Deep Representation for Abnormal Event Detection in Crowded Scenes

Yachuang Feng[1,2]
fengyachuang@opt.cn

Yuan Yuan[1]
yuany@opt.ac.cn

Xiaoqiang Lu[1]
luxq666666@gmail.com

[1]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
[2]University of Chinese Academy of Sciences, Beijing 100049, P. R. China

ABSTRACT

Abnormal event detection is extremely important, especially for video surveillance. Nowadays, many detectors have been proposed based on hand-crafted features. However, it remains challenging to effectively distinguish abnormal events from normal ones. This paper proposes a deep representation based algorithm which extracts features in an unsupervised fashion. Specially, appearance, texture, and short-term motion features are automatically learned and fused with stacked denoising autoencoders. Subsequently, long-term temporal clues are modeled with a *long short-term memory* (LSTM) recurrent network, in order to discover meaningful regularities of video events. The abnormal events are identified as samples which disobey these regularities. Moreover, this paper proposes a spatial anomaly detection strategy via manifold ranking, aiming at excluding false alarms. Experiments and comparisons on real world datasets show that the proposed algorithm outperforms state of the arts for the abnormal event detection problem in crowded scenes.

Keywords

Video surveillance; crowded scene; abnormal event detection; deep representation

1. INTRODUCTION

Abnormal event detection aims to identify abnormal events in surveillance videos automatically. In recent years, anomaly detection has attracted a surge of interest from both academia and industry. However, it is still quite challenging to design a general framework for abnormal event detection. This is because the definition of an anomaly varies in different applications. An abnormal event in one scene probably is normal in another. To deal with this problem, one common solution is to extract normal event patterns from training (normal) videos, and detect abnormal ones by finding events which diverge from these patterns.

In the past decade, significant improvements have been achieved by designing various valid video event features. For example, tracking [6] is usually carried out to analyse individual moving objects. Trajectory-based features are born with simplicity and high-level semantics. But they are limited by factors such as shadows and occlusion in crowded scenes. As alternatives, recent works extract event patterns at pixel-level, 2D patches or 3D blocks. For example, Reddy et al. [9] split video frames into non-overlapping cells, from which features of motion, size and texture are extracted. In [13], spatio-temporal interest points are described by the *histogram of gradient* (HoG) and *histogram of optical flow* (HoF). There are also many other video event features widely used for anomaly detection. One commonality of these features is that they are hand-crafted. Generally, designing an effective descriptor is time-consuming and difficult. Moreover, it is hard to decide which kind of feature is suitable for a specific situation.

Nowadays, deep learning [12] has become a hot topic, which learns features automatically from raw data. In many computer vision tasks, deep learning has demonstrated strong performance, such as object detection, image segmentation, and activity recognition. The reason why deep learning achieves inspiring performance is that meaningful and discriminative features can be adaptively extracted through multi-layer non-linear transformations. In consequence, abnormal event detection is expected to benefit from deep learning algorithms.

Recently, Xu et al. [11] propose a novel deep learning framework for abnormal event detection. In [11], appearance, motion, and their joint representations are learned via three *stacked denoising autoencoders* (SDAEs). Then anomaly scores are calculated by three one-class support vector machine classifiers with these learned features. Although they have learned effective video event features, there are still some shortcomings. 1) Only short-term motion information is considered, i.e. optical flow maps. As is well known, video events are highly complex. Static frames and short-term motion clues are insufficient to represent video events. Therefore, long-term temporal clues are crucial for learning meaningful regularities of video events. 2) Context information is ignored. In crowded scenes, moving objects are highly related to each other. In this case, discovering relationships among adjacent video events ought to be helpful for anomaly detection in crowded scenes.

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2967290

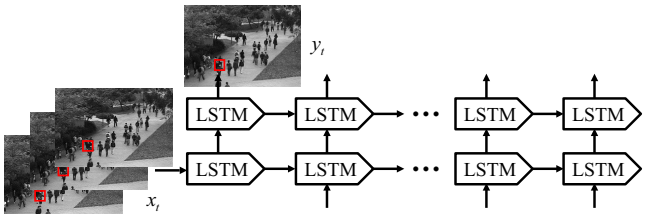

Figure 1: The illustration for learning long-term dependencies in video events.

In order to cope with these issues, this paper proposes a different deep learning framework for abnormal event detection. Similarly with [11], an SDAE model is trained to learn and fuse features from appearance, texture, and short-term motion clues. As for long-term temporal clues, Li et al. [4] model dynamics and appearances of crowds with the dynamic texture, which is a linear dynamical system. Due to the complexity of video events, one linear model might be insufficient. To improve the representation ability, multiple dynamic textures are mixed to model video events in [4]. However, this introduces more parameters and increases the complexity. In contrast, this paper models event regularities with a recurrent neural network, which adopts non-linear transformation and considers both the current input and the previous hidden state. Specifically, LSTM [10] is utilized in this paper because it is capable of learning long-term dependencies in video events. These dependencies are learned by an LSTM based prediction framework, as is shown in Fig. 1. The last three video frames are taken as the input to the LSTM, and the current frame is the predicted output.

The LSTM neural network learns meaningful regularities of video events, and dissimilarities in testing videos are treated as temporal anomalies. In order to incorporate the spatial context of individual moving objects, Li et al. [4] compute feature responses for surrounding annular windows, and detect anomalies as center-surround salient objects. In this paper, a much simpler and more effective strategy is proposed to make use of spatial contextual information. Based on the input video frame, a graph is constructed. In this graph, only adjacent samples in the spatial context are connected, and the weights are estimated as similarities in the feature space. After that, a graph-based manifold ranking algorithm is designed to identify anomalies. Since normal video frames contain no anomalies, we select some totally normal samples as queries and rank the rest according to their relevances to these queries. Samples with small ranking scores are treated as spatial anomalies. Furthermore, an adaptive weight is assigned to each sample in order to improve the distinctiveness between abnormal and normal events. As a result, false alarms are excluded and the identification of anomalies gets easier.

2. THE PROPOSED METHOD

2.1 Video Event Modeling

Inspired by [11], we use SDAEs to adaptively extract video event features. Specifically, we extract features from appearance, texture, and short-term motion clues, which are respected by image sequences, gradients and optical flows on

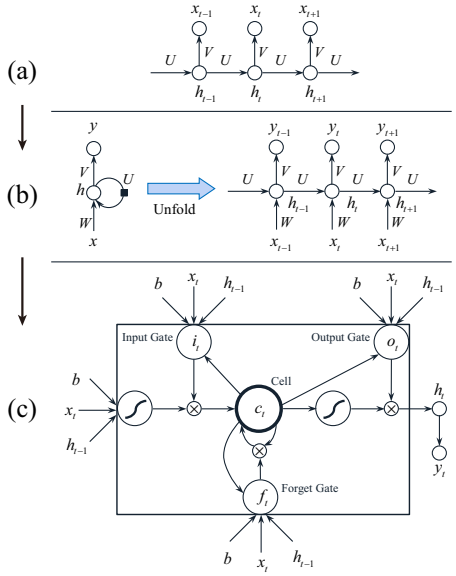

Figure 2: Frameworks for video event modeling. U, V, W are weight matrices. (a) Linear dynamic texture; (b) RNN; (c) LSTM.

horizontal and vertical directions, respectively. After that, another SDAE is learned to fuse these features.

In order to learn long-term dependencies in video events, an LSTM based prediction model is developed. Fig. 2 illustrates our motivation. Compared with linear dynamic textures, RNN is much more preferable because of its non-linear character, as well as taking both observations and hidden states as inputs to predict time dependencies. Equations related to the LSTM based prediction model are listed as follows.

$$g_t = \varphi(\mathbf{W}_{xg}\mathbf{x}_t + \mathbf{W}_{hg}\mathbf{h}_{t-1} + \mathbf{b}_g), \quad (1)$$
$$i_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i), \quad (2)$$
$$f_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f), \quad (3)$$
$$o_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o), \quad (4)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot g_t, \quad (5)$$
$$\mathbf{h}_t = o_t \odot \varphi(C_t), \quad (6)$$
$$\mathbf{y}_t = \mathbf{W}_y\mathbf{h}_t + \mathbf{b}_y. \quad (7)$$

φ denotes the *tanh* function, and σ represents the *sigmoid* function. $\odot$ stands for the element-wise multiplication. These gates (i_t, f_t, and o_t) ensure the capture of temporal dependencies. The objective function is minimizing

$$L(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2}\sum_t \|\mathbf{y}_t - \hat{\mathbf{y}}_t\|_2^2, \quad (8)$$

where $\mathbf{y}_t$ and $\hat{\mathbf{y}}_t$ are the predicted and true outputs at time t, respectively.

2.2 Abnormal Event Detection

With these normal event regularities, video events which disobey these rules are treated as temporal anomalies. The temporal anomaly score for sample $\mathbf{y}$ is computed as

$$\mathcal{T}(\mathbf{y}) = \frac{1}{2}\|\mathbf{y} - \hat{\mathbf{y}}\|_2. \quad (9)$$

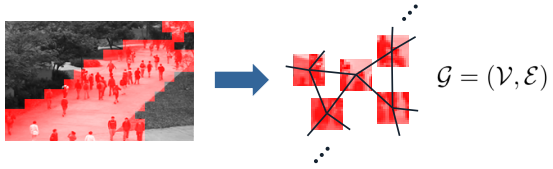

Figure 3: The graph model for manifold ranking.

In order to improve the detection accuracy using the spatial contextual information, a manifold ranking algorithm is designed. Firstly, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed from the motion region of an input frame. Vertices $\mathcal{V}$ are samples in the motion region, and edges $\mathcal{E}$ only connect adjacent samples in the spatial context, as is shown in Fig. 3. These edges are weighted by

$$\mathbf{W}_{ij} = \begin{cases} e^{-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\sigma_y^2}}, & \mathbf{y}_j \in \mathcal{N}(\mathbf{y}_i) \\ 0, & otherwise, \end{cases} \quad (10)$$

where $\mathcal{N}(\mathbf{y}_i)$ stands for neighbors of $\mathbf{y}_i$, and σ_y is a scale parameter. Subsequently, a ranking method is developed. Since normal video frames contain no anomalies, we select several normal ones as queries by

$$\widetilde{\mathcal{T}} = \begin{cases} 1, & \mathcal{T} < \tau \\ 0, & otherwise, \end{cases} \quad (11)$$

where τ is a threshold which can be estimated from training samples. The rankings for testing samples are evaluated by solving

$$\arg\min_{\mathcal{S}} \frac{1}{2} \sum_{i,j} \mathbf{W}_{ij} \|\mathcal{S}_i - \mathcal{S}_j\|^2 + \frac{\mu}{2} \sum_i \gamma_i \|\mathcal{S}_i - \widetilde{\mathcal{T}}_i\|^2. \quad (12)$$

The parameter μ controls the balance of the smoothness and fitting constraints. γ_i is an adaptive weight, which aims to improve the discrimination between abnormal and normal events. To this end, γ_i is set as

$$\gamma_i = \begin{cases} 1, & \mathcal{T} < \tau \\ \mathcal{T}_i / Z, & otherwise, \end{cases} \quad (13)$$

where Z is the maximum value of $\mathcal{T}$. Mathematically, Eq. (13) assigns large weights to normal ($\mathcal{T}_i < \tau$) and suspicious events (large $\mathcal{T}$), while small weights to others. In this case, temporal anomaly scores are preserved with large γ_i, and false alarms are smoothed. The spatial anomaly scores can be computed by

$$\mathcal{S} = \frac{\mu}{2} (\mathbf{D} - \mathbf{W} + \frac{\mu}{2} \mathbf{\Gamma})^{-1} \mathbf{\Gamma} \widetilde{\mathcal{T}}, \quad (14)$$

where $\mathbf{\Gamma}$ and $\mathbf{D}$ are both diagonal matrices, with $\mathbf{\Gamma}_{ii} = \gamma_i$ and $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$.

The final detection results are computed as

$$\mathcal{AE} = \mathcal{T} \odot (1 - \overline{\mathcal{S}}), \quad (15)$$

where $\overline{\mathcal{S}}$ is the normalized value of $\mathcal{S}$.

3. EXPERIMENTS

In order to verify the effectiveness of the proposed algorithm, qualitative and quantitative comparisons are provided on two real world datasets.

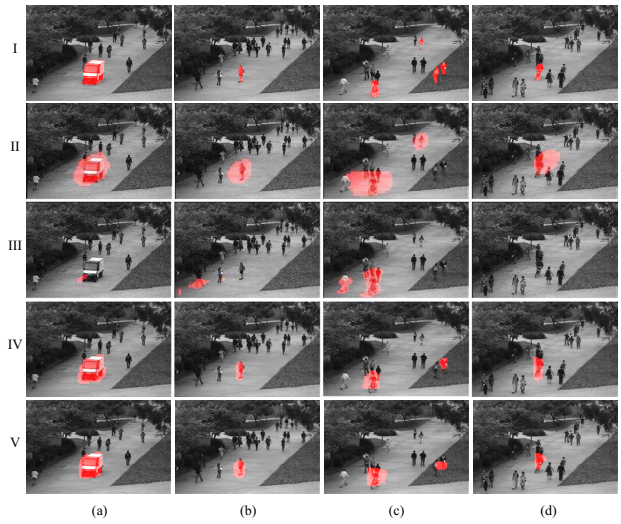

Figure 4: Examples of detected abnormal events on the UCSD ped1 dataset. (I) the ground-truth; (II) the MDT algorithm [7]; (III) the SF-MPPCA algorithm [7]; (IV) the SRC algorithm [2]; and (V) the proposed algorithm.

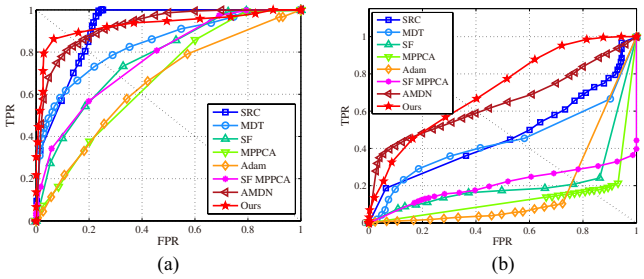

Figure 5: ROC curves for the UCSD Ped1 dataset. (a) Frame-level ROC; (b) Pixel-level ROC.

3.1 UCSD Ped1 Dataset

The UCSD Ped1 dataset[1] is taken from a surveillance camera overlooking a pedestrian walkway. The crowd density in the scene varies from sparse to extremely crowded. There are 34 and 36 video clips in the training and testing sets, respectively. In the testing dataset, all clips have frame-level labels, and 10 clips are provided with pixel-level ground-truth. Each short video clip is composed of 200 frames, with spatial resolution of 158×238.

We compare the proposed algorithm with some state of the art algorithms. Some visualized comparison results are shown in Fig. 4, in which detected abnormal events are labeled with red masks. The abnormal events include car, skater, biker, and so on. The proposed algorithm achieves competitive performance compared with others. The quantitative evaluations are measured on frame and pixel levels.

ROC Curve Comparison According to [4], as long as one pixel is detected as an anomaly in a testing frame, it is labeled as an abnormal frame. If the ground-truth is also abnormal, it is a true positive. Otherwise, it is a false pos-

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

Table 1: Comparisons on the UCSD Ped1 dataset.

	Frame-level		Pixel-level	
	EER	AUC	EDR	AUC
MDT [7]	25.0%	81.8%	45.0%	44.1%
MPPCA [3]	40.0%	67.0%	18.0%	13.3%
SF-MPPCA [7]	32.0%	76.9%	28.0%	20.5%
SF [8]	31.0%	76.8%	21.0%	21.3%
Adam [1]	38.0%	64.9%	24.0%	19.7%
SRC [2]	19.0%	86.0%	46.0%	46.1%
Lu [5]	15.0%	91.8%	59.1%	63.8%
H-MDT [4]	17.8%	–	64.8%	66.2%
AMDN [11]	16.0%	92.1%	59.9%	67.2%
SDAE+LSTM	13.5%	92.7%	61.9%	70.3%
Ours	11.1%	93.2%	63.5%	71.7%

itive. By altering the detection threshold, we produce a frame-level ROC curve as shown in Fig. 5(a). The proposed algorithm has obvious high *true positive rate* (TPR) when the *false positive rate* (FPR) is low. This is vital for real applications.

As for the pixel-level measurement, a detected abnormal frame is true positive if and only if more than 40% abnormal pixels are detected. A normal frame is false positive as long as one pixel is detected as abnormal. In Fig. 5(b), we compare the pixel-level ROC curves. The proposed algorithm achieves the best performance.

AUC, EER, and EDR Based on these ROC curves, three evaluation criteria are taken as quantitative indexes. *Area under curve* (AUC) is the area under the ROC curve; *Equal error rate* (EER) is the ratio of misclassified frames when the false positive rate equals the miss rate, i.e. the FPR at which FPR $= 1 -$ TPR; *Equal detected rate* (EDR) is the detection rate at EER, i.e. EDR $= 1 -$ EER. In Table 1, we report EER and AUC at frame-level, meanwhile EDR and AUC at pixel-level. These results indicate that the proposed algorithm outperforms AMDN by 1.1% and 4.5% at AUC values of frame-level and pixel-level, respectively. Moreover, we also verify the impact of the manifold ranking algorithm by comparing with the results without ranking, which is denoted as "SDAE+LSTM" in Table 1.

3.2 Avenue Dataset

The Avenue dataset[2] contains 14 kinds of abnormal events, including running, loitering, and throwing objects, etc. There are 16 video clips for training and 21 clips for testing. All testing clips have object-level ground-truth, i.e. labeling anomalies with rectangle regions. This dataset has 30,652 frames totally, with spatial resolution of 360×640.

Some visualized comparison results are shown in Fig. 6. "3D Gradient" is the result of the proposed algorithm replacing the SDAE features with 3D gradient features. We can find that both "3D Gradient" and the proposed algorithm exclude false detections. In Table 2, we compare accuracies under different detection thresholds. According to [5], it is a correct detection if the ratio $\frac{\text{Detected Anomaly} \bigcap \text{True Anomaly}}{\text{Detected Anomaly} \bigcup \text{True Anomaly}}$ exceeds θ. As is shown in Table 2, deep representation (SDAE) is better than hand-crafted features (3D gradient).

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

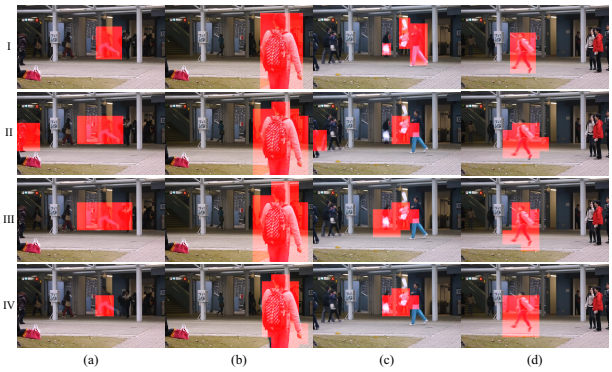

Figure 6: Examples of detected anomalies on the Avenue dataset. Abnormal regions are labeled with red masks. (I) the ground-truth; (II) Lu's algorithm [5]; (III) 3D Gradient; and (IV) the proposed algorithm.

Table 2: Comparisons on the Avenue dataset.

θ	Lu [5]	3D Gradient	SDAE+LSTM	Ours
0.2	70.0%	72.8%	73.7%	75.2%
0.3	67.3%	70.7%	71.9%	73.5%
0.4	63.3%	67.8%	68.4%	71.0%
0.5	59.3%	65.0%	67.6%	68.9%
0.6	57.5%	62.9%	65.7%	66.9%
0.7	55.7%	62.4%	63.8%	64.6%
0.8	54.4%	62.3%	63.2%	63.7%

Meanwhile, the manifold ranking algorithm improves the detection accuracies.

4. CONCLUSIONS

This paper presents a novel video abnormal event detection method based on unsupervised deep learning. Video event features are adaptively extracted and fused from appearance, texture, and short-term motion clues. In order to model long-term dependencies in video events, an LSTM based prediction model is proposed. Furthermore, a graph-based manifold ranking algorithm is designed to make use of the spatial contextual information. As a result, the discrimination between abnormal and normal events is increased, and false alarms are excluded. Experimental results on two surveillance video datasets demonstrate the effectiveness of the proposed algorithm, and show competitive performance compared with existing methods.

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 Nos. 61172143 and 61472413), the National Basic Research Program of China (973 Program) (Grant No. 2012CB719905), the Key Research Program of the Chinese Academy of Sciences (Grant No. KGZD-EW-T03), and the Open Research Fund of the Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences (Grant No. LSIT201408).

5. REFERENCES

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE TPAMI*, 30(3):555–560, 2008.

[2] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *IEEE CVPR*, pages 3449–3456, 2011.

[3] J. Kim and K. Grauman. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In *IEEE CVPR*, pages 2921–2928, 2009.

[4] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE TPAMI*, 36(1):18–32, 2014.

[5] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 FPS in MATLAB. In *IEEE ICCV*, pages 2720–2727, 2013.

[6] X. Lu, Y. Yuan, and P. Yan. Robust visual tracking with discriminative sparse learning. *Pattern Recognition*, 46(7):1762–1771, 2013.

[7] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *IEEE CVPR*, pages 1975–1981, 2010.

[8] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *IEEE CVPR*, pages 935–942, 2009.

[9] V. Reddy, C. Sanderson, and B. C. Lovell. Improved anomaly dtection in crowded scenes via cell-based analysis of foreground speed, size and texture. In *IEEE CVPR*, pages 55–61, 2011.

[10] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM Multimedia*, pages 461–470, 2015.

[11] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe. Learning deep representations of appearance and motion for anomalous event detection. In *BMVC*, pages 1–12, 2015.

[12] Y. Yuan, L. Mou, and X. Lu. Scene recognition by manifold regularized deep learning architecture. *IEEE TNNLS*, 26(10):2222–2233, 2015.

[13] B. Zhao, L. Fei-Fei, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In *IEEE CVPR*, pages 3313–3320, 2011.