

Real-Time Abnormal Event Detection in Complicated Scenes

Yinghuan Shi , Yang Gao

State Key Laboratory for Novel Software Technology
Nanjing Univeristy
Nanjing, China
yinghuan.shi@gmail.com , gaoy@nju.edu.cn

Ruili Wang

School of Engineering and Advanced Technology
Massey University
Palmerston North, New Zealand
r.wang@massey.ac.nz

Abstract—In this paper, we proposed a novel real-time abnormal event detection framework that requires a short training period and has a fast processing speed. Our approach is based on phase correlation and our newly developed spatial-temporal co-occurrence Gaussian mixture models (STCOG) with the following steps: (i) a frame is divided into non-overlapping local regions; (ii) phase correlation is used to estimate the motion vectors between successive two frames for all corresponding local regions, and (iii) STCOG is used to model normal events and detect abnormal events if any deviation from the trained STCOG is found. Our proposed approach is also able to update the parameters incrementally and can be applied in complicated scenes. The proposed approach outperforms previous ones in terms of shorter training periods and lower computational complexity.

Keywords—phase correlation; STCOG; abnormal event detection; real-time

I. INTRODUCTION

In recent years, abnormal event detection has attracted great research attention in computer vision. There are some research issues need to be addressed in this area such as (i) how to shorten training periods, (ii) how to reduce computational complexity. To address these issues, we propose a novel approach for real-time abnormal event detection in complicated scenes.

In general, previous approaches for abnormal event detection can be categorized into two groups: tracking-based and motion-based approaches. The tracking-based approaches [3], [11], [12], [15], [16] focus on the analysis of the trajectories of moving objects. However, real-time tracking of all moving objects in complicated scenes is too difficult to achieve in real-world situations [4]. Recently, motion-based approaches have been proposed to address the above problem, which can be classified into two sub-groups based on how to extract motion features: (i) the background-subtraction-based [4], [14]. These approaches are not able to perform well in complicated scenes either [14]. (ii) the optical-flow-based [7], [13], [17], [9], [2], [6], [8]. Few of them can both detect abnormal events in real-time and update parameters online.

Our novel approach is a kind of motion-based approach. Similar to other motion-based ones, a frame is divided into

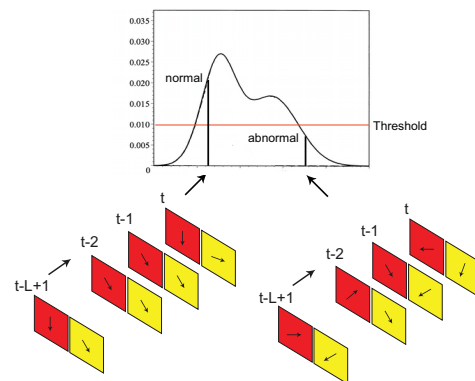


Figure 1. A pair-wise cuboid

non-overlapping local regions. Phase correlation is then applied to estimate the motion vectors between two successive frames for all corresponding local regions. After that, these frames can be seen as a 3D-spatial-temporal matrix, where each element of the matrix is a motion vector. As normal events always have regular spatial-temporal motion vectors, these normal events can be modeled by estimating the co-occurrence probability of the motion vectors between local regions with Spatial-Temporal Co-Occurrence Gaussian mixture models (STCOG). Similar with Markov Random Field (e.g., in [8]), in order to avoid falling into a high dimensional feature space, for a local region, pair-wise cuboids between it and its neighbors in the 3D-motion-vector matrix are modeled by STCOG. In Fig.1, the adjacent red and yellow regions within the recent L frames make up a pair-wise cuboid. The number of neighbors of a local region (N_c) is 4 in this paper. The Normal-State Probability (NSP) in the detection period of a local region is to average the 4 pair-wise co-occurrence probabilities. If NSP is high, the region is in a normal state. Our STCOG is trained with online K-means approximation [10], while NSPs of incoming cuboids are calculated in the detection period.

Our approach has the following *three advantages* comparing with previous ones: (i) lower computational cost due to adopting phase correlation and STCOG; (ii) the duration of training is shorter due to adopting the online K-means

approximation, and (iii) STCOG can also update parameters online and automatically.

II. PHASE CORRELATION

We apply phase correlation [5] to estimate the motion between two successive frames for every local regions. Phase correlation is based on the Fourier shift theorem and was originally proposed for translated images registration. In our approach, Fast Fourier Transform (FFT) is used to reduce the computational cost.

Supposing R_p^t and R_p^{t+1} are two sequentially observed regions at position p , time t and $t + 1$, respectively. The process is to (i) transform them to the frequency domain by applying discrete FFT, and denote them as f_p^t and f_p^{t+1} , respectively; (ii) normalize the phase correlation as follows:

$$\psi_p^{t+1} = (f_p^t \cdot f_p^{*,t+1}) / |f_p^t \cdot f_p^{*,t+1}| \quad (1)$$

note that \cdot is Hadamard product for f_p^t and $f_p^{*,t+1}$ (complex conjugate of f_p^{t+1}). (iii) obtain the normalized cross correlation ϕ_p^{t+1} by applying the Inverse Fast Fourier Transform (IFFT) for ψ_p^{t+1} . (iv) estimate the motion vector $(\Delta x, \Delta y)$ between R_p^t and R_p^{t+1} , where $(\Delta x, \Delta y) = \arg \max_{x,y} \phi_p^{t+1}$.

III. STCOG

We denote pair-wise cuboid $C_{p,q}^t$ at location p, q (p, q are the adjacent regions) and time t as follows: $C_{p,q}^t = \{M_p^t, \dots, M_p^{t-L+1}, M_q^t, \dots, M_q^{t-L+1}\}$, where M_p^t and M_q^t are motion vectors at location p, q and time t .

For $C_{p,q}^t$, the co-occurrence probability can be represented as follows:

$$P(C_{p,q}^t) = \sum_{i \leq K_{max}} \omega_{p,q}^{i,t} * g(C_{p,q}^t | \mu_{p,q}^{i,t}, \Sigma_{p,q}^{i,t}) \quad (2)$$

where i denotes the i^{th} Gaussian component in STCOG; g is Gaussian probability density function; $\omega_{p,q}^{i,t}$, $\mu_{p,q}^{i,t}$ and $\Sigma_{p,q}^{i,t}$ are the weight, mean value and covariance matrix of the i^{th} Gaussian in the mixture model at p, q and t . K_{max} is the maximum number of Gaussian components of STCOG. Finally, for each local region at p and t , we can calculate its NSP as follows:

$$P(C_p^t) = \frac{1}{N_c} \sum_{q \in N(p)} P(C_{p,q}^t) \quad (3)$$

where $q \in N(p)$ denotes that q is a neighbor of p , and N_c is the number of neighbors. The covariance matrix is set to be diagonal for reducing computational cost. The fixed length L is the duration in STCOG, which can be in the range of 3-8 frames. If L is too small, more false alarms can be caused, while if L is too large, it will increase computational cost. The maximum number of components K_{max} is determined by the complexity of a scene. $N_c = 4$, $L = 5$ and $K_{max} = 20$ are used in this paper.

In the training period, parameter estimation of Gaussian Mixture Models (GMM) traditionally uses Expectation

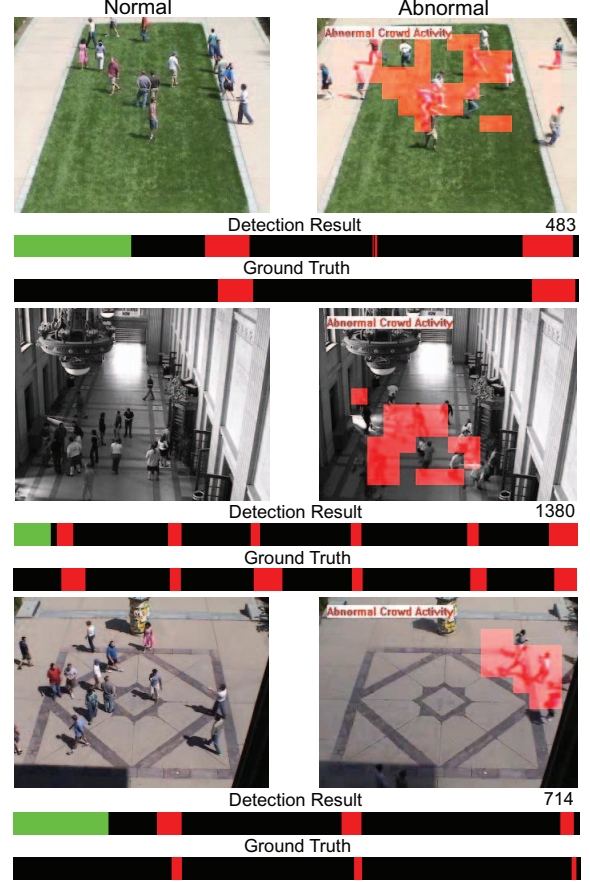


Figure 2. Detection results and ground truth on UMN dataset.

Maximization (EM) algorithm (e.g., in [14]), which is an iterative one with higher computational cost, and might converge into a local maximum. Inspired by [10], online K-means approximation is used to train our STCOG. For every incoming $C_{p,q}^t$, existing Gaussian components are checked and the k^{th} Gaussian with the highest probability in STCOG $k = \arg \max_i g(C_{p,q}^t | \mu_{p,q}^{i,t-1}, \Sigma_{p,q}^{i,t-1})$ is chosen, which should be within 2 standard deviations of this Gaussian simultaneously. After that, we can update the parameters of k^{th} matched Gaussian as follows:

mean value:

$$\mu_{p,q}^{k,t} = (1 - \beta) * \mu_{p,q}^{k,t-1} + \beta * C_{p,q}^t \quad (4)$$

covariance matrix:

$$\Sigma_{p,q}^{k,t} = (1 - \beta) * \Sigma_{p,q}^{k,t-1} + \beta * (\mu_{p,q}^{k,t-1} - C_{p,q}^t)^T (\mu_{p,q}^{k,t-1} - C_{p,q}^t) \quad (5)$$

weight:

$$\omega_{p,q}^{k,t} = (\omega_{p,q}^{k,t-1} + \Delta\omega) / (\sum_i \omega_{p,q}^{i,t-1} + \Delta\omega) \quad (6)$$

where β is the learning rate and $\Delta\omega$ ($0 < \Delta\omega < 1$) is the reward for this matched k^{th} Gaussian. For the

	Precision	Recall	AUC
Scene 1	0.9876	0.9520	0.9362
Scene 2	0.8618	0.9545	0.7759
Scene 3	0.9969	0.9209	0.9661
Social Force [9]	n/a	n/a	0.96
Pure Optical Flow [9]	n/a	n/a	0.84

Table I
PRECISION, RECALL AND AUC.

rest unmatched Gaussian, we only update their weights in STCOG like that: taking j^{th} Gaussian for example, $\omega_{p,q}^{j,t} = \omega_{p,q}^{j,t-1} / (\sum_i \omega_{p,q}^{i,t-1} + \Delta\omega)$. If the new observation $C_{p,q}^t$ matches none of all existing $K_e (K_e \leq K_{max})$ Gaussian components, we initialize a new Gaussian component and replace the lowest weight Gaussian (denoted as n^{th}) with it. Its mean value is set to $\mu_{p,q}^{n,t} = C_{p,q}^t$, and its covariance matrix and weight are initialized by the predefined values.

In the detection period, we calculate the NSP of a new cuboid. If NSP of a local region is lower than a predefined threshold, it will be considered as an abnormal event. Parameter updating in the detection period is the same as that in the training period.

IV. EXPERIMENTS

Our proposed method has been tested on two datasets: UMN dataset [1] and Adam's dataset [2]. Also, our approach has been compared with some previous algorithms in terms of computational complexity.

A. UMN Dataset

UMN dataset consists 3 different scenes of crowded escape events, and the total number of the frames of the video is about 2577 (483, 1380 and 714 for scenes 1-3, respectively). The original resolution of the UMN dataset is 320×240 , which is divided into 20×20 regions. 100, 90 and 120 frames are used to train STCOG for scenes 1-3, respectively. The detection results and ground truth labels can be seen in Fig.2, where the red regions are the abnormal events detected, and the label of "Abnormal Crowd Activity" in the top left corner is the ground truth label. The green bar denotes the frames for training STCOG; the black bar indicates normal events; the red one shows abnormal events.

In the detection results of all three scenes, all results are correct except only one false alarm that is generated at frame 307, 309 and 310 in scene 1 as shown in Fig.3. The reason for this false alarm is that: two pedestrians in the central region of the scene are walking towards each other, and partial occlusion occurs between them from frame 304. Thus, when extracting motion vectors of this region,



Figure 3. False alarm in scene 1

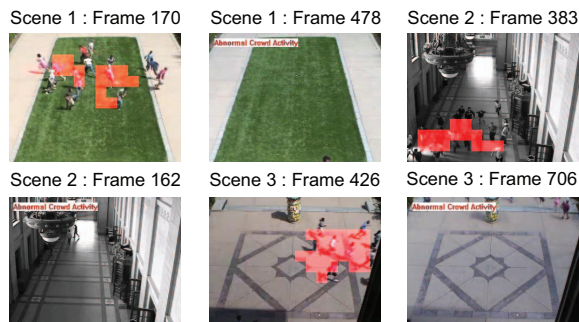


Figure 4. Time lags of the ground truth

our model considers it as a person walks and turns around suddenly.

Precision and recall of our approach are listed in Table.I on scenes 1-3 respectively. However, due to some time lags in the ground truth label, it deteriorates the precision and recall of our method (refer to Fig.4). For example, a suddenly running already occurs but the ground truth label does not appear (e.g., frame 170 in scene 1), some frames without obvious abnormal events have been labeled as abnormal (e.g., frame 478 in scene 1). We calculate the AUC (Area Under ROC) for every scene and compare our performance with [9], which also runs on the same dataset. The AUC of scene 2 of our approach is less than theirs mainly due to the time lags. While in scenes 1 and 3, we can archive similar detection results to theirs. However, the computational cost of our approach is much less than theirs.

B. Adam's Dataset

One video (43 mins about 64900 frames) provided by Adam *et al.* [2] is about an exit gate of a subway. The first five minutes are used to train STCOG. Compared with [2], our proposed approach can detect some spatial-temporal abnormal events that cannot be detected by their approach such as loitering, turning around, etc. In our experiments, we resized the frames from 512×384 to 320×240 and divided the new frames into 20×20 regions. Although there

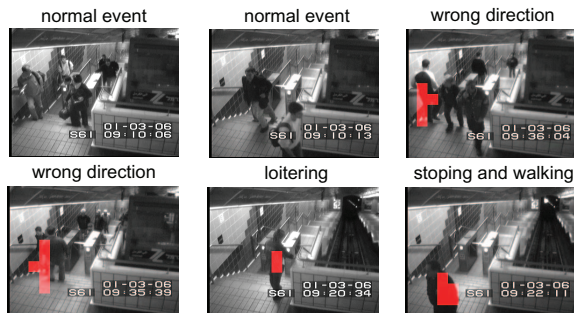


Figure 5. Abnormal events detected

are no ground truth labels in this video, the spatial-temporal abnormal events can still be detected by our approach (refer to Fig.5 for example).

C. Computational Costs

In motion-based approaches [6], [8], [2], majority of time is normally spent on motion estimation as the complexity of event classification is much smaller. Thus, we only analyze the time complexity of phase correlation. According to Section 2, assuming the amount of pixels in a local region is $N \times N$, time complexity of FFT or IFFT is $\Theta(N^2 \log N)$, element-wise matrix multiplication and maximum searching are both $\Theta(N^2)$. Therefore, the complexity of phase correlation is $\Theta(N^2 \log N)$ in all. Compared with block-matching [2], [8], which is a brute force searching method with high computational cost up to $\Theta(N^4)$, it often has to seek a tradeoff between accuracy and processing speed. Spatial-temporal gradient [17], [7], which can be fast calculated, has very high dimensional feature space and need some additional tools for reducing its dimension. In our experiments, using MATLAB code on Intel Duo 2.33GHZ CPU, we can archive 8-9 FPS by our approach, which can satisfy the requirements of real-time abnormal event detection *without dedicated hardware*. In addition, compared with [17], [9], [7], our approach can online update the model parameters, while compared with [8], [6], [13], our approach requires a shorter training period.

V. CONCLUSION

We proposed a novel framework for abnormal event detection that requires a short training period and has a fast processing speed. Phase correlation and STCOG are introduced here for their low computational cost. Compared with previous approaches, our approach can archive similar results but requires less computational cost. Our approach can satisfy the requirements of real-time abnormal event detection *without dedicated hardware*.

ACKNOWLEDGMENT

We would like to acknowledge support for this project from the National Science Foundation of China (NSFC

grant No.60775046 and No.60721002) and the National Grand Fundamental Research 973 Program of China (grant No.2009CB320700). Yinghuan Shi would like to thank Yongyan Cui and Xiaojia Pu for their comments on experiments, and Amit Adam for providing videos.

REFERENCES

- [1] Unusual event datasets of university of Minnesota, from <http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi>.
- [2] A. Adam, E.Rivlin, I.Shimoshoni and D. Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors, *PAMI*, vol.30, no.3, pp.555-560, 2008.
- [3] A. Basharat and A. Gritai and M. Shan, Learning object motion patterns for anomaly detection and improved object tracking, *CVPR*, pp.1-8, 2008.
- [4] Y. Benezeth, P.-M. Jodoin, V. Saligrama and C. Rosenberger, Abnormal events detection based on spatio-temporal co-occurrences, *CVPR*, pp.2458-2465, 2009.
- [5] E. De Castro and C. Morandi, Registration of translated and rotated images using finite Fourier transforms, *PAMI*, vol.9, no.5, pp.700-703, 1987.
- [6] T. Hospedales, S. Gong and T. Xiang, A Markov clustering topic model for mining behaviour in video, *ICCV*, 2009.
- [7] L. Kratz and K. Nishio, Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models, *CVPR*, pp.1446-1453, 2009.
- [8] J. Kim and K. Grauman, Observe locally, infer globally : a space-time MRF for detecting abnormal activities with incremental updates, *CVPR*, pp.2921-2928, 2009.
- [9] R. Mehran, A. Oyama and M. Shah, Abnormal crowd behavior detection using social force model, *CVPR*, pp.935-942, 2009.
- [10] C. Stauffer and E. Grimson, Learning patterns of activity using real-time tracking, *PAMI*, vol.22, no.8, pp.747-757, 2000.
- [11] X. Wang, X. Ma and E. Grimson, Unsupervised activity perception by hierarchical Bayesian models, *CVPR*, pp.1-8, 2007.
- [12] X. Wang, K. Tieu and E. Grimson, Learning semantic scene models by trajectory analysis, *ECCV*, no.3, pp.110-123, 2006.
- [13] X. Wang, K. Tieu and E. Grimson, Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models, *PAMI*, vol.31, no.3, pp.539-555, 2009.
- [14] T. Xiang and S. Gong, Video Behaviour profiling for anomaly detection, *PAMI*, vol.30, no.5, pp.893-908, 2008.
- [15] D. Zhang, D. Gatica-Perez, S. Bengio and I. McCowan, Semi-supervised adapted HMMs for unusual event detection, *CVPR*, pp.611-618, 2005.
- [16] X. Zhang, H. Liu, Y. Gao and D. Hu, Detecting abnormal events via hierarchical Dirichlet processes, *PAKDD*, pp.278-289, 2009.
- [17] H. Zhong, J. Shi and M. Visontai, Detecting unusual activity in video, *CVPR*, 819-826, 2004.