ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 15, Issue, 02, pp.4574-4579, February, 2024

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

DETECTING ANOMALIES IN VIDEOS USING SPATIOTEMPORAL AUTO ENCODER

Kusuma. S[1] and Kiran P[2]

[1]Department of Information Science & Engineering, Assistant Professor, MSRIT, Bengaluru, India
[1]Department of Computer Science & Engineering, Research Scholar, RNSIT, Affiliated to Visvesvaraya Technological University, Belagavi, Bengaluru, India
[2]Department of Computer Science & Engineering, Professor & Head, RNSIT, Bengaluru, India

DOI: http://dx.doi.org/10.24327/ijrsr.20241502.0858

ARTICLE INFO

Article History:

Received 14th January, 2023
Received in revised form 25rd January, 2023
Accepted 17th February, 2024
Published online 28th February, 2023

Keywords:

Anomaly Detection, Convolutional Neural Network, Long Short Term Memory, Auto encoder.

ABSTRACT

Suspicious event detection is one of the most focused areas of task in video analysis, which is the result of differentiating any event as abnormal or normal in the surveillance videos. As the differences between normal and abnormal events are uncertain, more discriminating methods or motion information need to be explored. There are three main classes of techniques to solve this problem unsupervised, supervised and semi-supervised. Recent work of applications in convolutional neural networks have shown significantly reliable results in identifying multiple types of objects present in the scene, which is given as an input in the form of an image with the help of convolutional layers. The downside of this is, it is a supervised learning model. An efficient method for detecting abnormalities in videos is proposed. It is semi supervised, progressing from existing supervised techniques. We aim to develop a spatiotemporal architecture consisting of two components, one for learning spatial information and the other for learning temporal information obtained from phylogeny of spatial features. The architecture proposed requires only normal event videos during training. The proposed architecture uses reconstruction error to make predictions.

INTRODUCTION

Over the past few decades, with speedy evolution of video data, there is pauperism not only for identification of objects and their behaviour, but also manually identifying anomalies which is a tedious task to perform and demands more manpower is required than is generally acquirable. The focus in particular is for detecting the abnormal behaviour in the vast amounts of ordinary data provided as input data, which is crucial for almost all kinds of applications in this field. Furthermore, whether a particular event could be considered anomalous or not is totally dependent on the context. For example, running in a park is considered normal whereas the same task performed in the classroom might be considered an anomaly. These extreme difficulties put any model under a challenge. In recent years the evolution of digitalization has led to increase in video surveillance as of forbes India article published on Aug 25th 2021. Security information and event management systems are only as effective as what they instrument. Abnormal event detection is an optimized and a fast technique to analyses a video, that parts between normal and abnormal events in a surveillance videos. India's capital city New Delhi ranks first with 1826.6 cameras per square mile, while Chennai has 609.9 cameras per square miles ranks third, London ranks second (1138.5 cameras) and Mumbai ranks eighteenth (157.4 cameras), thereby increasing the total video data to be constantly monitoring. In order to maintain citizen security, avoid the number of crimes happening on daily basis, the video surveillance has become one of the mandatory requirements in recent years as manually detecting anomalies is a meticulous job. With digitization everything is automated now and is slightly easier now to manage vast amounts of data which is applicable to abnormal event detection field as well. Abnormal event detection is crucial to detect any eccentric behaviour in an environment.

The framework consists of video data which has a set of general features. The deep learning approach is used to infer the data automatically from a long video footage. A deep neural network that comprises of a stack of convolutional auto encoders was used to process video frames in an unsupervised manner that capture special structures in the data that group together composed into video representation. This further is fed into a stack of convolutional temporal auto encoders to learn the regular temporal patterns. our method using convolution temporal auto encoder doesn't have restriction on domain either it requires any human efforts. The empirical results on public datasets depicts that the enactment of our implemented model is on par with current technologies.

Corresponding author:* **Kusuma. S
M S Ramaiah Institute of Technology, Research Scholar of Dept. of CSE, RNS Institute of Technology, Bangalore

The methodology used is based on the concept that when an abnormality occurs, the latest frames will be contradicting with the older frames. We shall train a model which can learn the features from the spatial and temporal encoder-decoder. The proposed model training will consist only of normal scenes, with the aim of deriving the value of reconstruction error between the given input and the generated output. By setting the threshold value for the sequence of frames generated as an output, our model identifies anomaly. There is no pre- set limit on what could be considered as an anomaly. Our approach consists of three stages which is further discussed in methodology

LITERATURE SURVEY

Abnormal events are very uncertain. Most of the abnormal events are not known priory making any model difficult to learn what exactly an abnormal event would be. Thus, the best approach is to train model only on normal events which are easily available.

Trajectories have been a well-known approach in the area of anomaly detection. This approach scrutinizes the normal events and gets familiar with the regular patterns and compares them with the test trajectory and finds the deviations. This technique is based on tracking and hence is not a reliable technique in crowded scenarios.

Several non-tracking approaches emerged after looking at the disadvantages of the tracking based approaches. These non-tracking approaches rely on drawing and examining the local features such as histograms, spatio-temporal features, optical flow. These are then clustered for finding abnormal events. Sparse reconstruction is another technique which rep- resents any event as a linear combination of feature representation in a trained dictionary.

With the emergence of deep learning models and their significant results in many areas, it has shown much improvement of abnormal event detection too. Deep learning models have customized hidden layers possessing the capabilities of learning the features on their own unlike previous methods. Convent has stack of convolutional layers which learns the features on its own and improves itself with more number of training data. It can efficiently determine the difference between the normal and abnormal events.

Convolutional neural network consists of stack of convolutional layers with a fully connected layer and a softmax classifier and aa convolutional auto encoder is essentially a convolutional neural network with its fully connected layer and classifier replaced by a mirrored stack of convolutional layer. Many models implemented with the stack of convolutional layers have shown the best use of learnt representation.

LSTM is another major unit which has shown its best performance in learning temporal features and works significantly well for time-series data. Videos are another type of time-series data. Thus, proving its worth presence in the proposed architecture. Though all the deep learning models have shown their capabilities, applying them in real time would be complex since samples related to abnormal events are very rare. With 2D convolution temporal information would be crashed making model very difficult to learn. Moreover, LSTM layers are memory intensive leading to high train and test times.

Table 1 Earlier contribution of researchers on abnormal event detection in videos

Title of the paper	Techniques used	Limitations	Dataset Used
Anomaly detection using convolutional spatiotempo- ralautoen- coder.[7]	Convolutional spatiotem- poral autoencoder.	High complexity.	Avenue, UCSD.
Spatiotemporal anomaly usingdeep learning for real-time video surveil- lance.[8]	Incremental Spatial temporal learner.	Requires hu- man feedback	UCSD, Avenue.
Real-world Anomaly Detection in Surveillance Videos.[13]	Multiple instance learning.	Long training time .	UCF Crime.
Two streamed fully convolutional networks.[9]	TS-FCN.	Data can be represented more clearly.	UCSD.
Learning spatiotempo- ral features using 3D convolutional networks.[16]	3D convolutional net- work(C3D).	-	UCSD, YUPENN, Avenue, Maryland.
Video classification using spatial- temporal features and PCA.[18]	The Gaussian Mixture Model.	Model complexity.	Commercial dataset.
Generative Neural Networks for Anomaly Detection in Crowded Scenesx.[14]	S2 - VAE.	Complexity.	UCSD, Avenue, PETS.
Two-stage Unsupervised Video Anomaly Detection usingLow-rank based Unsupervised One-class Learning with RidgeRe- gression.[10]	LR-UOCL-RR.	Not fit for feature reduction.	UCSD
Trajectory- Based Anomalous Event Detec- tion.[17]	Single-class (SVM) clustering.	Supervised learning approach.	Synthetic Data, Real-World Data.
Anomaly Detection in Videos Using Optical Flow and Convolutional Autoen- coder.[5]	ConvAE-LSTM.	Long training time.	Avenue, UCSD.
A Com-prehensive Survey ofVideo Datasets for Background Subtrac- tion.[12]	Gaussian models and fuzzy classification rules.	More compu- tation time.	UCSD, Avenue, PETS, UMN.
Anomaly detectionin video se- quences.[15]	Local and global autoencoders.	Complexity.	LAD.
Abnormal Crowd Behaviour Detection Using Motion Information, Images And CNN.[3]	CNN.	Needs both normal and abnormal events in train data.	Avenue, UCSD.
Unsupervised Anomaly Detection and Localization BasedOn Deep Spa- tiotemporal Translation Network.[4]	Generative AdversialNet- work (GAN) and Edge Wrapping (EW).	Considers un- known events as anomalous.	UCSD, Avenue.
Artificial Intelligence Image Recognition	RCNN.	Difficulty i n processing	PASCAL VOC.

Title of the paper	Techniques used	Limitations	Dataset Used
Method Based on CNN Algorithm.[2]		longer sequences.	
Spatiotemporal Unity Networking For Video Anomaly Detection.[6]	ConvLSTM.	Complexity.	Avenue, UCSD.
Abnormal Event Detection in Videos Based on Deep Neural Networks.[1]	Gaussian dis- tribution.	Partially supervised learning method.	UCSD, Avenue.
A Survey of Deep Learning- based Object Detec- tion.[11]	RCNN.	Difficulty in processing longer sequences.	Difficulty in processing longer sequences.

METHODOLOGY

A.Video to Image Frames

In the proposed methodology, as shown in Fig. 1, the given video input is converted into series of image frames. Videos can be given .mp4 or .avi formats. Other formats can also be provided.

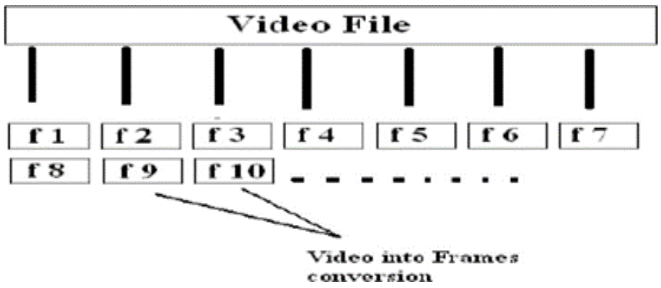

Fig.1 Video to image frames conversion

For analysing of a raw video it's to image frames are done first then those frames are converted to grey scale images for dimensionality reduction.

B. Pre-processing

All the frames drawn from the input video are resized to 227x227. Then the extracted frames are converted to grey scale for dimension reduction. The input images should not suffer from scaling issue for it to be consumed by the model properly. To overcome this, normalization is done. A batch of 10 continuous frames is provided as input to the model. Strides 1, 2, 3 are supplemented to increase the size of frame data.

C. Feature Learning

Our aim to develop a convolutional spatio-temporal auto encoder to learn the normal patterns from training dataset. Our design is bi componential.

D. Convolutional Auto encoder

Auto encoder is usually used for dimensionality reduction. Videos will consist of very large dimensionality. Not all the dimensions are necessary for prediction. Therefore, the auto encoder compresses the given input. As shown in Fig. 2, auto encoder consists of two parts, encoder and decoder. The encoder compresses the original input to low dimension in such a way that there is no information loss.

While compressing it, it will learn the features. The encoded information is given as an input to the decoder to reconstruct the input. The difference between original input and reconstructed output is called reconstruction error. While the compression and feature learning happens in auto encoder, the convolution function learns about structure of normal events and stores in it's memory, which is further used in prediction of abnormal events.

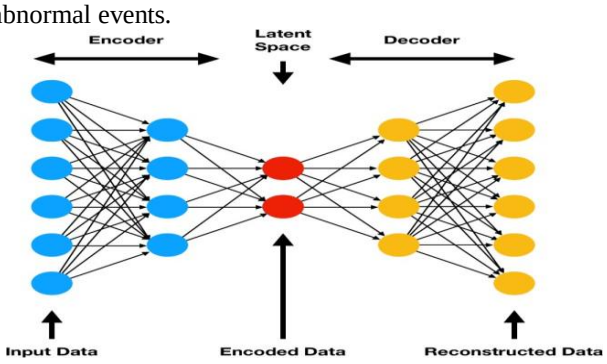

Fig. 2 Auto encoder

As shown in Fig. 3, in convolutional network, filters of desired size are used to extract the representation of required features from the input. Then dot product between filter matrix values and the input values are performed to obtain the desired output.

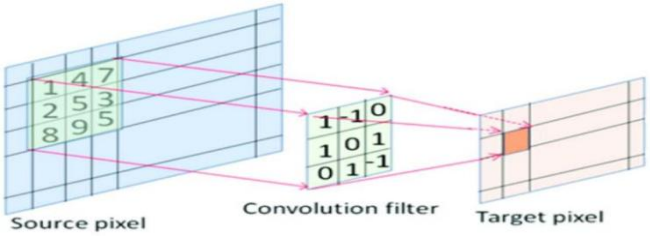

Fig. 3 Convolution Filter

E.Output Regularity Score

Once the model gains the capability it is trained for, we can compute our model's efficiency by feeding in test data and checking whether it performs its task accurately. The main challenge here is to make model generate as less false alarm as possible. By marginally changing the threshold value we can change the sensitivity of the model to reduce false alarm.

F. Thresholding

As shown in Fig. 4, the reconstructed clip is obtained from the spatial encoder-decoder and temporal encoder-decoder. Then the reconstruction error is computed which is the difference between original input and reconstructed input. A thresh- old value is defined on reconstruction error value. The value of threshold depends on requirements. The reconstruction error is compared with threshold, if it's less than or equal to it, it's considered as normal else abnormal. The reconstruction error of normal events will be very less because the network would have learnt the representations of normal events during training phase.

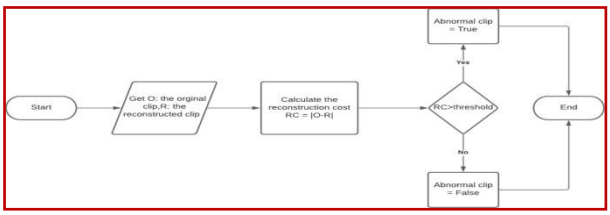

Fig. 4 Thresholding

G.Live Feed

The proposed model implements a live feed module where a live video could be fed as an input and the presence of abnormality could be identified if any.

DESIGN AND IMPLEMENTATION

A.Integrated Design

In the proposed method, a video input is given as a data set. Database set is either in the form of colour or grey scale. The initial step is to convert video frame into image frame after dimensionality reduction. The video data sets is passed on to auto encoder. An auto encoder will be used that learns efficient coding of input data. This data is validated and refined.

The functionality of a convolutional network is to map the information present in the input to matrices which best represents the spatial relationship by preserving as much as information as possible. The proposed architecture is as below. The input to the model is T which is sequence of frames and is fixed and the output is reconstructed input. It takes input of a sequence of length T as input, and output a reconstruction of the input sequence. The number 11x11 denotes dimension of output the output layer. The spatial encoder has the capability of consuming a single frame as an input at a time. After it processes 10(T) frames of data, encoded features of those 10 frames are appended and fed into the temporal encoder which performs motion encoding. The decoder acts similar to the encoders to reconstruct the video volume.

Recurrent Neural Network (RNN) used here works like a feed-forward network(FFN), excluding that the values of its output of input. The autoencoder consists of number of convolution and deconvolution layers for feature learning, it has many number of filters with the required size for the purpose. According to requirements in the filter stride numbers are specified. The convolutional network is a 3D one where height, width and padding of the filter is required for every characteristic learning.

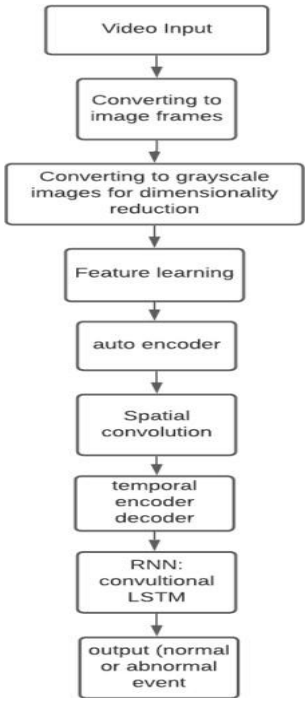

Fig. 5 System design flowchart.

B.Componential Design

An auto encoder is an artificial neural network (ANN), the purpose of it is to learn efficient coding of unsupervised data. This data is validated and refined. Auto encoding consists of two stages: encoding and decoding. It is used for dimensional reduction by generating less units of encoded output than units of input. We use back-propagation in an unsupervised manner to train this model. This is done by downplaying the reconstruction error of the decoding results from the original inputs. Auto encoder draws the features having more relevant and efficacious information than other communal methods like PCA.As referred in Fig. 6 the convolutional LSTM is used to keep memory of previously processed frames to detect motion in the scene. A LSTM has a cell to remember the value of processed frames.

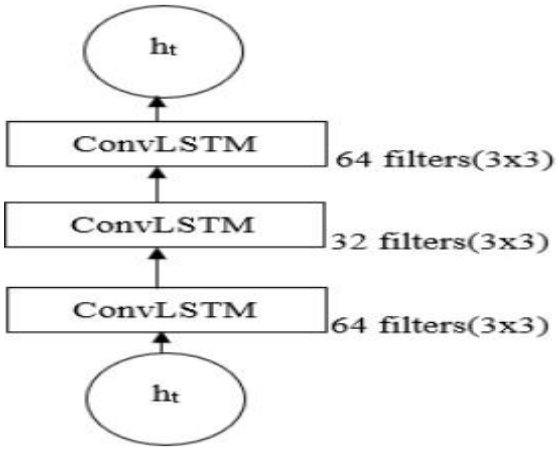

Fig. 6 Temporal encoder-decoder

As shown in Fig. 7, spatial encoder has convolution layers with filters to compress the image frames and to learn structure of image through feature representation.

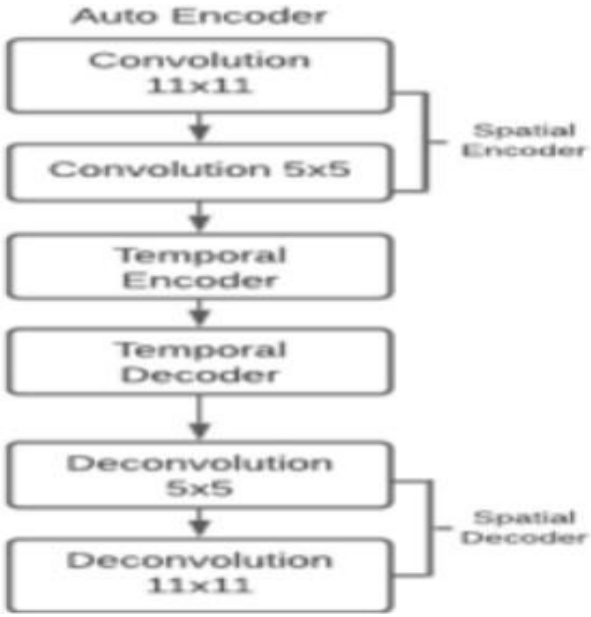

Fig. 7 Anomaly detection and identification

The size of the filters determines to what extent a feature learning can happen. The decoder has deconvolution layers to reconstruct video from the spatially encoded frames. It again has filters for the purpose. It extracts the encoded features and tries to reconstruct it. Every time a new normal image frame is seen, it learns from it and uses it in future for prediction of

abnormality. As shown in Fig. 8, A form of the Long-Short-Term- Memory(LSTM) architecture namely convolutional long short term memory (ConvLSTM) has been recently utilised for pre- diction of frames. Compared to the monotonic usual fully connected LSTM, ConvLSTM for both input to hidden and hidden-to- hidden connections. Convolutional long short term memory (ConvLSTM) requires fewer rates and produce better special feature maps. The formulas of the convolutional long short term memory (ConvLSTM) can be summarised as shown below. Equation (1) corresponds to the forget layer, layers (2) and (3) are where latest data is appended, layer (4) merges both old and new data, whereas layers (5) and (6) outputs the learnt features till that point of time to the Long-Short-Term-Memory(LSTM) unit at the succeeding time step.

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

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

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

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

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

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

The variable xt corresponds to the input vector, ht corresponds to the hidden state, and Ct corresponds to the cell state at a time t. W are the trainable weight matrices b are the bias vectors and the symbol corresponds to the Hadamard product.

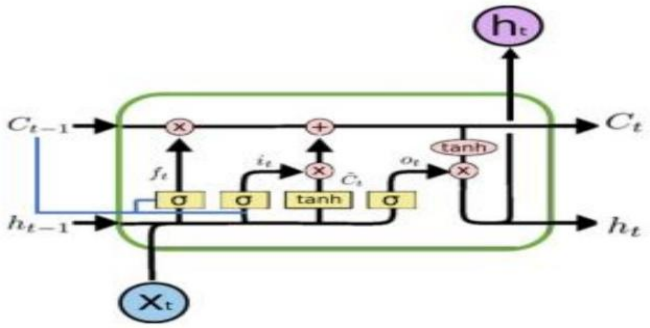

Fig. 8 The architecture of a Long-Short-Term-Memory (LSTM) unit.

EXPERIMENTAL RESULTS

The implemented model is tested on Avenue dataset. It works with UCSD dataset as well. Execution of the implemented model is as shown in the Figure 9. This model gives an option to set the number of epochs and threshold values according to the requirement. The model prints the accuracy and loss values which could be used to select the epochs and threshold values.

The model is implemented with live feed option where the live video could be fed as an input and the model outputs whether the input has abnormality. This model is performing well with live feeds to detect the normal and abnormality accurately as shown in the Figure 10.

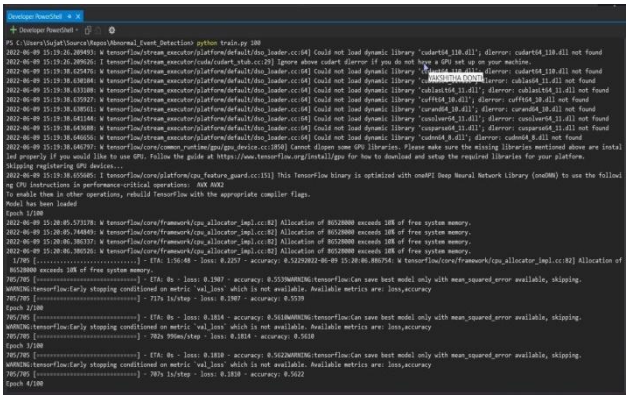

Fig. 9 Model Execution

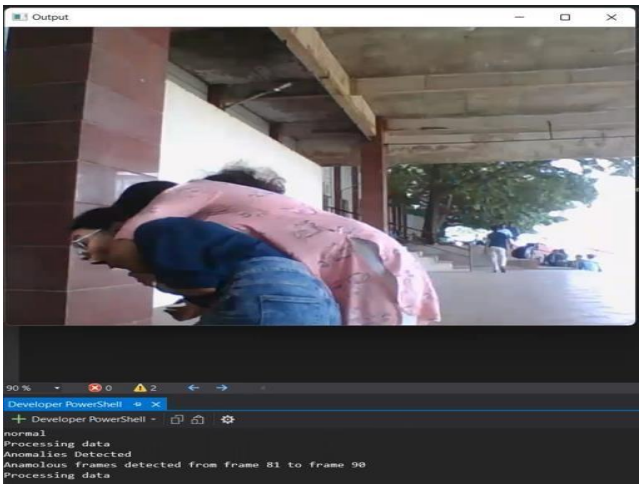

Fig. 10 Detecting abnormal event from the live feed

CONCLUSION

The technique proposed here can efficiently find out the abnormal events in a video input. Deep learning and CNN based this model works for both recorded video and Live video input. All video frames are converted to spatiotemporal sequence and then outliers are detected. As convolution operation is very effective in extracting optimum features from spatiotemporal data, appropriate convolutional layers as spatial feature and convLSTM as temporal feature extractor are used. As it is a semi-supervised approach and it exploits all the features from video inputs given, this model shows novelty and gives better accuracy. This model returns the frame number where the abnormal event is found on to the terminal display.

References

1. Kshitij Barsagade Sumeet Tabhane," Suspicious Activity Detection Using Deep Learning Approach ",1st IEEE International conference on Innovations in High-Speed Communication and Signal Processing (IEEE-IHCSP) 4-5 March, 2023979-8-3503-4595-7 03/2023 © 2023 IEEE
2. Abnormal Event Detection in Videos Based on Deep Neural Networks Qinmin Ma, 2021.
3. Artificial Intelligence Image Recognition Method Based on Convolu- tional Neural Network Algorithm, Youhui Tian, 2020.
4. Abnormal Crowd Behaviour Detection Using Motion Information Im- ages and Convolutional Neural Networks, Cem Direkoglu, 2020.

5. Unsupervised Anomaly Detection and Localization Based on Deep Spatiotemporal Translation Network, Supavadee Aramvith, 2020.
6. Anomaly Detection in Videos Using Optical Flow and Convolutional Autoencoder Elvan Duman and Osman Ayhan Erdem. December 2019.
7. Spatio-Temporal Unity Networking for Video Anomaly Detection, Yuanyuan Li, Yiheng Cai, Jiaqi Liu, Shinan Lang, And Xinfeng Zhang, 2019.
8. Anomaly Detection using Convolutional Spatiotemporal Autoencoder, Hemant Dhole, Mukul Sutaone, Vibha Vyas, 2019.
9. Spatiotemporal Anomaly Detection using Deep Learning for Real- time Video Surveillance Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu, 2019.
10. Two-streams Fully Convolutional Networks for Abnormal Event Detec- tion in Videos, Slim Hamdi, Samir Bouindour, Kais Loukil, Hichem Snoussi and Mohamed Abid, 2019.
11. Two-stage Unsupervised Video Anomaly Detection using Low-rank based Unsupervised One-class Learning with Ridge Regression, En Zhu, Siqi Wang, Jianping Yin 978-1-7281-2009-6 IEEE IJCNN 2019.
12. A Survey of Deep Learning-based Object Detection Licheng Jiao, Fellow, IEEE, Fan Zhang, Fang Liu, Senior Member, IEEE, Shuyuan Yang, Senior Member, IEEE, Lingling Li, Member, IEEE, Zhixi Feng, Member, IEEE, and Rong Qu, Senior Member, IEEE 2019.
13. A Comprehensive Survey of Video Datasets for Background Subtraction Rudrika Kalsotra and Sakshi Arora, 2019.
14. Real-world Anomaly Detection in Surveillance Videos Waqas Sultani, Chen Chen, Mubarak Shah, arXiv:1801.04264v3 2019.
15. Generative Neural Networks for Anomaly Detection in Crowded Scenes, Tian Wang, Member, Meina Qiao, Zhiwei Lin, Ce Li, Hichem Snoussi, Zhe Liu, Chang Choi, IEEE 2018.
16. Anomaly Detection in Video Sequences: A Benchmark and Computa- tional Model ISSN 1751-8644 Boyang Wan1, Wenhui Jiang1 , Yuming Fang1 , Zhiyuan Luo1 , Guanqun Ding, IET Research Journals. The Institution of Engineering and Technology 2015.
17. Learning Spatiotemporal Features with 3D Convolutional Networks Du Tran1,2, Lubomir Bourdev1, Rob Fergus1, Lorenzo Torresani2, Manohar Paluri1 1Facebook AI Research, 2Dartmouth College, 2015.
18. Trajectory-Based Anomalous Event Detection, Claudio Piciarelli, Mem- ber, IEEE, Christian Micheloni, Member, IEEE, and Gian Luca Foresti, Senior Member, IEEE November 2008.
19. Video Classification Using Spatial-Temporal Features and PCA, Li-Qun Xu and Yongmin Li, IEEE 2003.

How to cite this article:

Kusuma. S and Kiran P. (2024). Detecting Anomalies In Videos Using Spatiotemporal Auto Encode. *Int J Recent Sci Res.* 15(02), pp.4574-4579.

****