International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse011132022.pdf
https://doi.org/10.30534/ijatcse/2022/011132022

Anomalous Event Detection in Surveillance Videos using Spatio-temporal Autoencoders

Sai Kiran Singamaneni[1], Sujith Thota[2], Amutha Prabakaran M[3]
[1] Undergraduate, Computer Science and Engineering, Vellore Institute of Technology, India, singamaneni.saikiran2018@vitstudent.ac.in
[2] Undergraduate, Computer Science and Engineering, Vellore Institute of Technology, India, thota.sujith2018@vitstudent.ac.in
[3] Associate Professor, Computer Science and Engineering, Vellore Institute of Technology, India, amuthaprabakar.m@vit.ac.in

Received Date : April 09, 2022 Accepted Date : May 08, 2022 Published Date : June 06, 2022

ABSTRACT

Surveillance cameras are proliferating, and millions of devices are being used to capture endless footage of surveillance videos. With the advancements in computer vision and deep learning, we can now contemplate these videos to detect anomalies. In this paper, we propose to identify anomalies by using Spatio-temporal autoencoders. The autoencoders based on 3-D convolutional networks will identify anomalies in surveillance footage based on spatiotemporal features. Our architecture consists of two components, an encoder for spatial feature extraction, and a decoder for the reconstruction of frames. Then, abnormal events are identified based on reconstruction loss. We used the Avenue dataset and UCSD dataset for training and evaluation.

Keywords: Computer vision, Deep Learning, Spatiotemporal autoencoder, decoder, convolutional networks, Long short term memory(LSTM).

1. INTRODUCTION

Video surveillance and CCTV cameras are being deployed almost everywhere. They are used in houses, offices, parking places, schools, labs, traffic surveillance systems etc. We can leverage this video feed and use it to train deep learning models which can identify anomalous activities. To ensure public safety, we need to monitor these video feeds constantly for identifying violence, thefts, etc.

Nowadays these surveillance videos are monitored by people. Generally, an anomalous event occurs rarely compared to normal events and monitoring the whole video for these rare events is a tedious task, so to alleviate the waste of time and labour, The use of automated systems powered by artificial intelligence is in rising demand. The aim of this automated deep learning model is to timely monitor and signal an event that deviates from normal patterns. Real-world events are complex and anomalous events are even more diverse and complicated so it is so sensible that the algorithm which we use does not rely on any prior information about the events.

It is a challenging task to identify anomalies from surveillance feeds. For example, a person might think walking around in a crowded area is normal, other people might think it should be flagged as anomalous since it could be suspicious. These challenges are making it difficult for us to create more accurate models to detect real-world anomalies.

To create more accurate models, we need labelled videos, where the events are marked and do not involve any occluded scenes, such as crowded ones. The cost of labelling videos is very high. It is not guaranteed to cover all the past and future events, which makes models less likely to detect real-world anomalies. Therefore, many researchers turned to unsupervised methods, such as Spatio Temporal encoders, which require only unlabelled footage, with little or no anomalies to train and can predict real-world anomalies more accurately.

2. LITERATURE REVIEW

Video-based action detection and recognition have become a challenging area for research in the field of computer vision and deep learning. There are many challenges such as viewpoint variation, clustered backgrounds, camera motion, occlusions, and execution rate in human action recognition, many methods are proposed [1] to address these challenges. This paper reviews various state-of-the-art deep learning-based techniques proposed for human action recognition on various datasets. These datasets are of different types such as single viewpoint, multiple viewpoints, and RGB-depth videos.

Ajeet Sunil [2] proposes the model that determines activities in the video are usual or not. ability to classify and localise the activities in the videos using a single shot detector(SSD) algorithm with a bounding box, which is trained to detect the unusual activities and differentiate them from the usual ones for security surveillance applications. This model is further deployed in public places to improve the safety and security of people. The approach used in designing the SSD model is the transfer learning approach.

Annotating the anomalous segments of training videos is a very time-consuming process. To avoid doing this, Waqas Sultani [3], proposed to learn anomalies through a deep multiple instance ranking framework by leveraging weekly labelled training videos. So the training labels are at video level instead of clip-level. This approach begins with dividing surveillance videos into a fixed number of segments during training. These segments make instances in a bag. Using both positive (anomalous) and negative (normal) bags, they trained the anomaly detection model using deep MIL ranking loss.

D. Kumar [4] proposes a deep learning-based novel framework to identify human actions using the skeleton estimation technique. This is done by the pose estimation using a stacked hourglass network. This will provide the skeleton joint points of

humans. Transformations are applied to these skeleton points to make them invariable to rotation and position. Skeleton positions are identified using HGN based deep neural networks(HGN-DNN), and the feature extraction and classification are carried out to obtain the action class. Before classification, the skeleton actions sequence is encoded with the fisher vector.

Identifying if a person is texting, walking, or fighting from a surveillance feed is a difficult task. Cheng-Bin Jin [6], proposed a sub-action descriptor for action detection. This sub-action descriptor consists of three levels: posture, locomotion, and gesture level. This can detect and recognize the actions of multiple individuals in video surveillance using appearance-based temporal features with multi-CNN. The human action regions are detected by a frame-based human detector and a Kalman tracking algorithm. The action classifier is composed of three CNNs that operate on the shape, motion history, and their combined cues.

Tiriya [7] proposes a transfer learning approach by using VGG16 as a feature extractor and a convolutional neural network as a classifier which is smart video surveillance that could do automatic human behaviour classification that focuses on strategies for reducing the input features and training the model to classify human behaviours. This algorithm removes irrelevant features allowing the model to focus on essential details for better results.

Naik [11] proposes a technique for object detection in both images and videos using YOLO, a SOTA object detector; it is a smart CNN model that detects the objects perfectly. This method is used for multiple object detection in an image and video for traffic surveillance applications trained using a custom dataset created with Indian road traffic images. 9 different classes namely Car, Bus, Van, Truck, Two-wheeler, Auto, Person, Bicycle, and Mini truck are considered in creating a custom dataset. The proposed model gives an average accuracy of 98.32 percent.

3. METHODOLOGY

The anomaly detection method we proposed is based on the idea that when an abnormal event occurs, the most recent frames significantly change. So, we train a model that consists of a spatiotemporal autoencoder and a decoder. The model is trained on a weakly labelled video feed, to minimise reconstruction error, between normal video and abnormal video. After training the model accurately, the normal videos are expected to have low reconstruction error and abnormal videos are expected to have high reconstruction error. We can distinguish between normal and abnormal events by accurately setting a threshold value.

In the preprocessing stage, we extract frames from the raw video and resize them into 227 x 227. And pixel values for the images are set to 0 and 1. After that, frames are converted into grayscale images to reduce dimensionality.

After preprocessing the training data, the frames are passed on to the feature learning model to learn the regular patterns in the training videos. The feature learning model contains a spatial encoder, which is composed of two convolutional and two deconvolutional layers and the temporal encoder is a three-layer Convolutional long short term memory(LSTM) model. The architecture of the model is shown in Figure 1.

The system proposed consists of an autoencoder. Autoencoder is an unsupervised artificial neural network that learns how to efficiently compress and encode data and then learns how to reconstruct the data back from the reduced encoded representation to a representation that is as close to the original input as possible. Autoencoder, by design, reduces data dimensions by learning how to ignore the noise in the data.

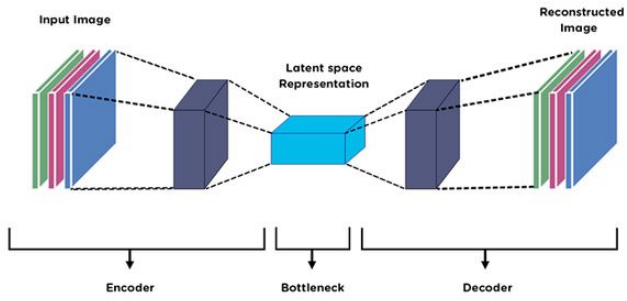

Figure 1 : Autoencoder architecture

The spatial encoder extracts the spatial information and temporal encoder extracts the temporal information, and then the decoder reconstructs the frames. The abnormal events are identified by computing the reconstruction loss using the Euclidean distance between the original and reconstructed batch.

Whether a frame is anonymous or not is classified by the reconstruction error. Here the threshold determines how sensitive the system should behave. If we set the threshold low the detection system will behave sensitively to the events where more signals will be triggered. Sometimes false positives also occur due to low threshold values. So we need to determine an optimal threshold based on datasets and real-world applications.

Once the model is trained, we can test it on testing videos to see if the model can accurately predict abnormal events while keeping the false positive rate low.

4. DATASET

We trained our model on the Avenue dataset and UCSD ped1 and ped2 datasets. All the training videos contained in the datasets contain normal events, and the testing videos contain both normal and abnormal events.

Avenue Dataset contains 16 training and 21 testing video clips. Each training video is a two-minute clip, that contains people walking, standing, etc on an avenue. Training videos only contain normal events, whereas testing videos contain both anomalous events and normal events.

The UCSD Anomaly Detection Dataset consists of ped1 and ped2 folders with each folder consisting of testing and training videos. The videos are of people walking on footpaths taken from an elevation. The abnormal events in the videos mainly include bikers, skaters, wheelchairs, small carts etc.

5. CONCLUSION

In this paper, We have successfully applied a deep learning model for the detection of anomalies in videos. We formulated this detection of anomalies as a spatiotemporal sequence outlier

detection problem. Thus by applying spatial feature extractor and temporal sequence ConvLSTM, we can solve the problem.

6. FUTURE WORK

Further, this can be investigated how to improve the results through active learning, i.e, having human feedback to update the model for better results and reduced false outputs Further this can be investigated how to improve the results by active learning, i.e, having human feedback to update the model for better results and reduced false outputs. A supervised module can be added to the current system, this supervised module works only on video segments filtered by our proposed method, and then train a model that classifies the anomalies when enough video data has been acquired.

7. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our supervisor Prof.Amutha Prabakaran, Senior Lecturer of the Department of Computer Science and Technology, Vellore Institute of Technology, India.

REFERENCES

1. Wu, D., Sharma, N., & Blumenstein, M. (2017, May). **Recent advances in video-based human action recognition using deep learning: A review**. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 2865-2872). IEEE.

2. Sunil, A., Sheth, M. H., & Shreyas, E. (2021, September). **Usual and Unusual Human Activity Recognition in Video using Deep Learning and Artificial Intelligence for Security Applications**. In 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-6). IEEE.

3. Sultani, W., Chen, C., & Shah, M. (2018). **Real-world anomaly detection in surveillance videos**. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6479-6488).

4. Kumar, D., Priyanka, T., Murugesh, A., & Kafle, V. P. (2020, December). **Visual Action Recognition Using Deep Learning in Video Surveillance Systems**. In 2020 ITU Kaleidoscope: Industry-Driven Digital Transformation (ITU K) (pp. 1-8). IEEE.

5. Nawaratne, Rashmika, et al. "**Spatiotemporal anomaly detection using deep learning for real-time video surveillance**." IEEE Transactions on Industrial Informatics 16.1 (2019): 393-402.

6. Jin, C. B., Li, S., & Kim, H. (2017). **Real-time action detection in video surveillance using sub-action descriptor with multi-cnn**. arXiv preprint arXiv:1710.03383.

7. Tiriya, A. A., & Zaveri, M. A. (2020, December). **Human Behaviour Classification for Video Surveillance Using CNN**. In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 769-774). IEEE.

8. Liang, J. (2020). **From Recognition to Prediction: Analysis of Human Action and Trajectory Prediction in Video**. arXiv preprint arXiv:2011.10670.

9. Bornia, J., Frihida, A., & Claramunt, C. (2020, December). **Detecting objects and people and tracking movements in a video using tensorflow and deep learning**. In 2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET) (pp. 213-218). IEEE.

10. Doshi, K., & Yilmaz, Y. (2021). **Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate**. Pattern Recognition, 114, 107865.

11. Naik, U. P., Rajesh, V., & Kumar, R. (2021, September). **Implementation of YOLOv4 Algorithm for Multiple Object Detection in Image and Video Dataset using Deep Learning and Artificial Intelligence for Urban Traffic Video Surveillance Application**. In 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-6). IEEE.

12. Revathi, A. R., & Kumar, D. (2017). **An efficient system for anomaly detection using deep learning classifiers**. Signal, Image and Video Processing, 11(2), 291-299.

13. Liu, W., Kang, G., Huang, P. Y., Chang, X., Qian, Y., Liang, J., ... & Chen, P. (2020). **Argus: Efficient activity detection system for extended video analysis**. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (pp. 126-133).

14. Sun, J., Shao, J., & He, C. (2019). **Abnormal event detection for video surveillance using deep one-class learning**. Multimedia Tools and Applications, 78(3), 3633-3647.

15. Öztürk, H. İ., & Can, A. B. (2021, January). **ADNet: Temporal Anomaly Detection in Surveillance Videos**. In the International Conference on Pattern Recognition (pp. 88-101). Springer, Cham.

16. https://www.warse.org/IJATCSE/static/pdf/file/ijatcse051122022.pdf