ISSN 2278-3091

Volume 9, No.5, September - October 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse241952020.pdf

https://doi.org/10.30534/ijatcse/2020/241952020

Bidirectional Convolutional LSTM Autoencoder for Risk Detection

[1], Khalid Housni[2]

[1] Ibn Tofail University, Faculty of Science, Kenitra,Morocco, iboulfrifi@gmail.com

[2] Ibn Tofail University, Faculty of Science, Kenitra,Morocco, Khalid.housni@uit.ac.ma

ABSTRACT

In this paper, we propose an unsupervised learning method for detecting risk in public space, using spatiotemporal auto-encoder combining convolutional neural networks to capture spatial features and Bidirectional LSTM to learn the evolution over time of the spatial features. This architecture learns normal features from training frames video of dataset and the events deviated from normal features are identified as risk. Experiments result show that the proposed method is more accurate and perform better than other state of art method such as HMM or bag of words.

Key words: Convolution neural network, LSTM, Auto-encoder, Risk detection

I. INTRODUCTION

The risk detection in video is an active area of activity recognition focuses on what is happening in the scene, and can be defined as the identification of abnormal patterns or abnormal motion in sequences video. Actually the most risk detection approaches focuses on unsupervised methods for many raison like difficulty to labeling data to situation with risk, the rarity to obtain a risk in sequences video, the challenging problem to define a situation with risk, and the constraint in detecting risk in real time.

In the past decade, a great work has already been done at this area of research, and can be classified to those categories: spatiotemporal descriptor based method [1], [2], [3], [4], statistical approaches [5], based trajectory method [6], [7], [8], [9], deep learning [10], [16], [17]or the combination of more than one of those categories.

In the last years, many risk detection approaches are based on spatiotemporal auto-encoder. This method of neural network is trained to produce an output same as input and it's trained in unsupervised manner for this raison it's suitable in real time application like risk detection in video.

Our approach present a model based on the convolutional neural network, Bidirectional LSTM and auto-encoder. The convolution neural network auto-encoder captures the local structure and the LSTM auto-encoder captures temporal information. This framework learn the normal characteristic from the normal training videos, and then the risks are detected as behavior deviated from the normal characteristic learned. This method captures the spatiotemporal correlation and performs consistently better than FC-LSTM and previous work.

2. RELATED WORK

Automated system to detect abnormal events has become an important domain of research due to his usage in real time application and the huge demand on video surveillance security.

Statistical framework approaches have modelled the motion variations in the form of HMM to describe the scene behaviour and identify the spatial and temporal relationships between motion patterns, and then the abnormal events are defined as statistical deviations in video frames [5].

Spatiotemporal descriptor based method are based on optical flow, its take features information from cuboids video and encoded those features.

Rensso and al. [11] propose a spatial temporal feature descriptor, called HOFM, this descriptor use optical flow information to characterize the normal behavior on the frames video, and applied nearest neighbor search to classify an a behavior as an normal event or not.

Deep learning based on artificial neural networks, have been applied to many research areas like computer vision, speech recognition, natural language processing…

Deep neural networks know a notable success in anomaly detection by responding to problematic not resolved by previous state of the art. And it learns the useful features directly from the input data and don't requires specific features to be extracted.

Zhou and al. [12], are used a spatial-temporal Convolutional Neural Networks (CNN) to extract spatial and temporal information, and extract appearance and motion information encoded in frames video. The spatial temporal convolutions are only applied on spatial temporal volumes of moving pixels to solve the local noise problem, and increase the quality detection

Hasan and al. [2], present an approach based on auto-encoders and The HOG and HOF motion features are used as input to the auto-encoder, the learned auto-encoder reconstructs regular motion in sequences video with low error, and using higher reconstruction error to identify abnormal events.

Ravanbakhsh and al. [13], proposed a method based on Generative Adversarial Nets (GANs) to learn representation of the normal pattern utilizing only normal training samples, and compare the real test-frame representations and the generated descriptions, if we get a difference the compared areas are identified as abnormal.

In Medel and al. [14], the network learns to predict normal evens similar to the training input videos, and using regularity scores derived from the reconstruction errors to identify the anomalies.

3. THE LSTL MODEL

The LSTM is a special RNN approach used for modeling long dependencies over past time, and resolve the problem of vanishing gradient and exploding problems in RNN. A stacked LSTMs can capture more accurate information and complex patterns.

The inner architecture of an LSTM unit is formulated as follow:

$$\boldsymbol{f_t} = \boldsymbol{\sigma}\left(\boldsymbol{W_f} \otimes [\boldsymbol{h_{t-1}}, \boldsymbol{x_t}] + \boldsymbol{b_f}\right) \quad (\mathbf{1})$$

$$\boldsymbol{i_t} = \boldsymbol{\sigma}(\boldsymbol{W_i} \otimes [\boldsymbol{h_{t-1}}, \boldsymbol{x_t}] + \boldsymbol{b_i}) \quad (\mathbf{2})$$

$$\hat{\mathrm{C}}_{\boldsymbol{t}} = \mathbf{tanh}(\boldsymbol{W_c} \otimes [\boldsymbol{h_{t-1}}, \boldsymbol{x_t}] + \boldsymbol{b_C}) \quad (\mathbf{3})$$

$$\boldsymbol{C_t} = \boldsymbol{f_t} \otimes \boldsymbol{C_{t-1}} + \boldsymbol{i_t} \otimes \hat{\mathrm{C}}_{\boldsymbol{t}} \quad (\mathbf{4})$$

$$\boldsymbol{o_t} = \boldsymbol{\sigma}(\boldsymbol{W_o} \otimes [\boldsymbol{h_{t-1}}, \boldsymbol{x_t}] + \boldsymbol{b_o}) \quad (\mathbf{5})$$

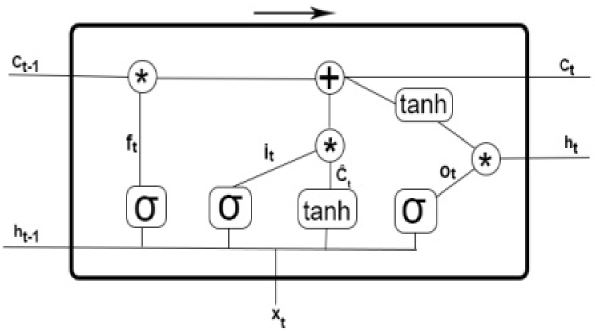
$$\boldsymbol{h_t} = \boldsymbol{o_t} \otimes \mathbf{tanh}(\boldsymbol{C_t}) \quad (\mathbf{6})$$

Equation (1) define the forget gate to reset the memory cell, Equations (2) and (3)denote the input and output gates, and essentially control the input and output of the memory cell.

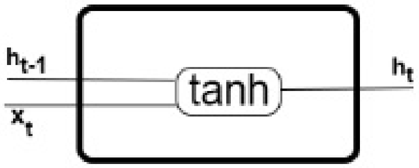
Equation (4) represents the memory cell that prevents the risk of vanishing gradient and exploding problems.

x_t denote the input at time t.

Figure 1 illustrate the difference between LSTM unit on the top and RNN on the bottom.



(a)



(b)

Figure 1: comparison between LSTM unit and RNN unit: (a) LSTM unit (b) RNN unit

Our approach use Bidirectional Long Short-Term Memory Networks to capture the most important semantic information and height level features in sequences video. In BLSTM the output of the forward and backward layer is combined at each time step to form one output, it's learn the past and future feature very fast and more accurate, the Figure 2 illustrate the BLSTM architecture:

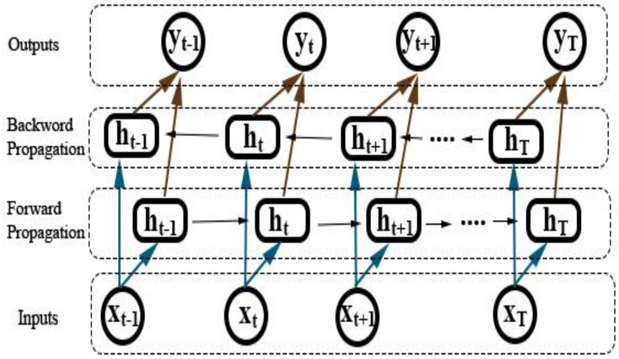


Figure2: BLSTM architecture

4. AUTOENCODER

An autoencoder is an unsupervised neural network formed by two steps: the first is the encoder that consists to reduce dimensionality of the input data, and the number of the output units layer is less than the output. The second is the decoder that consists to minimize the reconstruction error between the encoder result (hidden layer) and the original inputs. Figure 3 illustrates the autoencoder architecture.

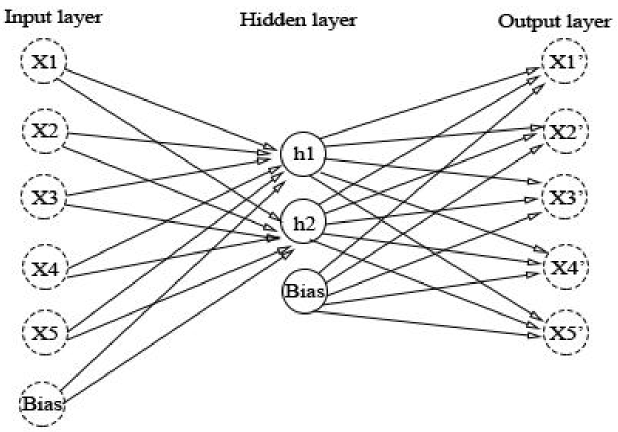


Figure 3: Autoencoder architecture

The autoencoder is more efficient in case ofnon-linear transformation compared to PCA.

5. THE PROPOSED METHOD

The proposed method focused on the reconstruction error by using an autoencoder composed by convolution layers and bidirectional LSTM. It's learns feature representation and modeling the temporal correlation. The autoencoder learns the normal behavior in video from the normal training videos, and then the risks are detected as behavior deviated from the normal characteristic learned.

Our model consists of two convolution layers followed by bidirectional LSTM layers, the Figure 4 illustrate the architecture of the proposed model.

The model is trained for 40 epochs with a batch size of 4, a dropout of 20 percent, Adam optimizer with lr=1e-4, decay=1e-5 and epsilon=1e-6.

We train our model on UCSD Ped1 dataset, the training videos contain videos without risk, and the testing videos contain both sequences video without risk and sequences video with risk.

Inspired by [9] the reconstruction error of the frame t is defined as follow:

$$e(t) = \|I(\ t) - f_w(I(t))\|_2 \qquad (7)$$

Where f_w the learned model by the proposed method, and the reconstruction error score is defined as follow:

$$RES = \frac{e(t) - e(t)_{min}}{e(t)_{max}} \qquad (8)$$

Frames with RES greater than a threshold α are classified as frames with risk. The Figure 5 illustrates the result of the proposed method on the dataset test UCSDped1.

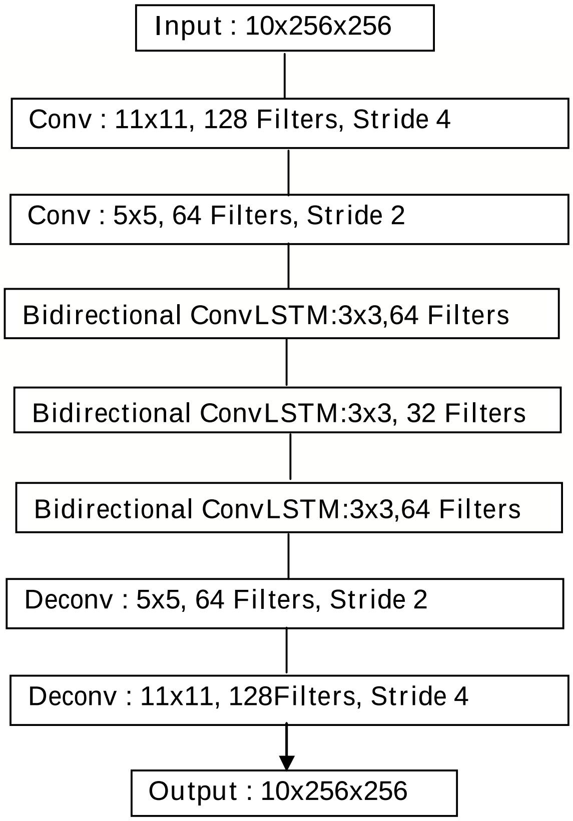

Input : 10x256x256
Conv : 11x11, 128 Filters, Stride 4
Conv : 5x5, 64 Filters, Stride 2
Bidirectional ConvLSTM:3x3,64 Filters
Bidirectional ConvLSTM:3x3, 32 Filters
Bidirectional ConvLSTM:3x3,64 Filters
Deconv : 5x5, 64 Filters, Stride 2
Deconv : 11x11, 128Filters, Stride 4
Output : 10x256x256

Figure 4: architecture of the proposed model

6. CONCLUSION

In this paper, we propose an unsupervised learning method for detecting risk in public space, using spatiotemporal auto-encoder combining convolutional neural networks to capture spatial features and Bidirectional LSTM to learn the evolution over time of the spatial features. The advantages of our method is than perform automatically and can be used in real time application. For future work we will focus on computing the threshold α automatically.

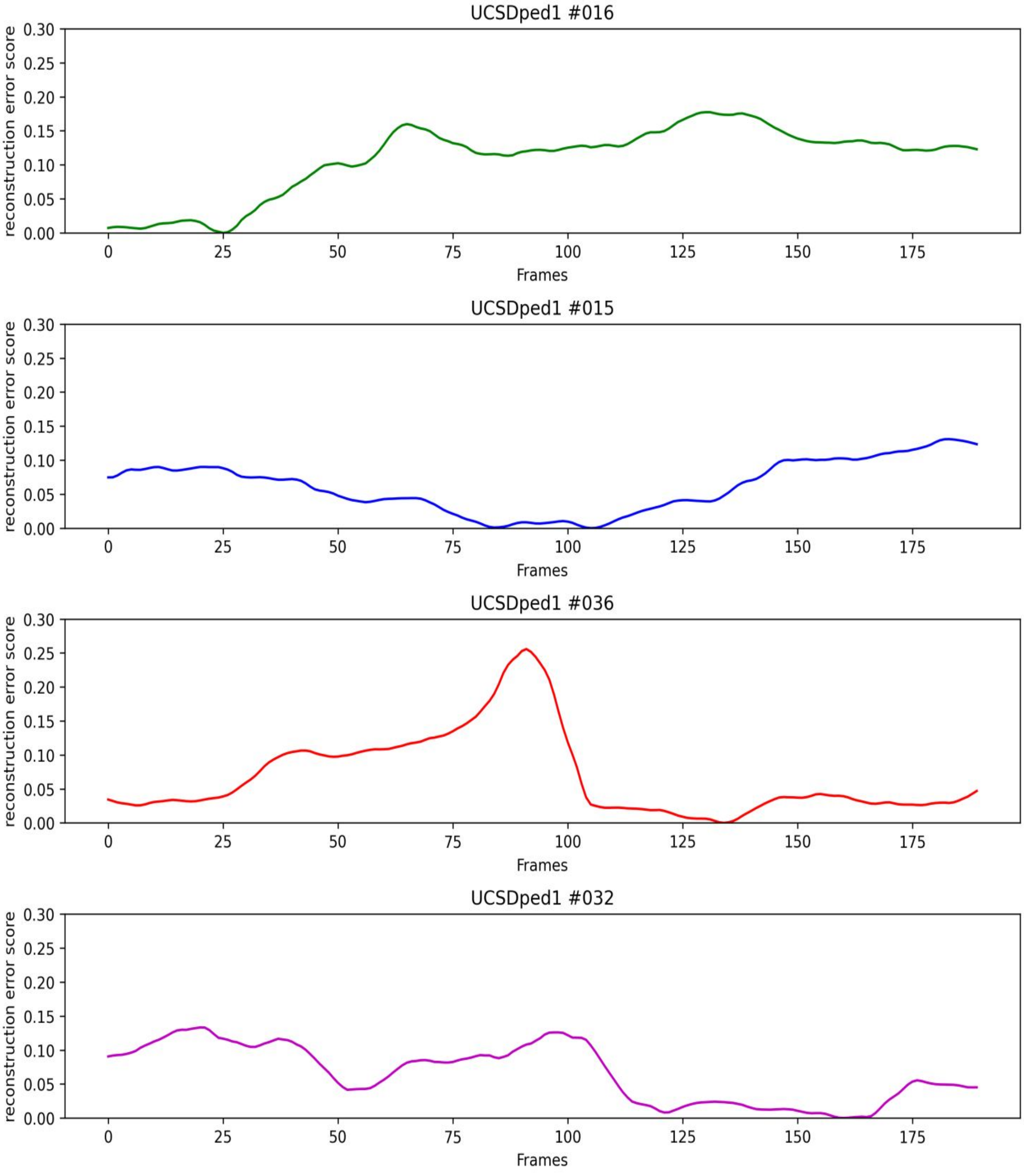


Figure 5: Result of the proposed method, frames with reconstruction error score greater than a threshold of 0.07 are classed as frames with risk.

REFERENCES

1. Y. Cong, J. Yuan, and Y. Tang, "**Video anomaly search in crowded scenes via spatio-temporal motion context**" *IEEE Trans. Inf. ForensicsSecur.*, vol. 8, pp. 1590–1599, 2013.
2. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, "**Learning temporal regularity in video sequences"**. *CVPR*. pp. 733–742 (June 2016)
3. M. J. Roshtkhari and M. D. Levine, "**Online dominant and anomalous behavior detection in videos**," *Proceedingsof the IEEE Computer Society Conference onComputer Vision and Pattern Recognition*, pp.2611–2618, 2013.
4. C. Lu, J. Shi, and J. Jia, "**Abnormal Event Detection at 150 FPS in MATLAB"***ICCV*, pp. 2720–2727, 2013.
5. L. KRATZ and K. NISHIMO, "**Tracking with local spatio-temporal motion patterns in extremely crowded scenes**," *International Conference on CVPR*, 2010.
6. X. Mo, V. Monga, R. Bala, and Z. Fan, "**Adaptive Sparse Representations for Video Anomaly Detection,**" *Circuits Syst. VideoTechnol. IEEE Trans.*, p. 1, 2013.
7. B. T. Morris and M. M. Trivedi, "**A survey of vision-based trajectory learning and analysis for surveillance,**" *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18. pp. 1114–1127, 2008.
8. Li, C., Han, Z., Ye, Q., Jiao, J."**Abnormal behavior detection via sparse reconstruction analysis of trajectory"**. *InternationalConference on Image and Graphics, ICIG*, pp. 807–810, 2011.
9. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, "**Learning temporal regularity in video sequences,"***CVPR*. pp. 733–742 , June 2016.
10. 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",***Signal Processing: Image Communication* 47, 358–368, sep 2016.
11. R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade and W. R. Schwartz"**Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos".***IEEE Transactions on Circuits and Systems for Video Technology*, pp. 673-682,2017
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".***Signal Processing: Image Communication* 47, 358–368,September 2016.
13. Ravanbakhsh, M.; Sangineto, E.; Nabi, M.; Sebe, N. "**Abnormal Event Detection in Videos using Generative Adversarial Nets"**. *IEEE International Conference on Image Processing (ICIP)*, 17–20, September 2017.
14. Medel, J.R.; Savakis, A. "**Anomaly detection in video using predictive convolutional long short-term memory networks"**, 2016.
15. Hochreiter, S., Schmidhuber, J."**Long short-term memory**",*Neural computation*, vol. 9(8), 1735-1780, 1997.
16. Mohamad Haider Abu Yazid, Mohamad Shukor Talib and Muhammad Haikal Satria, "**Heart Disease Classification Framework Using Fuzzy and Flower Pollination Neural Network".** *International Journal of Advanced Trends in Computer Science and Engineering*, vol 8(1.6), *135-139, 2019*.
17. Putu Doddy Heka Ardana, I Gusti Made Sudika, Ni Kadek Astariani and Gede Sumarda, **"Application of Feed Forward Back propagation Neural Network in Monthly Rainfall Prediction".** *International Journal of Advanced Trends in Computer Science and Engineering*, vol 8(1.5), 192-198, 2019.