Anomaly detection using surveillance videos

Prof. N Sheela[1], Disha Bhat[2], Nihal N Urs[3], Jeevanth D[4], Sanjana S[5]

[1] *Assistant Professor, Department of Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*
[2] *Undergraduate, Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*
[3] *Undergraduate, Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*
[4] *Undergraduate, Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*
[5] *Undergraduate, Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*

ABSTRACT

The videos captured in CCTV/surveillance cameras are not completely utilized in finding out all the anomalies that happens. We aim to detect any abnormal activities using surveillance camera footage, the abnormal activities can be from robbery, shoplifting, arson to fighting, accidents and shooting. We have a huge dataset of 1900 videos of both abnormal and normal videos captured in surveillance cameras. We use a feature extractor for every clip and train a FC network to get a well-trained model. We use c3d feature extractor to get all the required feature in a format of 4096 vector which can be easily used for further training the neural network. We consider to train both normal and abnormal videos as the model should be able to tell the difference between them. The final output of the FC neural network would be some anomaly score and higher the anomaly scores higher it is likely to be an abnormal event in that video. As the dataset is fairly huge, it will be a challenge to preprocess and normalize all the dataset but we also aim to provide the best results for detecting any anomalies in the video.

1. INTRODUCTION

In recent years, though there is continuous advancement in society, the criminal cases are increasing. People have begun to pay more attention to the safety of their lives and property. Real time crime behavior observation is the problem today. In the area of high population density, it is very difficult to continuously monitor videos to detect any anomalies, The aim of anomaly detection system is to timely signal an activity that deviates normal behavior in surveillance video and to identify the time window of an anomaly occurring.

2. LITERATURE SURVEY

The authors “Kothapalli Vignesh, Gaurav Yadav, Amit Sethi” have proposed a method to detect abnormal events for human group activities. They first subtract the background of each frame and then concatenate on the higher order learning only on the foreground. Then the features are extracted using a CNN, that is trained to classify between normal and abnormal frames. These feature vectors are fed into long short-term memory (LSTM) network to learn the long-term dependencies between frames. Finally, they classify the frames as abnormal or normal depending on the output of a linear SVM, whose inputs are the features computed by the LSTM. The limitation of this idea is that

the linear 4 SVM classifier is trained to identify only a particular abnormal activity (a person falling down/pushed). To train all abnormal activities many such classifiers have to be trained which is time consuming and a lot of overhead [1].

The authors "Trong-Nguyen Nguyen and Jean Meunier" proposed a deep convolutional neural network (CNN) for anomaly detection in surveillance videos. They trained only the normal events, so any event that it did not know was considered as an anomaly. The limitation of this model is that it will make a lot of false alarms as anything deviating from a normal trained image will be an anomaly and the model would not know what the real anomaly is [2].

The authors "Zheng Xu, Cheng Cheng, Vijayan Sugumaran" proposed an idea of how the population of people in the surveillance changes as an anomaly occurs. They divided each frame into blocks and used a foreground extraction algorithm to know about the population. When there is a sudden change in number of people or people running in only one direction, the model triggers an alert that there might be an anomaly that has occurred. The limitation of such a model is that it cannot identify which abnormal activity is caused. Another thing is that if a bunch of small kids are playing and running in the video frame, it would trigger an alert as the classifier does not know what the anomaly is [3].

The authors "Keval Doshi and Yasin Yilmaz" proposed a concept of "escape center" which is used to approximate the abnormal scattered behavior of the crowd. The algorithm to detect the single escape center was implemented. To further detect the possible location of anomalies in the actual scene, based on the single escape center, further research was conducted to obtain multiple localization algorithms for the escape center. They detected intrusion behavior using algorithms that directly judge on pixel coordinates. It was the same in the case of traditional algorithms. But this approach gave high accuracy in target detection and target tracking stage. Although it has achieved good detection results, it is suitable only for the detection of abnormal behavior of low and medium crowds [4].

2. DATASET

The Dataset we worked on consists of 1900 real world videos captured by surveillance cameras. This dataset includes different types of anomalies, it can be categorized into 13 different anomalies. Each of these anomalies has about 50 to 100 videos in then captured by the surveillance cameras. We have used Normal videos too so that the model can differentiate easily between an abnormal/anomaly video and a normal video. These 13 different categories include abuse, arrest, fighting, robbery, arson, assault, burglary, explosion, accidents, shooting, shoplifting, stealing, vandalism [5].

Fig 1. Dataset containing videos

2.1 Dataset Division

Our dataset is divided as a set of abnormal videos which has all the anomalies and another set is the normal videos. We need to test both normal and abnormal to find the best and accurate result of the accuracy of the model which is build. We are dividing the dataset as 75% used for training and 25% used for testing only. Testing dataset has a total about 290 videos. Among them 150 is normal videos and 140 is abnormal videos The result of testing the 150 normal videos and 140 abnormal videos is as shown in the table below

# of videos	Anomaly
50 (48)	Abuse
50 (45)	Arrest
50 (41)	Arson
50 (47)	Assault
100 (87)	Burglary
50 (29)	Explosion
50 (45)	Fighting
150 (127)	Road Accidents
150 (145)	Robbery
50 (27)	Shooting
50 (29)	Shoplifting
100 (95)	Stealing
50 (45)	Vandalism
950 (800)	**Normal events**

Fig 2. Division of dataset

3. DATA-PREPROCESSING

Pre-processing the data is the most important step in any project. The characteristics of the data present gives us the sense how accurate a model is going to be when we train it. We got a huge data from CVPR website which is about 95 gigs big. We first normalised all the video to be of a particular resolution so that when we extract a frame from it, we get the same size each time for any video. We trimmed the video to a resolution of 240x320 resolution, now each frame in the video will have same number of pixels. The videos may have different frame per second rate, so secondly, we changed the FPS (frame per second) to 30 so that there will be consistency in how much frames a particular video will have and how much it is used to train the model. Further a video is divided into 40 clips and these clips are further divided into 16 frames each as shown below.

4. METHODOLOGY

4.1 Feature extraction:

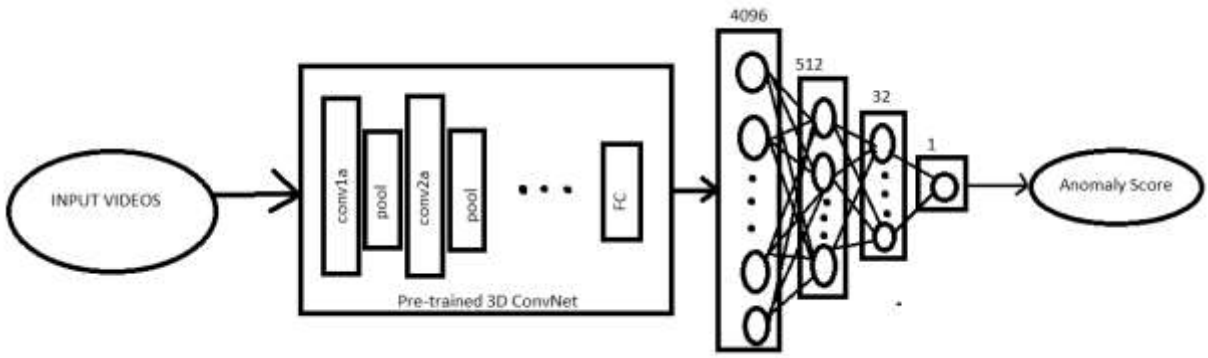

Fig 3. Structure of the C3D feature extractor

Feature extraction of the video is done by using pre trained weights which was developed by Facebook and later was made open sourced called C3D feature extractor. C3D is discriminative, compact, and efficient to compute. The C3D model is given an input video segment of 16 frames (after down sampling to a fixed size which depends on dataset used) and the outputs a 4096-element vector. C3D is obtained by training a deep 3D convolutional network on a large annotated video dataset. The dataset contains various concepts encompassing objects, actions, scenes and other frequently occurring categories in videos.

The 3D convolutions extracts both spatial and temporal components relating to motion of objects, human actions, human-scene or human-object interaction and appearance of those objects, humans and scenes. Thus, it is not limited to appearance representation. The feature is extracted for every 16-frame video clip, then we take the average of all the 16 frame clip features and store it in a file, this file contains 4096 element vector which is used further to train fully connected network. Thus, we get one .txt per video. This text file has stored features that can be used to train and build our model further.

4.2 Training

Using the features extracted in above step we input it into a 3-layer FC neural network. The first FC layer has 512 units followed by 32 units and 1 unit FC layers. 60% dropout regularization is used between FC layers. Dropout is used to avoid overfitting of the model. We use ReLU activation and Sigmoid activation for the first and the last FC layers respectively, and employ Adagrad optimizer with the initial learning rate of 0.001. We divide each video into 32 non-overlapping segments and consider each video segment as an instance of the bag. We randomly select 16

positive and 16 negative bags as a minibatch. We compute gradients by reverse mode automatic differentiation on computation graph using Theano. Then we compute loss and back-propagate the loss for the whole batch.

Our approach assumes that given normal and abnormal videos with video-level labels, the network will learn to predict the location of the anormal part in the video. In order to achieve this, the network should learn to produce high scores for abnormal part in video during training. For less iterations, the network produces high scores for both abnormal and normal video segments. Therefore, by increasing the number of iterations the network learns more videos, hence detects anomaly precisely. Here we are not using any segment level annotations, but still the network is able to predict the temporal location of an anomaly.

5. EXPERIMENTAL RESULTS

Our dataset is divided as a set of abnormal videos which has all the anomalies and another set is the normal videos. We need to test both normal and abnormal to find the best and accurate result of the accuracy of the model which is build. We are dividing the dataset as 75% used for training and 25% used for testing only. Testing dataset has a total about 290 videos. Among them 150 is normal videos and 140 is abnormal videos The result of testing the 150 normal videos and 140 abnormal videos is as shown in the table below.

Videos for testing	# of videos	Predicted correctly	Wrong prediction	Accuracy
Normal Videos	150	129	21	0.86
Abnormal videos	140	135	5	0.96

Table 1: Testing video details

As we can see out of 150 normal videos, model predicted 129 videos correctly and 21 videos incorrectly. From the 140 videos of the abnormal videos, it predicted 135 videos correctly and 5 videos incorrectly. So the overall accuracy of the model can be calculated by merging both the abnormal videos testing and normal videos testing,

Overall Accuracy = (Normal videos predicted correctly + Abnormal videos predicted correctly)/Total testing videos
Overall Accuracy = (129+135)/290
Overall Accuracy = 0.9103
Accuracy = 91.03%

5. CONCLUSION

We propose a deep learning approach to detect real world anomalies in surveillance videos. Due to the complexity of these realistic anomalies, using only normal data alone may not be optimal for anomaly detection. We attempt to exploit both normal and anomalous surveillance videos. To validate the proposed approach, a new large-scale anomaly dataset consisting of a variety of real-world anomalies is introduced. Our future work includes, doing necessary changes that helps in our accuracy and how robustly it predicts the abnormal activities. In this project we will use c3d feature extractor which was made open source by Facebook and we try other feature extractor which maybe more promising in helping us improve the accuracy of our model. The other thing which can be more improved it experimenting with the number of layers in the fully connected neural network and using different activation function for each layer and see which model structure gives best results. Since the videos we get in CCTV are not that good, so our model should be able to detect any abnormal activities even though the quality of the video is compromised. The model working in dim light can be a challenge as c3d feature extractor may not be able to

extract suitable features for the training of our model. This project can be used by the government itself for people's safety and to have law and order in the city, so it definitely has market value.

6. ACKKNOLWDGEMENT

Throughout the process of creating this software, we had to take some help and guidance from few people. They deserve our utmost gratitude. We would like to express our special thanks to our project guide, Prof. N Sheela, who gave us the opportunity to do this great project on the topic "Real-time Anomaly Detection in Surveillance Videos" and gave us guidance and consultations throughout. We would also like to thank our department head, Dr.M.P. Pushpalatha, who gave us a platform to showcase our work. We express our hearty gratitude to all of them. Lastly, we would like to say thanks to fellow teammates without whom this project would not be successful. We thoroughly enjoyed the process of making this software together.

REFERENCES

[1] Abnormal Event Detection on BMTT-PETS 2017 Surveillance Challenge by "Kothapalli Vignesh, Gaurav Yadav, Amit Sethi.

[2] Hybrid Deep Network for Anomaly Detection by "Trong-Nguyen Nguyen and Jean Meunier".

[3] Big data analytics of crime prevention and control based on image processing upon cloud computing by "Zheng Xu, Cheng Cheng, Vijayan Sugumaran".

[4] Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate by "Keval Doshi and Yasin Yilmaz".

[5] Dataset-https://visionlab.uncc.edu/download/summary/60-data/477-ucf-anomaly-detection-dataset