Anomaly Detection in Human Behavior using Video Surveillence

Neha Sharma
Dept of Computer Science Engineering
Ramaiah Institute of Technology
Bengaluru, India
neha2sharma321@gmail.com

Dr. J Sangeetha
Dept of Computer Science Engineering
Ramaiah Institute of Technology
Bengaluru, India
sangeethakirank@msrit.edu

Rohit Kumar
Dept of Computer Science Engineering
Ramaiah Institute of Technology
Bengaluru, India
greeshmaelachitaya@gmail.com

Mohammed Annan
Dept of Computer Science Engineering
Ramaiah Institute of Technology
Bengaluru, India
annankandlur@gmail.com

Abstract*—Traditional passive surveillance is proving ineffective as the number of available cameras for an operator often exceeds the operators ability to monitor them. Furthermore, monitoring surveillance cameras requires a focus that operators can only uphold for a short amount of time.Thus in this review paper the focus is on solving the problem of anomaly detection in video sequence through semi-supervised techniques. Each video is defined as sequence of frames.The model is trained with goal to minimize the reconstruction error which later on is used to detect anomaly in the test sample videos.The model was trained and tested on most commonly used benchmarking dataset-Avenue dataset[7].Experiment results confirm that the model detects anomaly in a video with a reasonably good accuracy in presence of some noise in dataset.

Keywords—video surveillance, anomaly detection, semi-supervised learning, unusual activity, video processing, abnormal behavior

I. INTRODUCTION

With the increasing number of anti-social activities that have been taking place, security has been given utmost importance off late and it is paramount that every citizen plays his share in warranting the safeguarding of our society.

Many organizations have mounted CCTV cameras for the constant monitoring and invigilation of people in public areas and their interactions. For a developed country such as ours, with a population of 1.33 billion, every person is captured by a camera ~ 70 times a day. A lot of video is spawned and stored for a certain time duration. A 704x576 resolution image which is recorded at 25 frames per second will produce roughly 20GB data per day. Since constant monitoring of data by humans to judge if the events are abnormal is a near impossible task, and requires a colossal workforce and constant attention and awareness, it calls for a need to automate the same. Also, there is a necessity to show which frame and which parts of the video contain the unusual activity which can aid in faster judgment of that anomalous activity being abnormal.

Further, the definition of an anomaly depends on what context is of interest. A video event is termed as an anomaly when something unusual happens in the video frame which does not confer to the usual norms. With the rapid growth of video data, there is an increasing need not only for recognition of objects but for detecting the unusual objects withal.

Types of anomalies :

- **Point Anomaly**: A lone instance of data is said to be anomalous if it is too far off from the remainder instances. *Business use case :* Detecting credit card fraud based on the amount spent from that card.
- **Contextual Anomaly**: The abnormality in this case is context definitive. This type of anomaly is frequent in time-series data. *Business use case:* Spending $200 on grocery and food every day during the holiday season is typical, but may be irregular otherwise.
- **Collective Anomalies:** A set of data samples/instances that simultaneously helps in detecting anomalies.

Meaningful events that are of interest in long video data, such as surveillance footage, often have low probability of anomalies occurring. As such, manual detection of these events is a tedious job that often requires more manpower than what is actually available.

Video data, by itself, is challenging to represent and model due to its high dimensionality, noise and a large variety of interactions. Anomalies are also highly contextual and the definition can be ambiguous. Now, there are copious successful cases where anomaly detection has worked well[1,2,7]. However, these methods work by exploiting labelled data which is infeasible and costly. One must record and classify past events and then train the model. This demands for an approach that is increasingly feasible to implement and doesn't burden the programmer

II. LITERATURE SURVEY

This section discusses the diverse research papers that are of consequence to this work and present the underlying features and inferences in them.

Khawaja M. Asim, Iqbal Murtza, and Asifullah Khan in [6] presents a supervised approach for dealing with the problem of detecting anomalies in videos. Taking into account the pixel based approach for identifying anomalies,

the authors have used k-mean clustering algorithm, accompanied by a posteriori probability based probabilistic model, and region intersection technique for discovering anomalies.

The algorithm consists of the following steps :

i. Selection of points of interest
ii. Interest points description
iii. Feature vector clustering
iv. Construction of an ensemble of key-points
v. Training and testing of the algorithm
vi. Region junction

The technique regards normal events as events having higher probabilities of occurrence. Thickly sampled points are delivered to a probabilistic model through the k-mean clustering to attain the probability of episodes.

The k-mean clustering technique is used for quantization of vectors, and is one of the most prevalent methods for cluster analysis. The algorithms dispenses n sample points in the feature space, randomly selects k points as cluster means, and then allocates every observation to the closest mean. The means of each cluster are updated iteratively.

Let there be n observations, [x1, x2,..,xn], where each xi depicts an observation which is a d-dimensional vector. The clustering algorithm partitions the n observation into a user-defined "k" number of clusters where k<=n. Mathematical representation of the algorithm is shown in equation 1.

$$\arg_s \min \sum_{i=1}^{k} \sum_{x_j \in s_i} \left\| x_j - \mu_i \right\|^2 \qquad (1)$$

A threshold value is applied for differentiating the anomalous events from ordinary ones. The ultimate results of abnormal event detection which are attained from multiple scales are put together through region assimilation. The amalgamation of results of multi-scale unusual event detection using region assimilation helps reduce false positive vigorously. The method is tested on the standard UCSD dataset, detecting anomalies with great success.

Chong and Tay in [1] talk about the most common feature for anomaly detection- video feature representation. Ample research has been done in finding the skillful anomaly detection but finding anomaly in videos is still an open challenge.This is because of large variations of its environment, human continual movement, high space-time complexity and the complex dimensionality of video data. The author also mentions that it is awfully difficult for any anomaly detectors to go with the supervised approach because one will have to train the model for every possible situations and the model will have incredible intricacies. Therefore,the author suggests that the semi-supervised approach could help in learning of the video data.

The author also discusses some semi supervised algorithms and optical flow-based descriptor in contrast to trajectory extraction which requires identifying and tracking of objects, while optical flow methods do not depend upon any such preconditions.

In [2], Sabokrou and Fathy discuss another novel approach for anomaly detection. The authors' proposed method can detect real-time anomalies in crowded scenes. Their work treats a video as a set of non-overlapping cubic patches, and is described using local and global descriptors. These descriptors only find the video properties from different features. Adopting simple and cost-effective Gaussian classifiers, the model can distinguish normal activities from the abnormal ones. The work also describes how these local and global features are defined, which is structure similarity between adjacent patches.

The algorithm commences with the creation of a sequence of frames from captured video and the classification of frames as local and global patches using Gaussian distributions after which the final decision is made regarding the nature of those frames.

In [3], B. Ravi Kiran and Ranjit Parakkal talk about a semi-supervised approach to detecting aberrations in videos.

Semi-supervised learning is a sphere of machine learning which makes use of unlabeled data alongside a small measure of labeled data. It is a blend of supervised and unsupervised learning approaches. Research has shown that a large amount of unlabeled data, when used with a small fraction of labeled data for training, can produce great improvements in the learning accuracies of machine learning models.

The procurement of labeled data for learning problems usually requires a skilled person or an experiment to determine the class labels associated with the instances. The cost linked with the labeling process may thus render a thoroughly labeled training set unattainable. In contrast, the acquisition of unlabeled data is comparably inexpensive. In such situations, a semi-supervised learning can prove to be of great practical significance.

The authors first review the deep convolutional architectures for representation of features along with the generative and predictive models for the task of detecting unusual patterns in the video footages. For representing an unusual activity, the authors make use of an "anomaly-mask" that highlights a suspicious activity in the given frame.

For reconstruction modeling, the paper discusses several dimensionality reduction techniques including :

Principal Component Analysis(PCA) : PCA tries to find the directions of maximal variance in the training dataset. Incase of videos, the model aims to achieve the spatial correspondence between pixel values which are segments of the vector representing a frame at a particular instant of time. Taking X as the input matrix having a non-zero mean, we find orthogonal projections that disassociate features in the training data using equation 2:

$$\min_{W^T W = I} \|X - (XW)W^T\|_F^2 = \|X - \hat{X}\|_F^2 \qquad -(2)$$

Here, WTW=I represents an orthogonal reconstructional of the input data matrix X, and the projection XW represents a vector in a lower dimensional space. This reduction in dimensionality captures the anomalous behavior in the samples, since they are not that well reassembled. The Mahalanobis distance between the reconstruction and the original input gives the anomaly score.

Autoencoders : An auto-encoder is an artificial neural network which is used to learn efficient data coding in an unsupervised fashion. They achieve to learn a representation

of a set of data for dimensionality reduction by training the mode to ignore useless or unnecessary data, often termed as noise. Now, along with the reduction lateral, a reconstruction lateral is also learnt, wherein the auto-encoder generates a representation closing resembling the original input from the reduced input which is devoid of irrelevant features or dimensions.

These reconstruction based predictive and generative models build representations to minimize the error of reconstruction from the normal distribution in learning models.

III. Problem Formulation

The following section discusses the problem of detecting unusual activities in frames effectively. The semi-supervised approach is the one that best suits our case. Even though supervised learning methods are the standard, and provide considerably good results, they are just not feasible for large datasets. A camera generates hundreds of gigabytes of video per day, and this video needs to be processed prior to the application of machine learning algorithms. The training dataset requires labelling, and this renders the task extremely time consuming and almost impractical. This leads us towards the unsupervised learning approach, but this method does not provide great results. So we finally stumble upon a hybrid method that takes the best of both worlds and provides the accuracy of supervised models, and the ease of practicality of the unsupervised ones.

A.An Illustrative example

The datasets considered in this work are listed as follows :

- The UCSD dataset [4] which consists of videos of people walking on pavements and footpaths where the appearance of objects like a cycle, or a car in the scene correspond to anomalous events taking place.
- Strolling of humans in unusual locations also amounts to an anomalous activity taking place.
- In the Avenue Dataset , anomalies correspond to strange actions such as a person propelling an object in the air, like papers or a bag.
- In Subway dataset, people moving in the incorrect direction are taken as anomalies.
- In recent times, controlled environment based LV dataset has been brought, with the laborious task of online video anomaly detection.

Fig. 1 depicts the working of the desired anomaly detection model on two datasets : UCSD and Avenue. In UCSD, since the presence of vehicles such a cycles or cars, or even people who are not on foot is considered as an anomaly, we see that the man cycling on the pavement in row one and the skateboarder in the row two are identified as doing something abnormal.

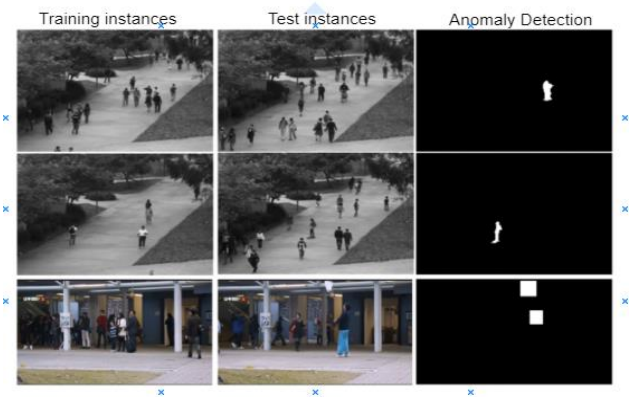

Fig 1. UCSD dataset (top two rows) , portrays the appearance of a cyclist or a skate-boarder on the pavement as an unusual activity. In the Avenue dataset(bottom row), the throwing of papers into the air by a person accounts for an anomaly.

The third row depicts the Avenue dataset wherein the propelling of an object in the air is identified as an anomalous event taking place. Since the person in the frame is throwing a bunch of papers, this act is identified by the model as an anomaly.

IV. Methodology

The method used in this approach is based on the difference between the older bunch of frames and the most recent ones to detect anomaly in the given video. The model is first trained on the normal videos(without any abnormal activities) with the goal in mind to minimize the reconstruction error between the input video sequence and the output video sequence reconstructed by the trained model with the help of an autoencoder. Once the model is trained, the reconstruction error for the normal video is less compared to videos with abnormal events. By setting up a threshold value on the error produced in the testing input, the model is able to detect abnormality in the scene.

Steps involved:

A. Data Preprocessing:

In this the first step is to convert the input video into frames. Then resize it to 227x227.Next, each frame is normalized by scaling the pixel value between 0 and 1.After that, the images are converted to greyscale to reduce the dimensions.We then clip out negative values if any and finally store them in numpy array for further processes.

B. Feature learning:

Feature Learning in general means a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data.To learn regular patterns from training videos, convolutional spatio temporal autoencoder [8] is used.The architecture consists of two parts (i) spatial autoencoder for learning spatial structures of each video frame Fig 2. (ii) temporal encoder-decoder for learning temporal patterns of the encoded spatial structures Fig 3.

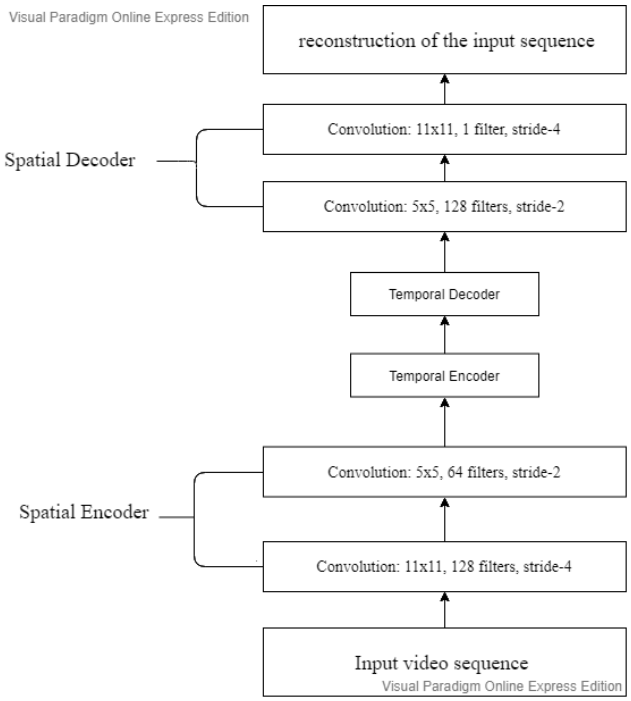

Fig 2.The architecture takes a sequence of length T as input, and output a reconstruction of the input sequence. The dimension is reduced as we go from input and we get back the output with reconstruction value that we tend to decrease in case of training the model using backpropagation algorithm

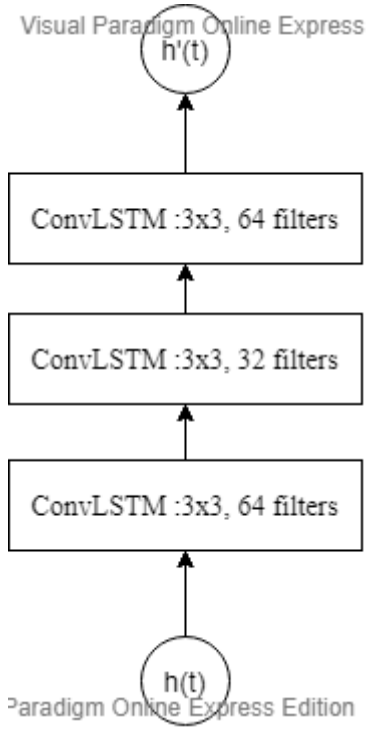

Fig 3:The zoomed-in architecture of temporal encoder-decoder which comprises of 3 ConvLSTM layers at time t.

C. Evaluation Metric:

To evaluate the performance of the model we feed in the testing data and check whether it is capable of detecting anomaly, making sure that false alarm rate is low. By computing the Euclidean distance between input frame and the reconstructed frame , the model flags the video as Anomalous or not depending upon the threshold value that is set during the testing period.

V. Experiment

A. Dataset:

We train our model on most commonly used benchmarking dataset : Avenue [5].In Avenue dataset, there are total 16 training and 21 testing video clips. Each clips duration vary between less than a minute to two minutes long. The normal scenes consist of people walking between staircase and subway entrance, whereas the abnormal events are people running, walking in opposite direction, loitering and etc. The challenges of this dataset include camera shakes and a few outliers in the training data. Also, some normal pattern seldom appears in the training data.

B. Model Parameters:

The model was trained by minimizing the reconstruction error of the input volume. Adam optimizer was used in order to set the learning rate automatically based on the model. Mini-batches of size 64 and the model was trained for 30 epochs. Hyperbolic tangent was used as the activation function for spatial encoder-decoder.

C. Environment and API's used:

Anaconda was used as the developing environment. Major APIs used in the experiment - numpy, keras, tensorflow, scipy.

VI. Result

The data was trained on a GTX 1050 GPU for 30 epochs. Each epoch took an ETA of 30 min approximately. Sequential model from Keras API was used to train and test on the Avenue dataset. The model detected anomaly from an anomalous video successfully with an accuracy of 0.77.

VII. Conclusion

In this review paper we have successfully applied deep learning approach to tackle the problem of video anomaly detection over the Avenue dataset. A spatial feature extractor and temporal sequencer ConvLSTM was used to solve this problem. The ConvLSTM layer not only preserves the advantages of FC-LSTM but is also suitable for spatiotemporal data due to its inherent convolutional structure. For the experiment Keras's predefined layers were used.By incorporating convolutional feature extractor in both spatial and temporal space into the encoding-decoding structure, we build an end-to-end trainable model for video anomaly detection. Although the model is able to detect anomaly from the benchmark dataset, real world scenarios could be more complex and thus there are chances of false alarm. Future work could be done on reducing the false alarm in case of complex environment.

References

[1] Chong, Yong Shean, and Yong Haur Tay. "Modeling representation of videos for anomaly detection using deep learning: A review." *arXiv preprint arXiv:1505.00523* (2015).

[2] Sabokrou, Mohammad, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette. "Real-time anomaly detection and localization in crowded scenes." In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 56-62. 2015.

[3] Kiran, B., Dilip Thomas, and Ranjith Parakkal. "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos." *Journal of Imaging*4, no. 2 (2018): 36.

[4] Mahadevan, Vijay, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. "Anomaly detection in crowded scenes." In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1975-1981. IEEE, 2010.

[5] Lu, Cewu, Jianping Shi, and Jiaya Jia. "Abnormal event detection at 150 fps in matlab." In *Proceedings of the IEEE international conference on computer vision*, pp. 2720-2727. 2013.

[6] Asim, Khawaja M., Iqbal Murtza, Asifullah Khan, and Naeem Akhtar. "Efficient and supervised anomalous event detection in videos for surveillance purposes." In *2014 12th International Conference on Frontiers of Information Technology*, pp. 298-302. IEEE, 2014.

[7] Taylor, Graham W., Rob Fergus, Yann LeCun, and Christoph Bregler."Convolutional learning of spatio-temporal features." In *European conference on computer vision*, pp. 140-153. Springer, Berlin, Heidelberg, 2010.