Abnormal Event Detection in Video Using Deep Learning

Ms. Deekshitha KN S.[1], MS. Harshitha k.[2], Mr. Mohammed Ibrahim[3] ,Ms. Nakshatra Gowda[4], Mr. Narayana M H[5] Mrs. Aruna M G[6], Dr. Malatesh S H[7]

[1,2,3,4]*Students, Department of CSE, M. S. Engineering College, Bangalore, India*

[5,6]*Associate Professor, Department of CSE, M. S. Engineering College, Bangalore, India*

[7]*Professor and Head of Department, Department of CSE, M. S. Engineering College, Bangalore, India*

***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 spatio-temporal 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-* Deep learning, Neural Network, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), ConvLSTM, Pre-processing and Thresholding.**

I. INTRODUCTION

Video surveillance is widely used for various fields, such as security guards, medical monitoring, traffic monitoring, etc. Among these research fields, anomaly detection plays an important role in discovering various irregularities. Since many first-hand video datasets are obtained without labels and abnormal event is hard to define in advance as well, unsupervised model is more practical and many existing methods about abnormality detection focus on learning the normal pattern, and then abnormal events are identified as those which deviate from the normal ones.

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 pre-defined heuristics, which makes the detection model difficult to generalize to different surveillance scenes.

Deep learning technologies have been widely used to detect abnormal event, including unsupervised methods and weakly supervised methods. Recently, another developing approach for video processing in the deep learning framework is two-stream networks, which have been successfully applied to video-based action recognition, often with state-of-the-art results. Despite the excellent performance, none of these methods considers the black-box problem brought by deep learning models.

II. EXISTING SYSTEM

[1] Liu et al. proposed a framework based on future frame prediction to detect anomalies. However, the prediction method can be sensitive to noise and perturbation, especially in scenes with illumination changes, leading to inferior robustness in anomaly detection.

[2] Qiang et al. proposed an anomaly detection model based on the latent feature space, combining the above two methods. In addition to detecting abnormal events from learning-based techniques,

[3] Yu et al. proposed a neuromorphic vision sensor, a natural motion detector for abnormal objects. Recently, several authors have presented abnormal video detection by the two-stream convolutional network.

[4] Simonyan and Zisserman proposed a two-stream network to recognize the actions of video objects.

SL .No	Title	Method	Advantages	Limitations	Proposed System
1	Robust realtime unusual event detection using multiplefixed location monitors	CNN	Binary class classifications can be studied for detection of abnormal event.	Prediction accuracy is very low. Real-life large-s cale surveillance projects is limited.	It is based on multiple local monitors which collect low-level statistics.
2	Sparse reconstruction cost for abnormal event detection	Sparse model	It is used to detect abnormal event, accuracy is 88 %	It considers only the sparse features. It consumes more time to	Introduced the sparse reconstruction cost (SRC) over the normal dictionary o measure the

				train the modal.	normalness of the testing sample.
3	Behavior recognitionvia sparse spatio-temporal features	Sparse spatio-temporal Auto encoder and decoder	Accuracy of over 80% for the given datasets.	It's not Supporting for the spatio-temporal layout of the features. It is limited only for a few datasets.	The features makes more manageable while providing increased robustness to noise and pose variation. They developed an extension of these ideas to the spatio-temporal.
4	Online detection of abnormal events using incremental coding length	unsupevised approach for abnormal event detection in videos	It supports for detection of online abnormal events.	It takes more computational resources and storage spaces.	Given a dictionary of features learned from local spatio temporal cuboids using the sparse coding objective.
5	Learning temporal regularity in video sequences	Autoencoder and decorder	It increases accuracy up to 82%.	It supports only temporal features and not spatial features, which results in more false detection.	We propose two methods- I) that are built upon the auto encoders for ability to work with no supervision. II) A fully convolutional feed forward autoencoder to learn both the local features and the classifiers as an end to end learning framework

Table.1. Literature Summary

III. PROPOSED SYSTEM

The general workflow for our method includes two streams (spatial stream and temporal stream), which learn features during the encoding stage, and then generate reconstructed sequences of the raw video sequence through decoding. The method can be considered as unsupervised learning scheme in which an autoencoder is trained on the normal data through reconstruction.

If an abnormal event occurs, the corresponding reconstruction error score is higher than the normal data since the model has not met the irregular pattern during training. Besides, we visualized the spatial model's convolutional layer features to identify ways that could help further understand and display the process of model learning at the object level to help people comprehend and trust the detection results of our model.

V. SYSTEM ARCHITECTURE

System architecture illustrates the steps in abnormal event detection system, System takes the video as input to train the model, before training system will convert video into frames, Our system will convert the frames into gray color and apply the CNN-LSTM to train the model. If you pass the input video system will detect the abnormal events occurred frames automatically.

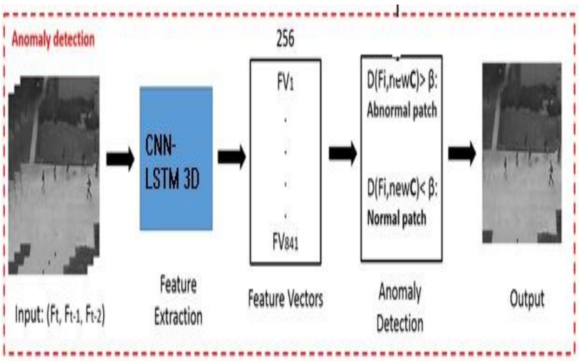

FIG.1. SYSTEM ARCHITECTURE

IV. 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. 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. Our approach consists of three main stages:

A. Pre- Processing

The task of this stage is to convert raw data to the aligned and acceptable input for the model. Each frame is extracted from the raw videos and resized to 227×227. To ensure that the input images are all on the same scale, the pixel values are scaled between 0 and 1 and subtracted every frame from its global mean image for normalization.

B. 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.

C. 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. Mathematically, convolution operation performs dot products between the filters and local regions of the input. Suppose that we have some n×n square input layer which is followed by the convolutional layer.

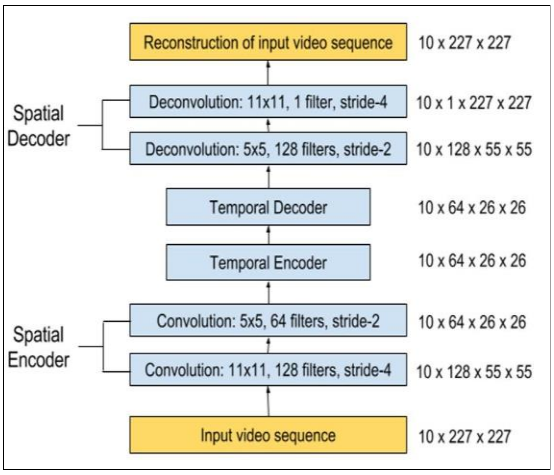

FIG.2 Spatial Decoder and Encoder

VI. RESULTS AND DISCUSSION

Deep learning technologies have been widely used to detect abnormal event, including unsupervised methods and weakly supervised methods. Recently, another developing approach for video processing in the deep learning framework is two-stream networks, which have been successfully applied to video-based action recognition, often with state-of-the-art results. Despite the excellent performance, none of these methods considers the black-box problem brought by deep learning models.

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.

We will investigate how to improve the result of video anomaly detection by active learning having human feedback to update the learned model for better detection and reduced false alarms. One idea is to add a supervised module to the current system, which the supervised module works only on the video segments filtered by our proposed method, then train a discriminative model to classify anomalies when enough video data has been acquired.

We have presented a prevailing method to detect abnormal events from videos to intensify detection ability and feature interpretability with a two-stream framework. Our approach fuses the visual appearances, behavioural characteristics, and motion of the video object and can determine abnormal events from many regular activities. To critically assess the robustness of detecting in capturing abnormal events, we performed several challenging data sets that allow our algorithm to operate robustly for long periods in various scenes, including crowded ones. Experiments have shown that our method is accurate and robust to noise. Furthermore, the visualization of feature maps semanticizes the internal logic.

Meanwhile, applying explainable deep learning methods to anomaly detection will be a future research direction. It has excellent benefits for handling abnormal events and even preventing abnormal events from happening in advance, which has great significance in public security.

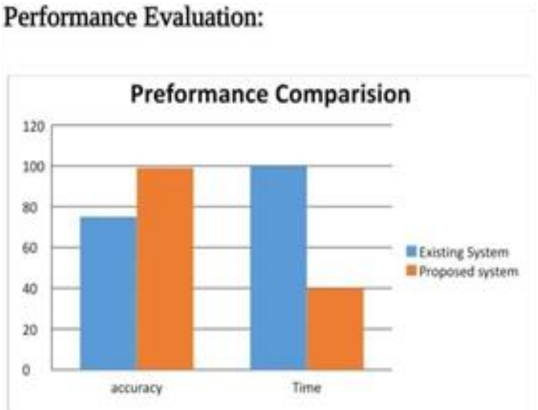

FIG 3: PERFORMANCE EVALUATION

VIII. CONCLUSION

We have presented a prevailing method to detect abnormal events from videos to intensify detection ability and feature interpretability with a two-stream framework. Our approach fuses the visual appearances, behavioural characteristics, and motion of the video object and can determine abnormal events sfrom many regular activities. To critically assess the rebustness of detecting in capturing abnormal events, we preformed several challenging data sets that allow our algorithm to operate robustly for long periods in various scenes, including crowded ones. Experiments have shown that our method is accurate and robust to noise. Furthermore, the visualization of feature maps semanticizes the internal logic.

REFERENCES

[1] M. Hasan, J. Choi, J.Neumann, A. K. Roy-chawdhury, and L.S. Davis, “Learning temporal regularity in video sequences” in proceedings of the 2016 IEEE conference on computer vision and pattern recognition(CVPR), oo. 733-742, IEEE, Las Vegas, NV, USA, June-2016.

[2] R.T. Tonescu, S. Smeureanu, M. Popescu et al., “Detecting abnormal events in video using narrowed normality clusterss”, in proceedings

of computer cision(WACV), pp. 1951-1960, IEEE waikoloa, HI, USA, January 2019

[3] A. Del Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos," in Proceedings of the European Conference on Computer Vision, pp. 334–349, ECCV, Amsterdam, Netherlands, October 2016.

[4] R. Tudor Ionescu, S. Smeureanu, B. Alexe et al., "Unmasking the abnormal events in video," in Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2895–2903, IEEE, Venice, Italy, October 2017.

[5] Y. Fan, G. Wen, D. Li. Et al., "Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder", computer vision and image understanding, vol. 195, article ID 102920, 2020

[6] Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(3), 555–560 (2008)

[7] Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3449–3456 (2011)

[8] Doll´ar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Proceedings - 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, VS-PETS. vol. 2005, pp. 65–72 (2005)

[9] Dutta, J., Banerjee, B.: Online detection of abnormal events using incremental coding length (2015), http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9923

[10] 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)

[11] Ji, S., Yang, M., Yu, K.: 3D Convolutional Neural Networks for Human Action Recognition. Pami 35(1), 221–31 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22392705

[12] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (June 2014)

[13] Kim, J., Grauman, K.: Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009. pp. 2921–2928 (2009)

[14] Kozlov, Y., Weinkauf, T.: Persistence1d: Extracting and filtering minima and maxima of 1d functions. http://people.mpi-inf.mpg.de/~weinkauf/notes/persistence1d.html,