International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Abnormal Event Detection in Videos using Spatiotemporal AutoEncoder

K. Butchi Raju[1], Sura Ganesh[2] , Bandaru Manas Naidu[3] , Penugonda Ganesh Venkata Shankar Gupta[4] , Mohammed Ansar Ul Haq[5]
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
raju_katari@yahoo.co.in[1] , ganesh030519@gmail.com[2] , manasnaidu946@gmail.com[3] , pandugupta769@gmail.com[4] , haqansar2000@gmail.com[5]

Abstract - ***The detection of abnormal events from surveillance video input is more difficult due to the small number of such occurrences that may occur over time. Through learning a supervised model, this can be achieved to detect such events from video sequences. The proposed system used Deep learning based Convolutional Long Short-Term Memory (Conv-LSTM) networks to learn video sequences as supervised models and to compute the regularity score from the learned frames. Any divergences on the computed scores represents the detection of events as the learning model has represented it through reconstruction of errors. The learned model can be chosen as the best model based on the detection performed through regularity scores. The learning videos of abnormal sequences are very less in real time video and which has to be determined through the model accurately. Thus, the proposed work is more challenging to get the event detection identification at a particular frame accurately. Auto encoder and decoders used in the proposed model. The model is also experimented on Avenue Dataset for detection of abnormal events. The Detection of Anomalous Occurrences from video sequences is accurately detected in experimental results.***

Key Words: Deep learning, Convolutional Neural Networks - Long Short-Term Memory (CNN-LSTM), Supervised learning, Regulatory scores

I. INTRODUCTION

Nowadays, large quantities of surveillance video are recorded, which helps in monitoring areas. Along with this there is an increased need for tracking or detection of objects or its nature to identify abnormal events occurring over the sequences of image frames. The detection of such abnormal activity in video sequence is quite challenging as it is a very small part over the large sequence. Monitoring through manuals means it is quite a difficult task. This explored new opportunities for deep learning implementation over the videos to identify such events. Detecting the abnormal behaviors of objects on the video sequences is more challenging as it requires a much more effective system.

Surveillance videos have very less probability of abnormal events to occur, detecting the events manually requires huge manpower. This prompted researchers to devise a system capable of detecting and segmenting frames with aberrant patterns. The available technology required a lot of processing time and overhead. Existing systems are more heuristics, which make the detection more complex.

Abnormal activities are the one, which can be defined as irregular activities and it can be applied in many real-world applications including security surveillance, ATM monitoring, health monitoring and more. Video processing is a very challenging task as it has a large dimension, noise and large number of objects and large features. Abnormal events are context-based events, which means a person when running in an ATM is considered abnormal, whereas when a person runs in a park is a normal event. Thus, defining an abnormal event is a challenging one. One can understand walking on a subway platform is normal, however, then other people can consider it as an anomaly since it could be abnormal. When it comes to automated learning, this human activity has become extremely difficult as machine learning algorithms do not understand video patterns that are generated for anomalies in real world surveillance.

Supervised machine learning or deep learning are the process, when the videos or images frames are labeled for its human action, which may or may not include any obstructed frames, such as a throng on the frames. These labeled sequences are normally learned by algorithm which is easy and detects the actions based on training videos. However, the process of marking it takes a long time and costs a lot of money. Even though, if it is labeled, it may not include past or future actions of humans. Surveillance video may not be enough to capture and categorize all types of suspicious activity. Some of the activities may occur in future, which may be new suspicious activity or unattended. There were studies which represented suspicious activity detection as a binary classification problem given good results of accurate prediction but they are limited for activities. There were more studies which represented abnormal activity as semi supervised or unsupervised models by using spatial temporal features.

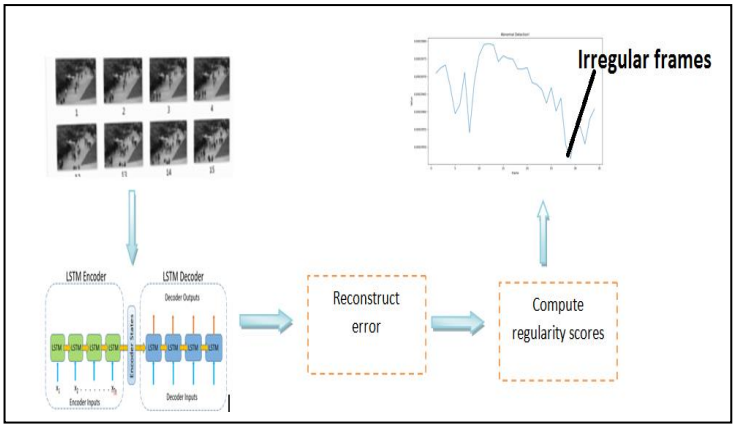

Figure 1: Overview - Abnormal Activity detection

The footage or sequences of abnormal events are very rare in videos and may not be sufficient to train the models. Thus, the recent research on abnormal events is considered by

researchers as semi-supervised or unsupervised models. The little or no supervision models have included spatiotemporal features and auto-encoding techniques for the study. These models required the sequences of videos with less sequence of abnormal or suspicious activities enough to learn and build the model. Thus, the data collection for these models are highly reliable as it is largely available and sufficient. In this proposed study, we considered the Avenue dataset. The objective of the proposed work is to build an efficient model, which can detect abnormal activities from crowded scene videos using extraction of spatiotemporal features and using encoder and decoders. For this, the proposed work used the Convolution network of Long Short term memory models, which is a type of recurrent neural network that remembers past sequences for Hidden Learning in layers.

The following chapters explains more on Literature survey studies on abnormal activity detection in the following chapter I, algorithm details and implementation methodology are discussed in III. In chapter IV, we discussed experimental results and analysis. In chapter V, work concluded with possible future extensions.

II. RELATED WORK

There were existing studies focused on anomalous event detection from videos, which are considered as supervised learning models with some labeled dataset. Some of the researchers have attempted a semi-supervised and unsupervised model for anomalous detection from crowd sequences. Some of them are discussed below.

The author [1] studied learning spatiotemporal features in deep learning methods, the author used 3D ConvNets and studied supervised learning models for action recognition. The author experimented with action similarity measures for different datasets. The optimal temporal kernel length was discovered using Con3D Data. However, these methods are only applicable for labeled video sequences as it was considered as a supervised learning model, a supervised learning model has their events clearly defined and may not have highly occluded scenes or crowded scenes.

Video classification using CNN algorithm was proposed in [2], the author evaluated with YouTube videos 1M sport dataset with 487 classes. The advantage of spatiotemporal studies has been used in this study and given good performance results. Fusions with time information, such as single, multiple, and early frames, are incorporated into the training model. The study was considered as supervised learning as the classes are labeled. Labeling a huge quantity of data has a significant overhead.

Anomalous event detection by histogram of oriented social force was proposed by Yen.et.al in [3]. HOSF is used to extract features from video frames in this case. Dictionary of code words was created with the first few frames and these frames are used for training. For every active particle, HOSF features were extracted and dictionary creation was done. The frames are labeled as normal. The author used dictionary learning techniques labeled with normal frames and done as supervised learning, whereas spatiotemporal features are more effective in our proposed work.

Wang et al [4] studied abnormal events in videos utilizing motion vector direction and magnitude. Motion information entropy obtains abnormal motion from the image frames. After the entropy information extraction, a Gaussian regression distribution model was applied for detecting abnormal events from the frames.

Abnormal events from Surveillance footage by Latent Dirichlet Allocation (LDA) analysis was proposed in [5]. As the storage is a challenging job for surveillance videos, the compressed domain is used in this study. LDA model analysis the motion information and displaced frame difference was computed on the training module. The results show that the model has achieved high accuracy of abnormal event detection.

Another work on compressed domains was studied in [6], the author used High efficiency video coding (HVEC) and motion intensity features extracted from surveillance frames of footage. Anomalies are detected through MEI very fast as the compression domain has provided reliability and less overhead. Extraction of MEI data from frames is the pre-process and further they are analyzed for anomalies by providing motion energy information. The detection time was less in this model and the model was effective on detection of abnormal events in a compressed domain.

Autonomous shuttles mobility environments are considered in this study [7]. The author proposed abnormal passenger behavior detection to avoid petty crimes over the passenger vehicle. Some petty crimes such as bag snatching, fighting is classified from the input frames. The model is implemented with an LSTM classifier for bidirectional studies, spatiotemporal encodes and an LSTM classifier using spatiotemporal features. Pose estimation and tracking is done, the body key points features were collected and fed to LSM classifier for classification. The model achieved good accuracy.

There was more research available in the area of anomalies activity prediction, as discussed above, some of them used feature extraction like HOSF, HVEC, MEI and used for classification. Some of the studies used conventional CNN algorithms for learning and detection. Some of the studies also used auto encoder and decoders for training. Most existing studies were considered as supervised learning, as they used labeled dataset for classification. However, the proposed work has a limited or no supervision model, as the dataset is taken from a large sequence of videos and not labeled. The challenges in detecting unlabeled video sequences are studied in the proposed work. Multiple layers in the hidden layer with autoencoder achieves the feature learning effectively.

III. METHODOLOGY

The advantage of proposed work is that it is considered domain free, which means it can be applied for any area of application if developed and fine-tuned. The application areas range from surveillance, human activity analysis, abnormal events considering spatial features. The efficiency of detection of anomalies in the frames is good in the proposed work and can be applied on real world datasets. The following figure represents the overall architecture of proposed work.

The proposed work abnormal events detection from crowded scene videos using their spatiotemporal features for learning through Conv-LSTM algorithm. The proposed work carried out four major steps 1. Data Collection, 2. Data Preprocessing 3. Training and 4. Detection of abnormal events from videos.

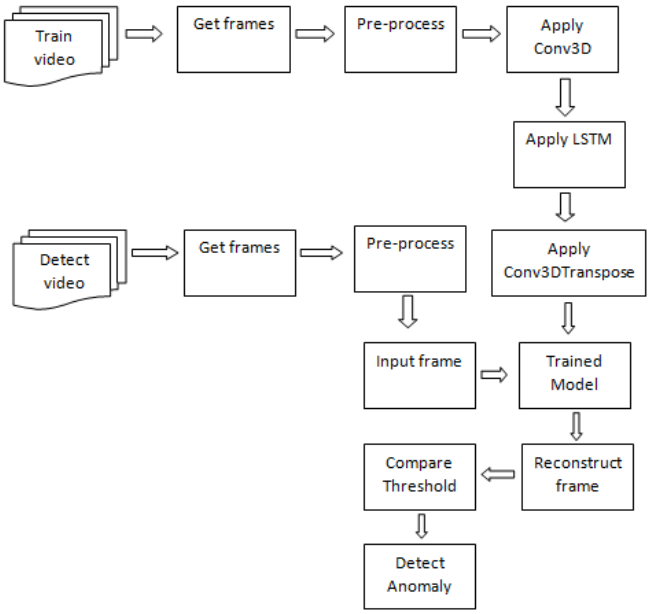

Figure 2: Overall architecture of Abnormal Activity detection

DATASET COLLECTION

Avenue dataset is collected and used for this study. The proposed work is a novel study of learning spatiotemporal features from long sequences of video by Conv-LSTM model by using auto encoders and decoders. Avenue dataset contained a total of 37 videos with little abnormal activity on each video. The following figure shows the collection of datasets as video sequences.

Figure 3: Collection of Avenue dataset

DATA PRE-PROCESSING

The pre-processing processes are completed here in order to make the dataset ready for the algorithm to learn from. These are sequences of pre-processing steps involved in this work, which considered splitting the videos into image frames, image resizing, feature extraction and converting to NumPy array values. The following image shows the video is split into image frames for processing.

The video is split into frames for a minimum of 20 frames. The time limit considered is one second. As the Avenue dataset is captured for enough duration, each video can be reconstructed with a minimum 20 frames to process. Each frame from the long video is taken, then scaled to 227x227 pixels for input. The input images are ensured to have the same size thus scaled between 0 and 1 and subtracted from the mean value of image to bring it as normalized. The mean value is computed by averaging the pixel data in every pixel of each input frame. The next step the images are converted to grayscale values. The pre-processed images are normalized to mean and unit variance and stored as NumPy values. The extracted features are constructed as NumPy array values and stored as training.npy. In the following module, this file will be examined for input to Conv-LSTM.

Figure 4: Video Split into frames in Pre-processing

CONV-LSTM TRAINING

The below figure represents the overall architecture of the auto encoder and decoder in the LSTM model. Encoder takes process input and given output by decoder module as represented below.

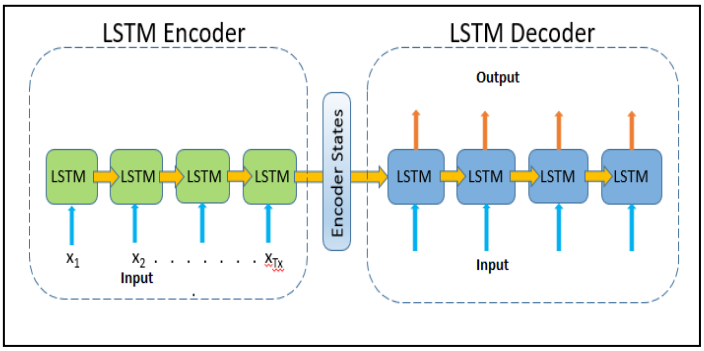

Figure 5: Architecture of Conv-LSTM

The AnomalyDetector.h5 output file was created by the trained Conv-LSTM model which can be used for detection of abnormal activities from videos.

Convolutional network is used here to exploit the nature of extracting the learning features automatically. This model has the quality of preserving spatial relation on pixels by learning image features. The input parameters for learning are given are number of filters, filter size, the number of layers for training. The number of filters used is 32, 64 and 128. The kernel size represented is 3x3 and the number of epochs used for training 3.

In the hidden layer, LSTM memorized the back propagation error, allowing it to parse extended sequences and obtain propagation error at a higher level. Convolutional is used for both inputs to hidden layers and also input hidden to hidden layer connections. This model requires only a few weights and

gives better results as the spatial features are considered. This module takes input sequences of length T and outputs reconstructed sequences. The output size in each layer is represented. Auto encoder takes input as a single frame at a time and finally T is equal to 10 after processing. The encoded features are aggregated and given to encoding. The decoder mirror reconstructs and outputs the volume. The following output screen represents the loss and accuracy generated over 3 epochs of training model.

```
Instructions for updating:
Use tf.cast instead.
Epoch 1/3
2/2 [==============================] - 9s - loss: 0.2995 - acc: 0.6011
Epoch 2/3
2/2 [==============================] - 2s - loss: 0.2475 - acc: 0.5718
Epoch 3/3
2/2 [==============================] - 2s - loss: 0.2184 - acc: 0.5920
Training Finished
```

Figure 6: Conv-LSTM model training

ABNORMAL EVENT DETECTION

The measure of regularity score in this module helps to identify the irregular frames. The normal score represents the regular frames and highly divergences in values represent the irregular frames, in which anomalies are found. The error reconstruction in each processed frame is given as abnormal events. The threshold can be determined using the output graph in order to detect anomalies and generate appropriate alarms at that given frame. The frames are reconstructed as video and the frames, where the abnormal events detected are given with alarm.

The error computed here is the mean square error for reconstruction. The computed divergence value is set as threshold and we define that if computed loss is less than the most divergence values, then frames are considered as irregular frames.

The error values computed differ from video to video from the range of 0.0003 to 0.0004 for Avenue dataset video. Once the anomaly is detected through these error values, we embed a watermark on the frames to give alarm for the abnormal event detection from the crowded sequences. This can easily be identified by the monitoring centre in real world applications.

The proposed anomalous event detection is implemented in Python 3.6.4 with environment Keras and Tensorflow and other mandatory libraries namely OpenCV and matplotlib. Avenue Dataset is used and we trained and tested with two different videos. The Windows application in TKinter is designed to ease user interaction with the application of event detection. On the designed interface, users need to load the input video, which is divided into frames for pre-processing. When the train is started these frames are received and processed with 3 epochs. Then the detection module finds the divergences in error for anomalous detection.

The following images shows the output of anomalous detection from video frames in a surveillance video given as input. A person riding a bike in the opposite way is revealed to be anomalous in this footage.

Figure 7: Anomaly detected in Video

The following images show the output of anomalous activity detected from Avenue dataset video. In this video, a person coming and throwing a bag upside down is detected as an anomalous event.

Figure 8: Anomaly detected in Avenue dataset Video

IV. RESULTS AND DISCUSSIONS

Implementation is done with Avenue Dataset and surveillance video. The two videos are trained and evaluated independently. Surveillance video showed a divergence error value 0.0004 and Avenue video tested was showing a divergence error of 0.0003. The dataset loaded was divided into 20 frames and each frame is processed by the Conv-LSTM algorithm.

To get an optimized model, the experimental set up was given only 3 epochs as the divergence of error in the results were accurate to detect the anomalous at the correct input image frames. It is clearly observed that from the above table, the loss function is reduced with epoch drastically and the

accuracy reduced little over 3 epoch, thus we considered the experimental setup with an optimized value.

Conv-LSTM model with 3 epoch has given the accuracy and loss values as given below

Epoch	Accuracy (in %)	Loss(MSE)
1	60.11	0.2995
2	57.18	0.2475
3	59.28	0.2184

Table 1: Experimental Analysis

The following figure shows the model loss computed as Mean Square error (MSE) by the proposed Conv-LSTM model for the avenue dataset. From the below figure, it is observed that the divergence in value 0.0003 where irregular frames are detected by the algorithm. While representing these values to the detection frames, the frames number are identified and watermarked to reconstruct the resultant video as shown in figures 7 and 8.

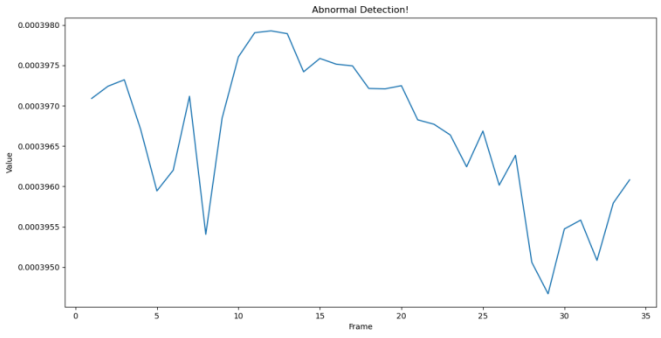

Figure 9: Regularity score achieved by the CNN-LSTM algorithm

V. CONCLUSIONS

The proposed work, Abnormal events from video sequences was implemented in a Deep learning framework. The model used the Conv-LSTM network. The spatiotemporal sequence is considered as outlier identification and spatial features are extracted along with temporal features. These features are fed into the Conv-Long short-term memory model. The model used auto encoders and decoders in extraction of features from the spatiotemporal domain. The model built is considered a semi-supervised learning model. With a divergence detection on the provided error values, the model generated good results on event detection. Experimental results show the performance of detection on Avenue dataset.

In future, the work can be further extended for incorporating human feedback for more accurate detections. The models can be trained and achieved with supervised models to classify the events more accurately.

REFERENCES

[1] Tran, Du & Bourdev, Lubomir & Fergus, Rob & Torresani, Lorenzo & Paluri, Manohar. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. 4489-4497. 10.1109/ICCV.2015.510.

[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725-1732, doi: 10.1109/CVPR.2014.223.

[3] S. Yen and C. Wang, "Abnormal Event Detection Using HOSF," 2013 International Conference on IT Convergence and Security (ICITCS), 2013, pp. 1-4, doi: 10.1109/ICITCS.2013.6717798.

[4] Z. Wang, C. Hou, B. Li, T. Chen, L. Yao and M. Song, "Global Abnormal Event Detection in Video via Motion Information Entropy," 2018 2nd URSI Atlantic Radio Science Meeting (AT-RASC), 2018, pp. 1-4, doi: 10.23919/URSI-AT-RASC.2018.8471516.

[5] A. K. Diop, "Real-Time Abnormal Event Detection in the Compressed Domain of CCTV Systems by LDA Model," 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), 2020, pp. 221-226, doi: 10.1109/ICICSP50920.2020.9232052.

[6] Y. Zhang and H. Chao, "Abnormal Event Detection in Surveillance Video: A Compressed Domain Approach for HEVC," 2017 Data Compression Conference (DCC), 2017, pp. 475-475, doi: 10.1109/DCC.2017.32.

[7] Tsiktsiris, Dimitris et al. "Real-Time Abnormal Event Detection for Enhanced Security in Autonomous Shuttles Mobility Infrastructures." Sensors (Basel, Switzerland) 20 (2020)

[8] W. Sultani, C. Chen and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6479-6488, doi: 10.1109/CVPR.2018.00678.

[9] S. Swarnalaxmi, I. Elakkiya, M. Thilagavathi, A. Thomas and G. Raja, "User Activity Analysis Driven Anomaly Detection in Cellular Network," 2018 Tenth International Conference on Advanced Computing (ICoAC), 2018, pp. 159-163, doi: 10.1109/ICoAC44903.2018.8939064.

[10] M. Irfan, L. Tokarchuk, L. Marcenaro and C. Regazzoni, "ANOMALY DETECTION IN CROWDS USING MULTI SENSORY INFORMATION," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639151.

[11] Chaudhary, Sarita & Khan, Mohd & Bhatnagar, Charul. (2018). Multiple Anomalous Activity Detection in Videos. Procedia Computer Science. 125. 336-345. 10.1016/j.procs.2017.12.045.

[12] A. CHIBLOUN, S. EL FKIHI, H. MLIKI, M. HAMMAMI and R. OULAD HAJ THAMI, "Abnormal Crowd Behavior Detection Using Speed and Direction Models," 2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC), 2018, pp. 197-202, doi: 10.1109/ISIVC.2018.8709192.