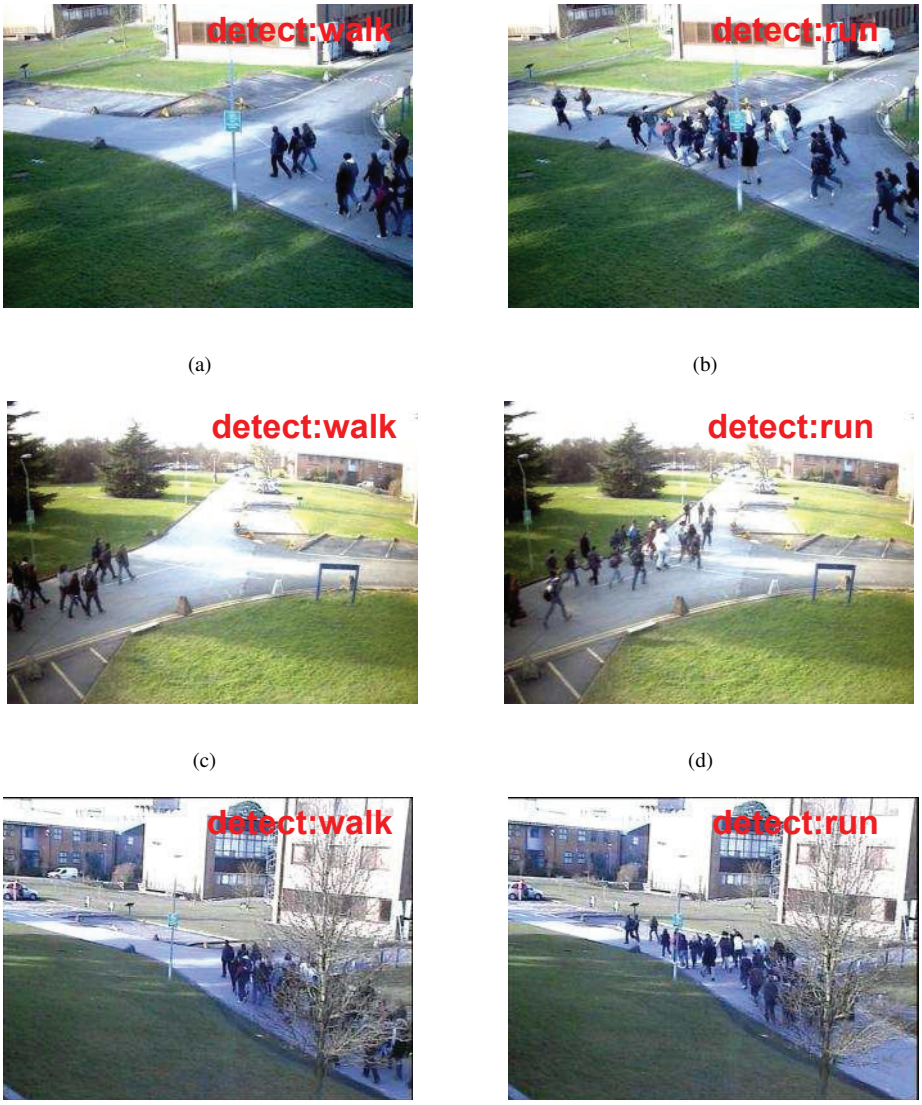

Fig. 6: Detection results of the normal and abnormal scenes of *Time 14-16* captured by thedistributed camera network, the individuals are moving from right to left. (a)(c)(e): The detection result of one normal frame, the individuals are walking in one direction. (b)(d)(f): The detection result of one abnormal frame, the individuals are running in one direction.

[4] A. C. Sankaranarayanan, R. Chellappa, and R. G. Baraniuk, "Distributed sensing and processing for multi-camera networks," in *Distributed Video Sensor Networks*. Springer, 2011, pp. 85–101.

[5] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 3161–3167.

[6] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid *et al.*, "Evaluation of local spatio-temporal features for action recognition," in *Proceedings of British Machine Vision Conference (BMVC)*, 2009.

[7] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*. IEEE, 2011, pp. 3169–3176.

[8] T. S. Haines and T. Xiang, "Delta-dual hierarchical dirichlet processes: A pragmatic abnormal behaviour detector," in *Proceedings of IEEE International*

Fig. 7: Detection results of the normal and abnormal scenes of *Time 14-16* captured by the distributed camera network, the individuals are moving from left to right. (a)(c)(e): The detection result of one normal frame, the individuals are walking in one direction. (b)(d)(f): The detection result of one abnormal frame, the individuals are running in one direction.

Conference on Computer Vision (ICCV), 2011, pp. 2198–2205.

[9] J. Varadarajan and J.-M. Odobez, "Topic models for scene analysis and abnormality detection," in *Proceedings of the 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, 2009, pp. 1338–1345.

[10] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 1250–1257.

[11] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 6, pp. 865–878, 2012.

[12] T. Wang, J. Chen, Y. Zhou, and H. Snoussi, "Online least squares one-class support vector machines based abnormal visual event detection," *Sensors*, no. 12, pp. 17 130–17 155, 2013.

[13] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1, pp. 185–203, 1981.

[14] T. Wang and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 988–998, June 2014.

[15] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[16] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "Svm and kernel methods matlab toolbox," Perception Systèmes et Information, INSA de Rouen, Rouen, France, 2005.

[17] V. Vapnik, *The nature of statistical learning theory*. Springer, 2000.

[18] V. N. Vapnik, *Statistical learning theory*. Wiley: New York, NY, USA, 1998.