OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

AUTOENCODER BASED ANOMALY DETECTION IN SURVEILLANCE VIDEOS

Sahil Digikar, Abhijit Chaudhari, Pratik Angre, Rajat Pathak
SKN Sinhgad Institute of Technology And Science, Lonavala
Department of Computer Engineering

Abstract: Surveillance videos are able to capture a variety of realistic anomalies. In this paper, we propose to learn anomalies by exploiting both normal and anomalous videos. To avoid annotating the anomalous segments or clips in training videos, which is very time consuming, we propose to learn anomaly through the deep multiple instance ranking framework by leveraging weakly labeled training videos, i.e. the training labels (anomalous or normal) are at video-level instead of clip-level. In our approach, we consider normal and anomalous videos as bags and video segments as instances in multiple instance learning (MIL), and automatically learn a deep anomaly ranking model that predicts high anomaly scores for anomalous video segments. Furthermore, we introduce sparsity and temporal smoothness constraints in the ranking loss function to better localize anomalies during training.

Keywords: *Anomalies, Surveillance videos, Auto-encoders*

---

1. INTRODUCTION

An estimation suggests that by the end of 2021 more than 1 billion closed-circuit televisions (CCTV) will be installed all over the world. Law enforcement agencies, transportation systems, environment monitoring systems etc use video surveillance systems for security and monitoring purposes For example, transportation agencies use surveillance cameras for traffic management and detection of traffic rules violations. However, due to the massive volume of data generation these monitoring systems are unable to deliver expected results. This has resulted in several errors and under utilization of available surveillance infrastructure hence we need to develop a highly accurate computer vision deep learning algorithm which can detect anomalies in surveillance videos at real time. Real time anomaly detection in surveillance videos plays an important role in ensuring safety and security, hence we also have to make video surveillance systems capable of making important decisions on their own. Events such as accidents, fire in remote areas, robbery, violence an-expected events may need immediate actions to prevent or control the situation which can be achieved by real time detection of such anomalous events.

According to the existing anomaly detection literature anomalies are considered as the unusual events which do not conform to the pattern learned by the algorithm. but for practical implementations, it is not possible to get a dataset which consists of all the possible patterns/events. Often,it is challenging to differentiate between usual and anomalous events. For example, in many publicly available datasets a previously unseen event such as a person riding a bike is considered as an unusual event, yet under different conditions, the same event can be labeled as normal. Hence we need a system which can update its usual and anomalous events continually.

2. LITERATURE SURVEY

- In the paper "Abnormal Event Detection in Video using hybrid spatio-temporal encoder", an unsupervised dynamic sparse coding approach is used.
- This method is based on online sparse constructability of query signals from an atomically learned event dictionary which forms a sparse coding base.
- The paper "Unsupervised learning approach for abnormal event detection" uses two approaches: pure model and hybrid model.

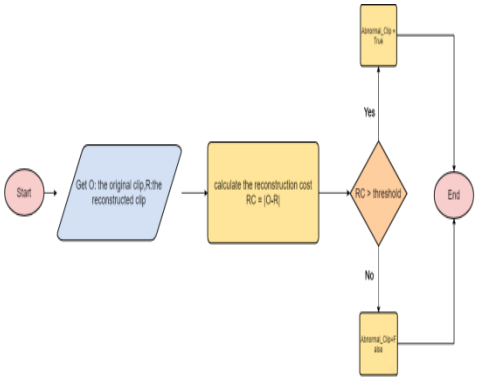
- Hybrid model alleviate the gradient vanishing problem and strengthen feature and encourage feature reuse in both spatial flow and temporal flow
- There are several methods for abnormal event detection according to the applications .

- Few of them are probability based methods or frame based classification methods. Some are rule based
- Abnormal event detection in videos can be done using autoencoders (Variational , Convolutional), generative adversarial nets and Convo nets.
- Some deep learning techniques like Gaussian Mixture Model , Multiple instances Learning can be implemented .
- Currently most methods consist of some steps: Feature computation, transformation of the aggregated features to certain domains, building of model and implementing detection.

3. PROPOSED SYSTEM AND ARCHITECTURE

Our Proposed autoencoder based algorithm has two streams. In the first one we learn common appearances and spatial structure/patterns using Convolutional Autoencoders. The second stream is to determine a connection between each input pattern and its corresponding motion represented by an optical flow of xy magnitude and displacement. The skip connections present in U-Net are useful for image translation since it directly transforms low-level features (e.g. edge, image patch) from original domains to the decoded ones. Such connections are not employed in the appearance stream because the network may let the input information go through these connections instead of emphasizing underlying attributes via the bottleneck. Our model does not use any fully-connected layer, so it can theoretically work on images of any resolution. In order to simplify the model as well as make it appropriate for possible further extensions, we fixed the size of the input layer as 128 × 192 × 3. The image size is set to a ratio of 1:1.5 instead of 1:1 as in related works (e.g.
[11, 42, 25]) in order to preserve the aspect of objects in surveillance videos.

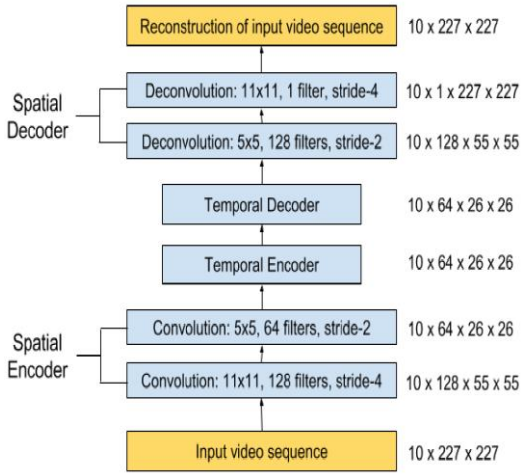
1. After the model is trained with a larger number of videos , the normal scene video is expected to have a low reconstruction error while videos containing abnormal scenes are expected to have a high reconstruction error.
2. By thresholding on the errors produced by each testing video inputs ,the system will be able to detect the anomalies.
3. The test video frames are categorized into abnormal or normal frames using the reconstruction error.



3.1 AUTOENCODER

An autoencoder is a type of artificial neural network used to learn efficient data codings in an unsupervised manner. The autoencoder consists of two parts:

1. The Encoder: Used for learning efficient representations of the input data (x) called the encoding f(x). The last layer of the encoder is called the bottleneck, which contains the input representation f(x).
2. The Decoder: produces a reconstruction of the input data r = g(f(x)) by making use of the encoding in the bottleneck.



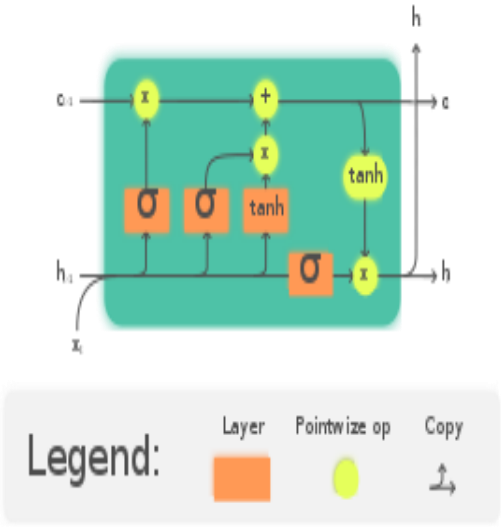
3.2 FEATURE LEARNING

We propose a convolutional autoencoder for learning the regular patterns in the training data.

Our proposed architecture consists of two parts:

1. Spatial autoencoder - which can learn spatial patterns structure of video frames.
2. Temporal encoder-decoder - which can learn temporal pattern structure of videos.

Conv Long short term memory (LSTM)



source: wikipedia

1. Created for sequence prediction problems with spatial inputs, like images or videos.
2. Output is influenced not just by the current input but also by the previous series of inputs to the model
3. Extracted features are given as an input to the LSTM model for the output generation.

3.3 ANOMALY DETECTION

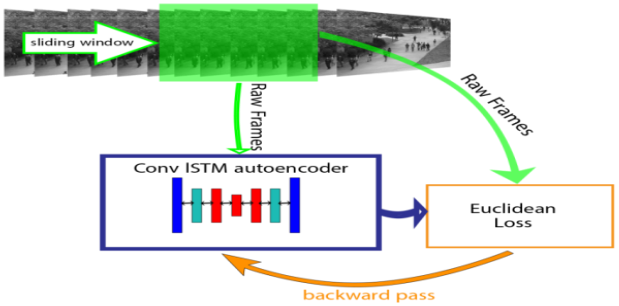
Our main aim is to detect anomalies in videos with much less detection delays and lowering the false positive rate at the same time. For video surveillance, we can say that any anomalous event could persist for an unknown period of time.hence we are using a sequential anomaly detection framework which is more suitable for such an instance. However, as we don't have any prior knowledge about the unusual event that might occur in a video, hence we can't use a traditional parametric model which uses probabilistic models and data. Thus, a non-parametric anomaly detection model would be the best choice for detection of anomalous activities/events.

Training:

Mathematically, convolution operation performs dot products between the filters and local regions of the input.

Suppose that we have some n x n square input layer which is followed by the convolutional layer. If we use an m x m filter W, the convolutional layer output will be of size (n-m+1)x(n-m+1).

A convolutional network learns the values of these filters on its own during the training process



The spatial encoder takes one frame at a time as input, after which T = 10 frames have been processed, the encoded features of 10 frames are concatenated and fed into a temporal encoder for motion encoding. The decoders mirror the encoders to reconstruct the video volume.

Testing:

Once the model is trained, we can evaluate our model's performance by feeding in testing data and check whether it is capable of detecting abnormal events while keeping false alarm rate low.

The reconstruction error of all pixel values I in frame t of the video sequence is taken as the Euclidean distance between the input frame and the reconstructed frame:

$$e(x,y,t) = \left\| I(x,y,t) - f_W\left(I(x,y,t)\right)\right\|_2$$

$$e(t) = \sum_{(x,y)} e(x,y,t)$$

Where fW is the learned weights by the spatiotemporal model. We then compute the abnormality score sa(t) by scaling between 0 and 1.

Subsequently, regularity score sr(t) can be simply derived by subtracting abnormality score from 1:

sa(t) = (e(t) - e(t)min) /e(t)max……(7)

sr(t) = 1 - sa(t)……(8)

3.4 DATASETS:

Our proposed method is first tested on CHUK avenue, UCSD pedestrian and ShanghaiTech campus datasets which are publically available video anomaly datasets.

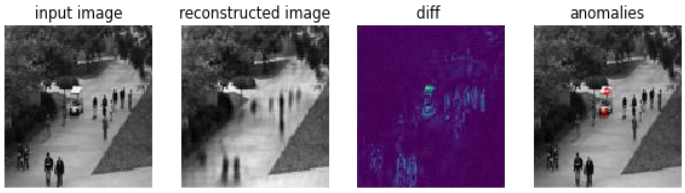
UCSD Ped2: Total of 16 training and 12 test data videos are present in UCSD pedestrian dataset, every video has a resolution of 240 x 360. Vehicles such as bicycles, skateboarders and wheelchairs crossing pedestrian areas are considered as anomalous events in the datasets.

Avenue: The avenue dataset has 16 training and 21 test videos where frame resolution in the video is 360 x 640. The unusual activities here are represented by people throwing objects, scrambling and running.

ShanghaiTech: Among available datasets ShanghaiTech is the most challenging and complex dataset. Resolution of the video frame is 480 x 856 and It has 330 training and 107 test videos from 13 diverse scenes,which make this dataset more different than others.

4. RESULTS AND DISCUSSION

Even though existing datasets such as UCSD, Avenue, ShanghaiTech provide a good baseline for comparison of video surveillance frameworks, they lack some critical aspects. Firstly, they have an underlying assumption that all nominal events/behaviors are covered by the training data, which might not be the case in realistic implementation. Secondly, there is an absence of temporal continuity in the test videos, i.e., most videos are only a few minutes long and there is no specific temporal relation between different test videos. Moreover, Dataset lacks many external factors such as light exposure, brightness, and weather conditions which affects the quality of the images. Hence, we also test our autoencoder based model on a publicly available CCTV surveillance feed.



5. CONCLUSION AND FUTURE WORK

- An efficient method is used to detect the anomalies in the videos consisting of crowded scenes
- We proposed convolutional spatiotemporal architecture for abnormality detection.
- This includes data preprocessing ,feature learning and building of trained models to test the abnormal events in videos.
- It is all about the reconstruction error
- We use an autoencoder to learn regularity in video sequences.
- The intuition is that the trained autoencoder will reconstruct regular video sequences with low error but will not accurately reconstruct motions in irregular video sequences.
- In future, an additional task can be applied to model to get the human feedback and update its learning for better prediction and reduced false alarm rates
- User Interface can be added which allows the user to watch the anomaly video at any time and any number of times
- Immediate alarms can be generated when anomalies are detected thus making humans more alert.

REFERENCES

[1] Oluwatoyin P. Popoola ,Kejun Wang "Video-Based Abnormal Human Behavior Recognition"—A Review Article in IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews) · November 2012

[2] Chong, Yong Shean & Tay, Yong Haur. (2017). Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder. 189-196. 10.1007/978-3-319-59081-3_23.

[3] Waqas Sultani1 , Chen Chen2 , Mubarak Shah2 1Department of Computer Science, "Real-world Anomaly Detection in Surveillance Videos", University of Central Florida (UCF) arXiv:1801.04264v3 [cs.CV] 14 Feb 2019

[4] Abnormal event detection in crowded scenes based on deep learning" Zhijun Fang & Fengchang Fei & Yuming Fang & Changhoon Lee & Naixue Xiong & Lei Shu & Sheng Chen(2018). DOI 10.1007/s11042-016-3316-3

[5] Raksha S, B G Prasad "Anomalous Human Activity Recognition in Surveillance Videos"International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S7, July 2019

[6]Nadeem Anjum and Andrea Cavallaro. Multi Feature object trajectory clustering for video analysis. IEEE Transactions on Circuits and Systems for Video Technology, 18(11):1555–1564, 2008. 2

[7] Michele Basseville and Igor V Nikiforov. ` Detection of abrupt changes: theory and application, volume 104. Prentice Hall Englewood Cliffs, 1993. 4

[8] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and Rene Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 1932-1935

****

e- National Conference

On

Advances in Modern Technologies of Multidisciplinary Research in Engineering Field (AIMTMREF)

[20th -21st May, 2021]

In association with ISTE , IETE and CSI

Address for Correspondence SKN Sinhgad Institute of Technology and Science Lonavala, Pune. 410 401, MS, India.

Website: www.sinhgad.edu

****