Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder

Yong Shean Chong(✉) and Yong Haur Tay

Lee Kong Chian Faculty of Engineering Science,
Universiti Tunku Abdul Rahman, 43000 Kajang, Malaysia
yshean@1utar.my, tayyh@utar.edu.my

Abstract. We present an efficient method for detecting anomalies in videos. Recent applications of convolutional neural networks have shown promises of convolutional layers for object detection and recognition, especially in images. However, convolutional neural networks are supervised and require labels as learning signals. We propose a spatiotemporal architecture for anomaly detection in videos including crowded scenes. Our architecture includes two main components, one for spatial feature representation, and one for learning the temporal evolution of the spatial features. Experimental results on Avenue, Subway and UCSD benchmarks confirm that the detection accuracy of our method is comparable to state-of-the-art methods at a considerable speed of up to 140 fps.

Keywords: Anomaly detection · Feature learning · Regularity · Autoencoder

1 Introduction

With the rapid growth of video data, there is an increasing need not only for recognition of objects and their behaviour, but in particular for detecting the rare, interesting occurrences of unusual objects or suspicious behaviour in the large body of ordinary data. Finding such abnormalities in videos is crucial for applications ranging from automatic quality control to visual surveillance.

Meaningful events that are of interest in long video sequences, such as surveillance footage, often have an extremely low probability of occurring. As such, manually detecting such events, or anomalies, is a very meticulous job that often requires more manpower than is generally available. This has prompted the need for automated detection and segmentation of sequences of interest. However, present technology requires an enormous amount of configuration efforts on each video stream prior to the deployment of the video analysis process, even with that, those events are based on some predefined heuristics, which makes the detection model difficult to generalize to different surveillance scenes.

Recent effort on detecting anomalies by treating the task as a binary classification problem (normal and abnormal) [12] proved it being effective and accurate, but the practicality of such method is limited since footages of abnormal

F. Cong et al. (Eds.): ISNN 2017, Part II, LNCS 10262, pp. 189–196, 2017.
DOI: 10.1007/978-3-319-59081-3_23

events are difficult to obtain due to its rarity. Therefore, many researchers have turned to models that can be trained using little to no supervision, including spatiotemporal features [3,11], dictionary learning [10] and autoencoders [7]. Unlike supervised methods, these methods only require unlabelled video footages which contain little or no abnormal event, which are easy to obtain in real-world applications.

This paper presents a novel framework to represent video data by a set of general features, which are inferred automatically from a long video footage through a deep learning approach. Specifically, a deep neural network composed of a stack of convolutional autoencoders was used to process video frames in an unsupervised manner that captured spatial structures in the data, which, grouped together, compose the video representation. Then, this representation is fed into a stack of convolutional temporal autoencoders to learn the regular temporal patterns.

Our proposed method is domain free (i.e., not related to any specific task, no domain expert required), does not require any additional human effort, and can be easily applied to different scenes. To prove the effectiveness of the proposed method we apply the method to real-world datasets and show that our method consistently outperforms similar methods while maintaining a short running time.

2 Methodology

The method described here is based on the principle that when an abnormal event occurs, the most recent frames of video will be significantly different than the older frames. Inspired by [2], we train an end-to-end model that consists of a spatial feature extractor and a temporal encoder-decoder which together learns the temporal patterns of the input volume of frames. The model is trained with video volumes consists of only normal scenes, with the objective to minimize the reconstruction error between the input video volume and the output video volume reconstructed by the learned model. After the model is properly trained, normal video volume is expected to have low reconstruction error, whereas video volume consisting of abnormal scenes is expected to have high reconstruction error. By thresholding on the error produced by each testing input volumes, our system will be able to detect when an abnormal event occurs.

2.1 Feature Learning

We propose a convolutional spatiotemporal autoencoder to learn the regular patterns in the training videos. Our proposed architecture consists of two parts — spatial autoencoder for learning spatial structures of each video frame, and temporal encoder-decoder for learning temporal patterns of the encoded spatial structures. As illustrated in Fig. 1, the spatial encoder and decoder have two convolutional and deconvolutional layers respectively, while the temporal encoder is a three-layer convolutional long short term memory (LSTM) model.

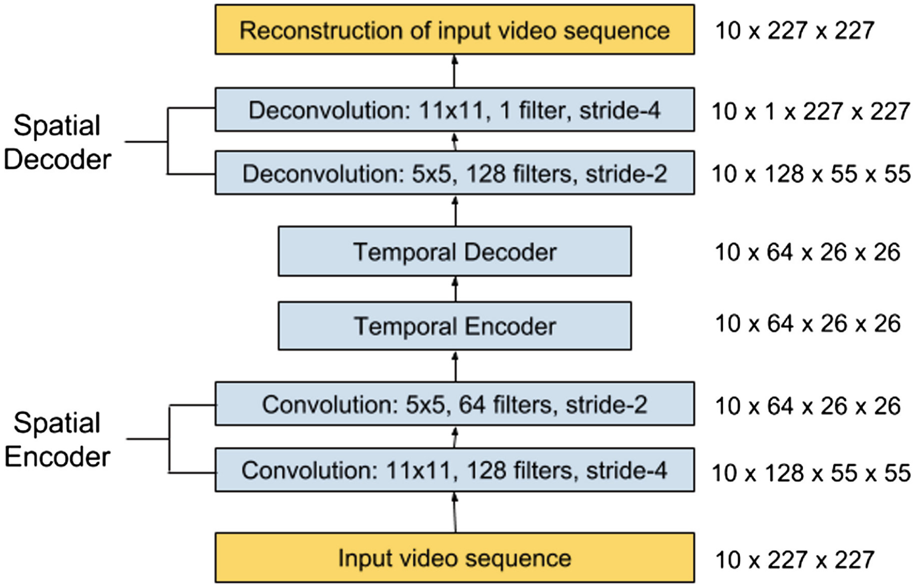

Fig. 1. Our proposed network architecture. It takes a sequence of length T as input, and output a reconstruction of the input sequence. The numbers at the rightmost denote the output size of each layer. The spatial encoder takes one frame at a time as input, after which $T = 10$ frames have been processed, the encoded features of 10 frames are concatenated and fed into temporal encoder for motion encoding. The decoders mirror the encoders to reconstruct the video volume.

Convolutional layers are well-known for its superb performance in object recognition, while LSTM model is widely used for sequence learning and time-series modelling and has proved its performance in applications such as speech translation and handwriting recognition.

Autoencoder. Autoencoders, as the name suggests, consist of two stages: encoding and decoding. It was first used to reduce dimensionality by setting the number of encoder output units less than the input. The model is usually trained using back-propagation in an unsupervised manner, by minimizing the reconstruction error of the decoding results from the original inputs. With the activation function chosen to be nonlinear, an autoencoder can extract more useful features than some common linear transformation methods such as PCA.

Spatial Convolution. The primary purpose of convolution in case of a convolutional network is to extract features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data. Suppose that we have some $n \times n$ square input layer which is followed by the convolutional layer. If we use an $m \times m$ filter W, the convolutional layer output will be of size $(n - m + 1) \times (n - m + 1)$.

Convolutional LSTM. A variant of the long short term memory (LSTM) architecture, namely Convolutional LSTM (ConvLSTM) model was introduced by Shi et al. in [8] and has been recently utilized by Patraucean et al. in [6] for video frame prediction. Compared to the usual fully connected LSTM (FC-LSTM), ConvLSTM has its matrix operations replaced with convolutions. By using convolution for both input-to-hidden and hidden-to-hidden connections, ConvLSTM requires fewer weights and yield better spatial feature maps. The formulation of the ConvLSTM unit can be summarized with (7) through (12).

$$f_t = \sigma(W_f * [h_{t-1}, x_t, C_{t-1}] + b_f) \tag{1}$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t, C_{t-1}] + b_i) \tag{2}$$

$$\hat{C}_t = tanh(W_C * [h_{t-1}, x_t] + b_C) \tag{3}$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \hat{C}_t \tag{4}$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t, C_{t-1}] + b_o) \tag{5}$$

$$h_t = o_t \otimes tanh(C_t) \tag{6}$$

In contrast to the FC-LSTM, the input is fed in as images, while the set of weights for every connection is replaced by convolutional filters (the symbol $*$ denotes a convolution operation). This allows ConvLSTM work better with images than the FC-LSTM due to its ability to propagate spatial characteristics temporally through each ConvLSTM state. Note that this convolutional variant also adds an optional 'peephole' connections to allow the unit to derive past information better.

2.2 Regularity Score

Once the model is trained, we can evaluate our models performance by feeding in testing data and check whether it is capable of detecting abnormal events while keeping false alarm rate low. To better compare with [2], we used the same formula to calculate the regularity score for all frames, the only difference being the learned model is of a different kind. The reconstruction error e of all pixel values in frame t of the video sequence is taken as the Euclidean distance between the input frame $x(t)$ and the reconstructed frame $f_W(x(t))$:

$$e(t) = ||x(t) - f_W(x(t))||_2 \tag{7}$$

where f_W is the learned weights by the spatiotemporal model. We then compute the abnormality score $s_a(t)$ by scaling between 0 and 1. Subsequently, regularity score $s_r(t)$ can be simply derived by subtracting abnormality score from 1:

$$s_a(t) = \frac{e(t) - e(t)_{min}}{e(t)_{max}} \tag{8}$$

$$s_r(t) = 1 - s_a(t) \tag{9}$$

3 Experiments

3.1 Datasets

We train our model on five most commonly used benchmarking datasets: Avenue [3], UCSD Ped1 and Ped2 [4], Subway entrance and exit datasets [1]. All videos are taken from a fixed position for each dataset. All training videos contain only normal events. Testing videos have both normal and abnormal events.

3.2 Results and Analysis

Quantitative Analysis: ROC and AUC. Table 1 shows the frame-level AUC and EER of our and of other methods on all five datasets. We outperform all other considered methods in respect to frame-level EER.

Table 1. Comparison of area under ROC curve (AUC) and Equal Error Rate (EER) of various methods. Higher AUC and lower EER are better. Most papers did not publish their AUC/EER for avenue, subway entrance and exit dataset.

Method	AUC/EER (%)				
	Ped1	Ped2	Avenue	Subway entrance	Subway exit
Adam [1]	77.1/38.0	−/42.0			
SF [5]	67.5/31.0	55.6/42.0			
MPPCA [4]	66.8/40.0	69.3/30.0		N/A	
MPPCA+SF [4]	74.2/32.0	61.3/36.0			
HOFME [9]	72.7/33.1	87.5/20.0	N/A	81.6/**22.8**	84.9/17.8
ConvAE [2]	81.0/27.9	**90.0**/21.7	70.2/25.1	**94.3**/26.0	80.7/9.9
Ours	**89.9/12.5**	87.4/**12.0**	**80.3/20.7**	84.7/23.7	**94.0/9.5**

We also present a run-time analysis on our proposed abnormal event detection system, on CPU (Intel Xeon E5-2620) and GPU (NVIDIA Maxwell Titan X) respectively, in Table 2. The total time taken per frame is well less than a quarter second per frame for both CPU and GPU configuration.

Qualitative Analysis: Visualising Frame Regularity. Figures 2, 3, and 4 illustrate the output of the proposed system on samples of the Avenue dataset, Subway entrance and exit scenes respectively; our method detects anomalies correctly in these cases even in crowded scenes.

From Fig. 5, it is easy to see that our method has detected more abnormal events with fewer false alarms compared to [2]. Also, as observed in Fig. 6, our method is able to produce higher regularity score during normal activities and lower scores when there are abnormalities.

Table 2. Details of run-time during testing (second/frame).

	Time (in sec)			
	Preprocessing	Representation	Classifying	Total
CPU	0.0010	0.2015	0.0002	0.2027 (∼5fps)
GPU	0.0010	0.0058	0.0002	0.0070 (∼143fps)

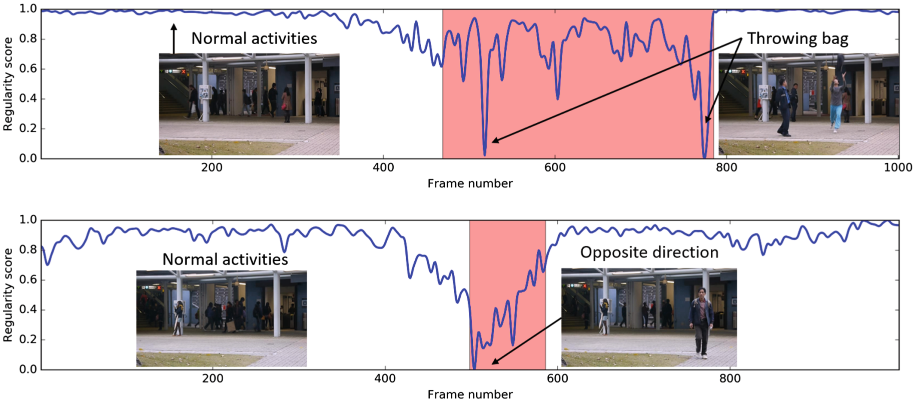

Fig. 2. Regularity score of video #5 (top) and #15 (bottom) from the Avenue dataset.

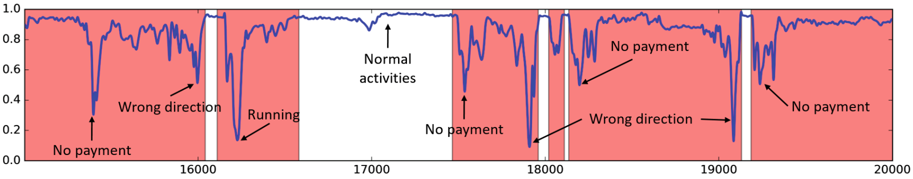

Fig. 3. Regularity score of frames 115000-120000 from the Subway Entrance video.

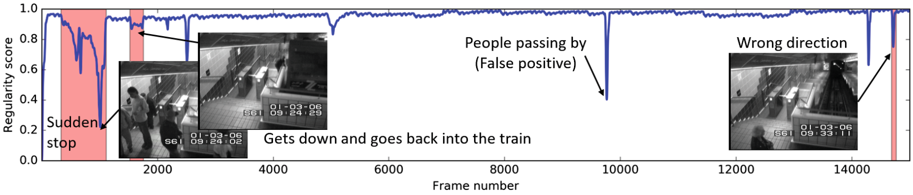

Fig. 4. Regularity score of frames 22500-37500 from the Subway Entrance video.

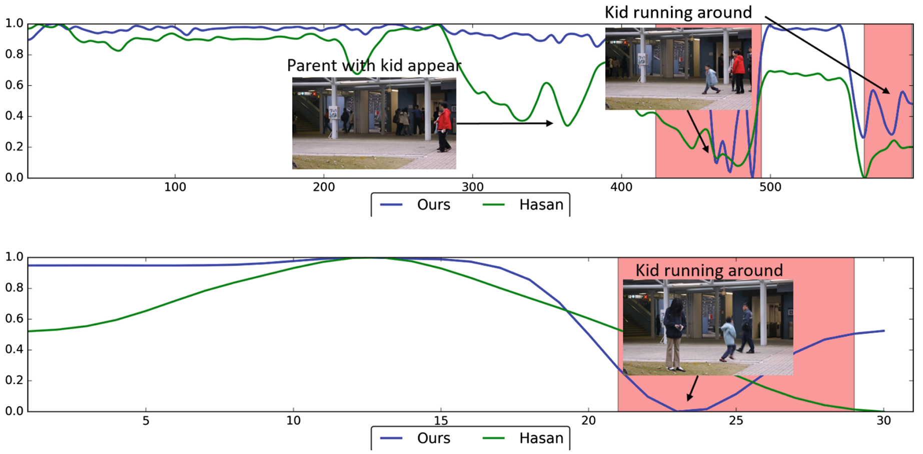

Fig. 5. Comparing our method with ConvAE [2] on Avenue dataset video #7 (top) and #8 (bottom). Best viewed in colour.

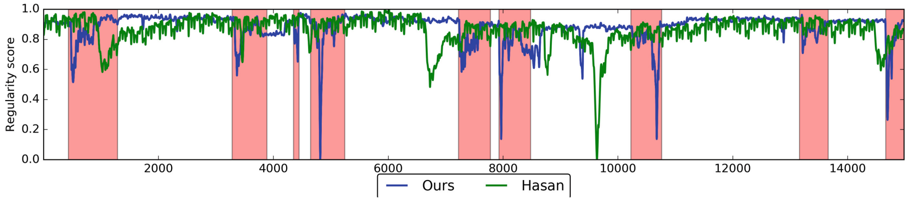

Fig. 6. Comparing our method with ConvAE [2] on Subway Entrance video frames 120000-144000. Best viewed in colour.

4 Conclusion

In this research, we have successfully applied deep learning to the challenging video anomaly detection problem. We formulate anomaly detection as a spatiotemporal sequence outlier detection problem and applied a combination of spatial feature extractor and temporal sequencer ConvLSTM to tackle the problem. The ConvLSTM layer not only preserves the advantages of FC-LSTM but is also suitable for spatiotemporal data due to its inherent convolutional structure. By incorporating convolutional feature extractor in both spatial and temporal space into the encoding-decoding structure, we build an end-to-end trainable model for video anomaly detection. The advantage of our model is that it is semi-supervised – the only ingredient required is a long video segment containing only normal events in a fixed view. Despite the models ability to detect abnormal events and its robustness to noise, depending on the activity complexity in the scene, it may produce more false alarms compared to other methods.

References

1. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans. Pattern Anal. Mach. Intell. **30**(3), 555–560 (2008)
2. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–742, June 2016
3. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: 2013 IEEE International Conference on Computer Vision, pp. 2720–2727, December 2013
4. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1975–1981 (2010)
5. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, pp. 935–942 (2009)
6. Patraucean, V., Handa, A., Cipolla, R.: Spatio-temporal video autoencoder with differentiable memory. In: International Conference on Learning Representations (2015), pp. 1–10 (2016). http://arxiv.org/abs/1511.06309
7. Sabokrou, M., Fathy, M., Hoseini, M., Klette, R.: Real-time anomaly detection and localization in crowded scenes. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 56–62, June 2015
8. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W., Woo, W.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 802–810. NIPS 2015. MIT Press, Cambridge, MA, USA (2015). http://dl.acm.org/citation.cfm?id=2969239.2969329
9. Wang, T., Snoussi, H.: Histograms of optical flow orientation for abnormal events detection. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS, pp. 45–52 (2013)
10. Yen, S.H., Wang, C.H.: Abnormal event detection using HOSF. In: 2013 International Conference on IT Convergence and Security, ICITCS 2013 (2013)
11. Zhao, B., Fei-Fei, L., Xing, E.P.: Online detection of unusual events in videos via dynamic sparse coding. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3313–3320 (2011)
12. Zhou, S., Shen, W., Zeng, D., Fang, M., Wei, Y., Zhang, Z.: Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Sig. Process. Image Commun. **47**, 358–368 (2016)