ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 6, 145-150 (2021)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Video Risk Detection and Localization using Bidirectional LSTM Autoencoder and Faster R-CNN

Idir Boulfrifi*, Mohamed Lahraichi, Khalid Housni

Faculty of Science, Ibn Tofail University, Kenitra, 14000, Morocco

ARTICLE INFO

Article history:
Received: 09 November, 2021
Accepted: 05 December, 2021
Online: 30 December, 2021

Keywords:
Convolution neural network
Fast R-CNN
LSTM
Auto-encoder

ABSTRACT

This work proposes a new unsupervised learning approach to detect and locate the risks "abnormal event" in video scenes using Faster R-CNN and Bidirectional LSTM autoencoder. The approach proposed in this work is carried out in two steps: In the first step, we used a bidirectional LSTM autoencoder to detect the frames containing risks. In the second step, for each frame containing risks, we first used Faster R-CNN to extract all the objects containing in the scene and then for each object detected we check whether it represents a risk or not. In other words, in testing phase, the frames with events deviated from normal features learned in training phase are detected as risk. To locate objects representing risk, only the objects detected by Fast R-CNN deviated from normal feature are classified as risk. Experimental results demonstrate that the proposed method can reliably detect and locate the object representing risk in video sequences.

1. Introduction

The security of public spaces has become a very important area in recent years, hence the need to develop an automated surveillance system capable of analysing video scenes, exactly detect and locate anomalies. In order to respond to this, demand several approaches have been proposed based on sparse coding [1]-[3] or deep learning techniques [4]-[6]. The last one is divided on two types of learning techniques [1], the supervised setting which requires both normal and abnormal labelled training samples, but it is difficult to obtain a training labelled data set, moreover, unsupervised methods avoid excessive manual labelling, it can only be applied on the specific scenes because they are using underlying data and prior knowledge to design limited distributions.

In recent years, deep learning has become an important domain and have been applied for a diverse set of tasks, the anomaly detection is one, most research on this area was based on RNN and CNN network architectures, such as deep generative models such as variational autoencoder (VAE) [7], generative adversarial networks (GANs) [8], Long Short-Term memory networks (LSTMs) [9], deep learning have resolved the weakness of traditional method.

In real life there is some challenging problems to detect anomalous video, like the lack of clear definition of anomalies, difficulty of scene segmentation, high density object with random motions, occlusions, and the fact that a risk appears rarely in short time. This paper focused on unsupervised risk detection and localization approaches based on convolutional neural network algorithms.

Our approach present a model based on the convolutional neural network, Bidirectional LSTM auto-encoder and Fast C-RNN. The convolution neural network auto-encoder captures the local structure and the LSTM auto-encoder captures temporal information, it learns the normal patterns from the normal training videos then the risks are detected as events deviated from the normal patterns learned. To detect risk in frame we compute de difference between the original frame and the reconstructed frame, and performing a threshold error will detect the frame contains risk. In the second steep the Fast C-RNN extract objects from the frame with risk, to localize risk in this frame the error is only computed between objects detected in normal frame and those objects in the reconstructed frame. Fig. 1 illustrates an overview of the proposed method.

The reminder of this paper is organized as follows: The second section presents the methodology, the third section shows experimental results, and finally a conclusion.

*Corresponding Author: Idir Boulfrifi, iboulfrifi@gmail.com

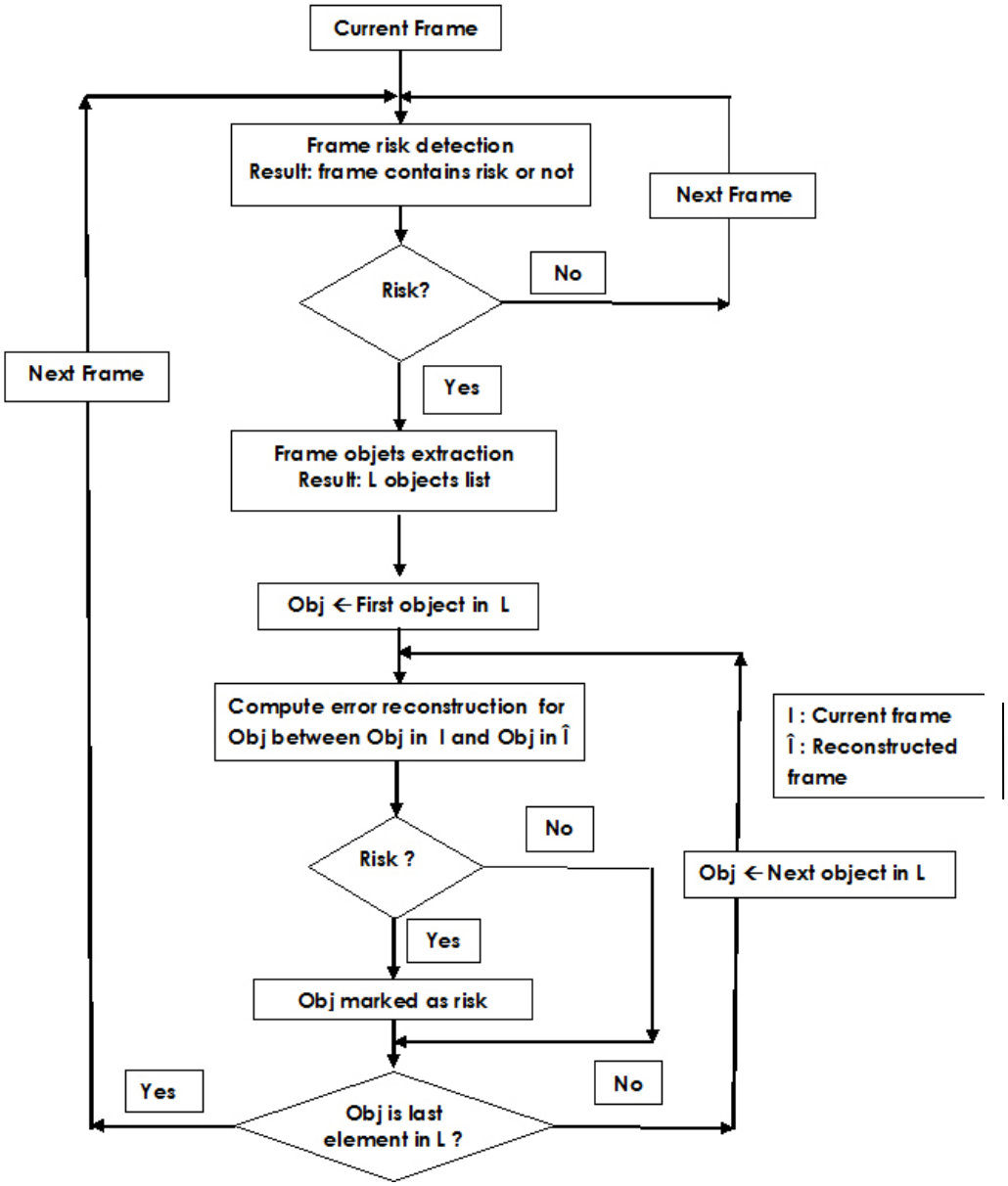


Figure 1: Overview of the proposed approach

2. Related Work

In this section, we present some of the previous works on anomaly detection based on deep learning. In [4], the authors have been introduced a cascade of auto-encoders, based on two novel cubic-patch-based anomaly detectors, the first one based on power of an auto-encoder on reconstructing an input video patch and the second is based on the power of sparse representation of an input video patch.

To analyze spatial and temporal information, the spatial temporal Convolutional Neural Networks have been used to capture spatial and temporal features encoded in frames video. The convolution is only performed in spatial temporal volumes of moving pixels to resolve the problem of local noise, and increase detection accuracy [10]. In [11], the authors have integrated the HOG and HOF motion features as input to the autoencoder to learn the reconstruction of regular motion in video frames, in the reconstruction step, the higher error is classified as abnormal events. Hence in [5], the authors have proposed an unsupervised deep learning framework for anomalous event detection in complex video scenes based on a three-stream architecture (spatial, temporal and their joint representation) by employing the auto-encoder to learn the features, furthermore in [6], the authors built a novel model called spatio-temporal autoencoder (STAE), which learn video representation automatically using deep neural networks and extract features from both spatial and temporal dimensions by using 3-dimensional convolutions.

Whereas Generative Adversarial Nets (GANs) [12] takes as the first input the normal frames and produces corresponding optical-flow images. As the second input, GANs takes the real optical-flow of normal frames and outputs an appearance reconstruction. The abnormal areas are detected by computing local differences between the reconstructed appearance and motion and the normal frames and real optical-flow in order.

In [13], the authors have been presented a composite Conv-LSTM network able to reconstruct input frames, and predict future frames. The anomalous video segments are detected using a regularity evaluation algorithm at the model's output, video sequences containing normal events have a higher regularity score since they are similar to the data used to train the model, while sequences containing abnormal events have a lower regularity score.

3. Methodology

3.1. The LSTM Model

To resolve the problem of vanishing gradient and exploding in recurrent neural network (RNN), Hochreiter and al in [14], have been presented a neural network called Long Short-Term Memory (LSTM) for modelling long dependencies over time, and learn more semantic information and complex features. The LSTM unit is formulates with the equations as follow:

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

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

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

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

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

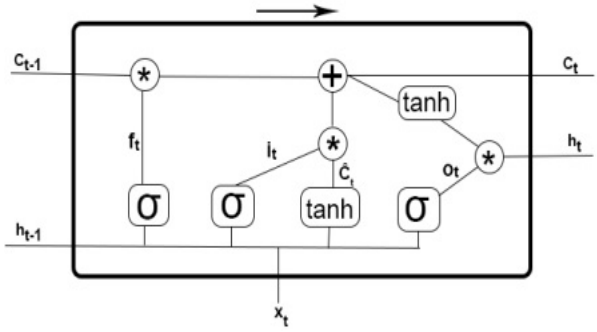
$$\boldsymbol{h_t} = \boldsymbol{o_t} \otimes \boldsymbol{tanh}(\boldsymbol{C_t}) \quad (6)$$

Eq. (1) define the forget gate to reset the memory cell, Eq. (2) and Eq.(3) denotes the input and output gates, and essentially control the input and output of the memory cell. Eq. (4) represents the memory cell that prevents the problem of vanishing gradient and exploding problem in RNN.

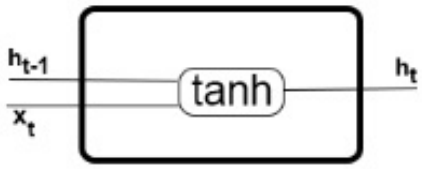
x_t denote the input at time t.

Figure 2 illustrates the difference between LSTM unit and RNN neural network.

To capture the most important semantic information and height level features in sequences video, we use Bidirectional Long Short-Term Memory (BLSTM) Networks architecture. The basic idea of BLSTM that the output of the forward and backward layer is combined at each time step to form one output, it's learned the past and future feature very fast and more accurate, the Fig. 3 illustrate the BLSTM architecture:



(a)



(b)

Figure 2: LSTM unit on the top and RNN unit on the bottom

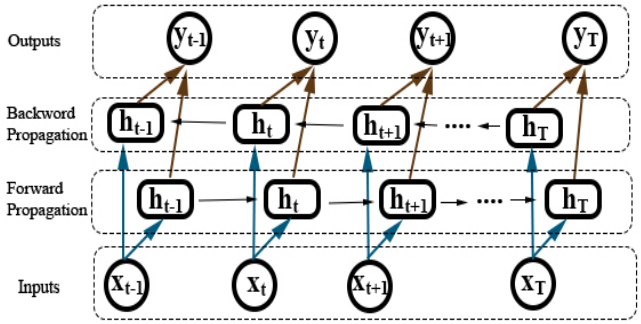


Figure 3: The architecture of BLSTM

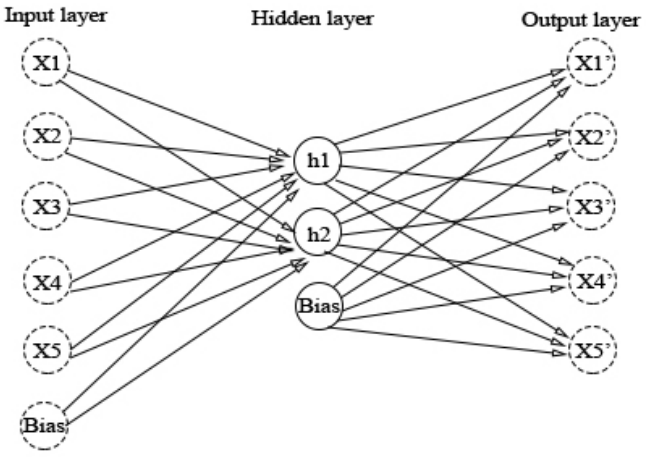


Figure 4: Autoencoder architecture

3.2. The Autoencoder

An autoencoder is an artificial neural network which performed in an unsupervised learning context, it can be seen as

the set of two components, an encoder and a decoder, the encoder that consists to reduce dimensionality of the input data, in order to represent them in a new space, the number of their input units layer is less than the output. Whereas the decoder, reconstruct data from the encoding by minimizing the reconstruction error between the encoder result (hidden layer) and the original inputs. The Fig.4 shows the autoencoder architecture.

In comparison with the PCA, the autoencoder is more efficient in case of non-linear transformation.

3.3. Faster C-RNN

Object detection is the process of finding objects in frames video, there is many approaches treating this area of research, on this paragraph we will focused on the neural network Faster R-CNN approach [14].

Faster R-CNN is a region based neural network, the first steep use RPN fully convolutional network that take an image as input and outputs proposal regions with an objectness score for each one the second steep integrate the Fast R-CNN network to classify those regions.

The RPN network is a complete convolutional network which slides on the feature map to indicate for each position whether there is an object or not, without taking into account the class of the object.

In order to have a system that is robust to translation and to scale, RPN uses an anchor-based algorithm. For each position of the sliding window on the feature map, 9 anchors are placed. The anchors are all centered on the sliding window, only their scale and ratio change (there are three scales and three ratios (1: 1, 2: 1 and 1: 2), which makes the 9 anchors. Each anchor is processed through the convolutional layers of the RPN and the networks produce the probability that this anchor represents an object and potentially an offset to correct the dimensions of the anchor. Faster R-CNN generate region proposals directly in the network instead of using an external algorithm, that's make it faster, accurate and useful in real time detection.

4. The Proposed Method

For risk detection and localization in space public, we have presented an approach, which consists of recurrent neural network and convolution neural network. The proposed method focused on two pre-trained models. The Bidirectional LSTM Autoencoder [15], this model learns the normal behavior from normal training video frames. The risks on testing data are detected as behavior deviated from the normal characteristic learned, by performing a threshold on reconstruction error. The second model is the Fast RCNN [14], a pre-trained object detection model which used to detect and extract objects in video frames, the model was trained on the MS-COCO dataset, and the output of this model is the bounding box of objects detected in frames video.

The first model consists of two convolution layers followed by bidirectional LSTM layers in the Fig. 5. Inspired by [17], the reconstruction error of the frame t is defined as follow:

$$\boldsymbol{e(t)} = \|\boldsymbol{I}(\boldsymbol{t}) - \boldsymbol{f_w}(\boldsymbol{I(t)})\|_2 \quad (7)$$

where $\boldsymbol{f_w}$ the learned model and the reconstruction error score is defined as follow:

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

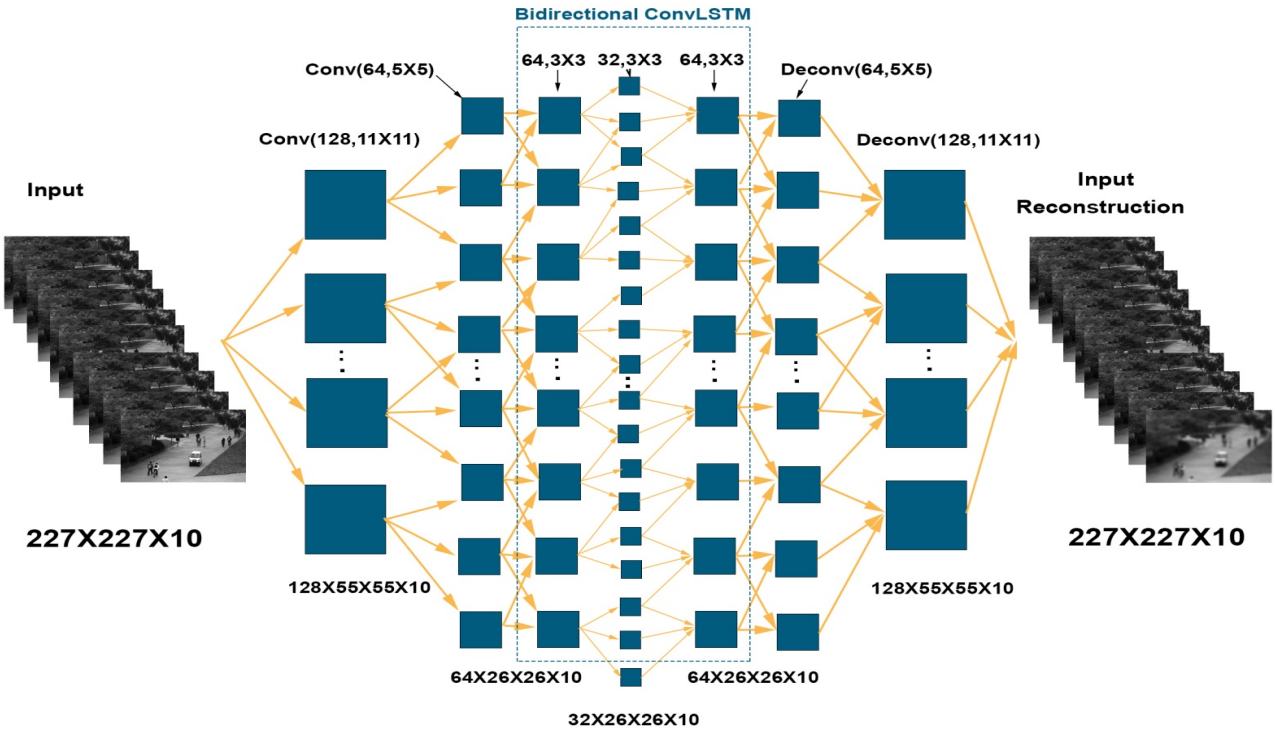


Figure 5: Architecture of the first model

To locate objects representing risk, we apply equation (8) to compute error construction of objects detected in frames with risk. In other words, the reconstruction error scores between bounding box of object in the original frame and the same bounding box in reconstructed frame must be greater than a threshold α to classify the object as abnormal, the experimental value of α is 0.005.

5. Experimental Results

The model was trained and tested on the UCSD Ped1 datasets, the training video frames are divided into temporal cuboid of 10 frames, the resolution of each frame is 227x227, and the pixels value are normalized to take value between 0 and 1. The same processing data was performed to the testing video frames and the reconstructed error was computed for every cuboid of testing data.

The training videos contain videos without risk, and the testing videos contain both sequences video without risk and sequences video with risk. The computer used in this works has NVIDIA K80 with 12GB Memory tensorflow 2 python Library. To trains the model 40 epochs was used with a batch size of 4, a dropout of 20 percent, Adam optimizer with lr=1e-4, decay=1e-5 and epsilon=1e-6.

The Fig 6, 7 and 8 illustrates the result of our approach. (a) Represents the image difference between the original and the reconstructed frame, (b) shows objects of original frame, and the reconstructed objects, the object which has a reconstruction error greater than a threshold (0.005) is classified as abnormal object.

(a) Input frame on the left, reconstructed frame on the middle, and the difference between input and reconstructed frame on the right

(b) Extracted objects on the top, reconstructed frame on the bottom
Figure 6: The folder test 19 in UCSDped1, frame 150

(a) Input frame on the left, reconstructed frame on the middle, and the difference between input and reconstructed frame on the right

(b) extracted objects on the top, reconstructed frame on the bottom
Figure 7: The folder test 36 in UCSDped1, frame 120

(a) Input frame on the left, reconstructed frame on the middle, and the difference between input and reconstructed frame on the right

(b) Extracted objects on the top, reconstructed frame on the bottom
Figure 8: The folder test 19 in UCSDped1, frame 120

The Tab. 1 demonstrates the reconstructed error of detected objects. The object 5(car), where error greater than the threshold (0.005) in frame 2 was located as risk.

Table 1: Reconstruction Error for Detected Objects

	Obj1	Obj2	Obj3	Obj4	Obj5
Frame1	0.0037	0.004	0.0028	0.0032	0.0033
Frame 2	0.0043	0.003	0.0038	0.0041	0.0054
Frame 3	0.0035	0.0029	0.0028		

6. Conclusion

This work proposes an automated deep learning-based approach to detect and locate risks in public space; it exploits both convolution neural network and recurrent neural network in order to learn the spatial and temporal semantic features from frames video. Firstly, Bidirectional LSTM Autoencoder is used to detect abnormal frames, and then, a Fast R-CNN is applied to locate exactly objects representing risks in frames containing risks. Experiments were tested on UCSD datasets, and prove that the proposed approach detect and locate risks accurately. This approach performs automatically without any involvement, which makes it useful for real time video surveillance.

Finally, our future work consists to treat this idea in one convolutional neural network model.

References

[1] M. Sabokrou, M. Fathy, M. Hoseini, "Video anomaly detection and localization based on the sparsity and reconstruction error of auto-encoder," Electronic Letter, **52**, 1122 – 1124, 2016, doi: 10.1049/el.2016.0440.

[2] X. Mo, V. Monga, R. Bala, and Z. Fan, "Adaptive Sparse Representations for Video Anomaly Detection," Circuits Syst. Video Technol. IEEE Trans, **1**, 2013, doi**:** 10.1109/TCSVT.2013.2280061.

[3] C. Li, Z. Han, Q. Ye and J. Jiao, "Abnormal behavior detection via sparse reconstruction analysis of trajectory," Proceedings-6th International Conference on Image and Graphics, ICIG, 807–810, 2011, doi: 10.1109/ICIG.2011.104

[4] Hu, Jingtao, E. Zhu, S. Wang, X. Liu, X. Guo, J. Yin, "An Efficient and Robust Unsupervised Anomaly Detection Method Using Ensemble Random Projection in Surveillance Videos," Sensors 19, **19**, 4145, 2019, doi:10.3390/s19194145.

[5] D. Xu, E. Ricci, Y. Yan, J. Song, N. Sebe. "Learning deep representations of appearance and motion for anomalous event detection," BMVC, 1-12, 2015,doi: 10.5244/c.29.8.

[6] Y. Zhao, Y. Deng, B. Shen, C. Liu, Y. Lu, H. Hua, X.S. "Spatio-temporal autoencoder for video anomaly detection,**"** Proceedings of the 25th ACM International Conference on Multimedia, Silicon Valley, CA, USA, 23–27, 1933–1941, 2017, doi: 10.1145/3123266.3123451.

[7] M. Welling, D. Kingma, "Stochastic gradient vb and the variational auto-encoder," Proceedings of the 2nd International Conference on Learning Representations (ICLR), 2014.

[8] J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing system, 2672–2680, 2014, DOI:10.1145/3422622.

[9] S. Hochreiter, S., Schmidhuber, J. "Long short-term memory," Neural computation, **9**(8), 1735-1780, 1997, doi:10.1162/neco.1997.9.8.1735.

[10] M. Sabokrou, M. Fayyaz, M. Fathy , Z. Moayed, R. Klette, "Spatial temporal convolutional neural networks for anomaly detection and localization in crowded scenes," Signal Processing: Image Communication, **47**, 358–368, September 2016, doi:10.1016/j.image.2016.06.007.

[11] M. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L. S. "Learning temporal regularity in video sequences," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),733–742, (June 2016), doi:10.1109/CVPR.2016.86.

[12] M. Ravanbakhsh, E. Sangineto, M. Nabi, N. Sebe, "Abnormal Event Detection in Videos using Generative Adversarial Nets," In Proceedings of the IEEE International Conference on Image Processing (ICIP),17–20, September 2017, doi:10.1109/APSIPAASC47483.2019.9023261.

[13] J. Medel, A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," 2016.

[14] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: To- wards realtime object detection with region proposal networks," Advances in Neural Information Processing Systems, 91–99, 2015, doi:10.1109/TPAMI.2016.2577031.

[15] K. Boulfrifi, K. Housni, "Bidirectional Convolutional LSTM Autoencoder for Risk Detection," International Journal of Advanced Trends in Computer Science and Engineering, **9**, 85-89, 2020, doi:10.30534/ijatcse/2020/241952020.

[16] 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, **18**, 1114–1127, 2008, doi:10.1109/TCSVT.2008.927109.

[17] M. Hasan, J. Choi, J. Neumann, K. Roy-Chowdhury, S. Davis "Learning temporal regularity in video sequences," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),733–742 , June 2016, doi:10.1109/CVPR.2016.86.