Suspicious Activity Detection Using Surveillance Footage

Group Name-Tranquilizer
Kamal Budhory
Subhojit Sarkar
Debayan Dutta
kamalbudhory18@gmail.com
sarkarsubhojit73@gmail.com
ddebayan09@gmail.com

June 26, 2022

Abstract

In today's world there is a concern of every individual to be safe in the environment they are living. The main reason behind this concern is increase in the activities that relates to possible threats and crime. Suspicious activities are one of the biggest problem when it comes to the potential risk it brings to humans. It deals with identifying the pattern and events that vary from the normal stream. In a surveillance paradigm, these events range from abuse to fighting and road accidents to snatching and many other unusual activities. With the increase in criminal activities in urban and suburban areas, it is necessary to detect them to be able to minimize such events. As we know the surveillance in early days was done manually by humans and was a tiring job as suspicious activities were uncommon compared to the usual activities. The frequency of reporting of unusual activities by using surveillance cameras may be low or high and to wait for the unusual activity to happen is a cumbersome job. The manpower to be deployed for such work has to be more as single person can't focus for longtime looking at a small screen. The approaches reported can be generally categorized as handcrafted and deep-learning based. Most of the reported studies address binary classification i.e. anomaly detection from surveillance videos. But these reported approaches did not address other anomalous events e.g. abuse, fight, road accidents, shooting, stealing, vandalism, robbery etc from surveillance videos. With the arrival of intelligent surveillance systems, various approaches were introduced in surveillance. We focus on analyzing some cases, those if ignored could lead to high risk of human lives, which are detecting abuse, arrest, assault, burglary, explosion, fighting, stealing, robbery e.t.c on frames of surveillance footage. By this project we will present a

neural network model that can identify such crimes. The overall objective of this project is to design such a model which can classify the suspicious activities.

Here, we focus on analysing 13 cases, which if ignored can cause a potential risk to human lives, Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accidents, Robbery, Shooting, Shoplifting, Stealing, Vandalism along with Normal Class as the $14^(th)$ class. We present a Dual CNN deep neural network model that can detect mentioned unusual activities and helps in possible elimination of the threat.
Code has been made available here

1 Introduction

The crime rate in today's world is increasing on a daily basis. The real problem arises when the crimes go unnoticed and lead to possible threat in humans. There's always a concern of every individual to be safe in the environment they are living. People have started taking steps to ensure safety for them and people living n the surroundings. Most of Developed cities, Urban area, Metropolitan Cities, suburban areas and even in the village area surveillance cameras have been installed. The installation of surveillance cameras has been a very big step towards overcoming this problem. Though the surveillance cameras has been installed still the problem remains as surveillance footage are recorded one and the crimes takes place way back than what we are seeing in the recording. This problem can be overcome by making a person sit in front of the screen and have a continuous surveillance. This solves the problem of detecting crimes in the real time. Though we have deployed a person to have continuous look at the screen but to look at the screen 24x7 is very tiring job to do. We can solve this problem by deploying more manpower to have a look at the screen in different shifts but it will be a costly solution.

The problem can be solved in better way which is cost effective as well as independent of a person looking at the screen. A security system that is fully automatic and detects the possible threat in real time and sends an alert to the control room so that helping aid can be provided on time and the threat can be eliminated. Using a deep neural network we can build a model that helps us to solve the problem of crime detection and also helps us to get rid of difficulties we were facing while deploying the above mentioned solutions.

The Field of computer vision is growing a lot in today's scenario. The computer vision helps us to build a deep neural network model to solve many problems related to video data points and image data points. Using the deep neural network we can build a model that helps us to solve the problem of crime detection and also helps us to get rid of difficulties we were facing while deploying the above mentioned solutions.

Suspicious human activity recognition from surveillance video is an active research area of image processing and computer vision. We know that videos are nothing but just a collection of images arranged in sequence in which the action is taking place in the video. The images so arranged are also known as frames and we can also say that a video is set of

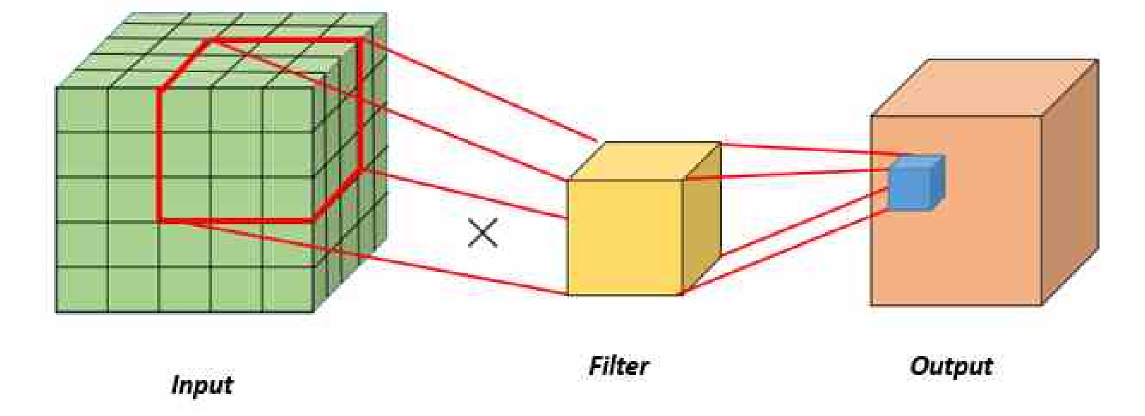

Figure 1: A general 3D CNN model
[7]

frames arranged sequentially in order of their occurrence in the video. As we have an idea that video is collection of frames or images, so the video classification is similar to image classification. For the purpose of classifying images we just need to define a neural network which extracts the features from the images and use the extracted features to classify the images. Similarly, we can do for video classification, first we need to divide the videos into frames then pass the frames to the defined neural network and extract the features from it. Using the extracted features we can classify the videos. The video coming from the surveillance cameras will follow the same set procedure to first train the deep neural network model. The trained model will then be used to detect the action that is taking place in the real time and detect the crime is happening.

2 Literature review

In Literature, several works are discussed that propose the recognition of human activities. The first approaches used to integrate human activities were based on the human joints trajectories (Campbell and Bobick, 1995; Niyogi and Adelson, 1994). These methods need specific techniques to detect parts of the body or to track them in each image. Another approach that has already been explored is based on Bag-of-Words (BoW) [1], this type of proposal requires a large storage space for the less frequent features, besides the need to combine with other techniques for classification and extraction of Features. With the success of Deep Networks, in special the AlexNet in 2012 , the deep models are being explored recently by different researchers. Among the deep architecture models, Convolutional Neural Networks (CNNs) gained attention because of their ability to learn contextual relationships between features . This type of architecture has already achieved great results in domains such as digit recognition, speech recognition, emotion recognition. Some authors have tried to apply the CNN for action recognition in videos. In the work of Ji et al[2]. , they proposed the use of Convolutional Neural Network with two flows for the recognition of the actions, the first is the raw frame and the second is the optical flow with the temporal information of the movement between the frames. In the work of Wang et al[3]. , they have developed a two-channel Convolutional network based on RGB frames. The first input of the model is the raw image and the second is the optical flow extracted from the motion in order to predict the trajectories. In the work of Lin et al[4]. , they use a CNN to decompose temporal information by separating groups by sub-actions into RGB-D videos. In the work of Chron et al[5]. , in order to determine the action in videos, they proposed a descriptor based on body posture. Initially, they calculated the optical flow between frames and motion in each x and y direction. With the information of the directions, they calculate the movement of each part of the body. Finally, they use the raw frame and the flow of motion as input to a CNN for action recognition. Other authors have analyzed the option of using the 3D Convolutional Neural Network (3DCNN) for the recognition of human activities in videos. This kind of architecture can process temporal information, what has meaning for applications in videos . In the work of Ji et al. , they used gray scale images, gradient and the optical flow along the x and y axes, taking these values as input in a 3DCNN for the recognition of actions in secure videos.The latest work is done by Facebook AI research team in 2019. They use a 'slowfast' model to analyze human activity[6] .It presents a novel method to analyze the contents of a video segment, achieving state-of-the-art results on two popular video understanding benchmarks .— Kinetics-400 and AVA. At the heart of the method is the use of two parallel convolution neural networks (CNNs) on the same video segment — a Slow pathway and a Fast pathway.

3 Proposed methodology

SlowFast uses a slow, high-definition CNN (Fast pathway) to analyze the static content of a video while running in parallel a fast, low-definition CNN (Slow pathway) whose goal is to analyze the dynamic content of a video. The technique is partially inspired by the retinal ganglion in primates, in which 80% of the cells (P-cells) operate at low temporal frequency and recognize fine details, and 20% of the cells (M-cells) operate at high temporal frequency and are responsive to swift changes. Similarly, in SlowFast the compute cost of the Slow pathway is 4x larger than that of the Fast pathway.

3.1 Slowpathway

Both the Slow and Fast pathways use a 3D ResNet model, capturing several frames at once and running 3D convolution operations on them.The Slow pathway can be any convolution model (e.g., [12, 49, 5, 56]) that works on a clip of video as a spatiotemporal volume. The

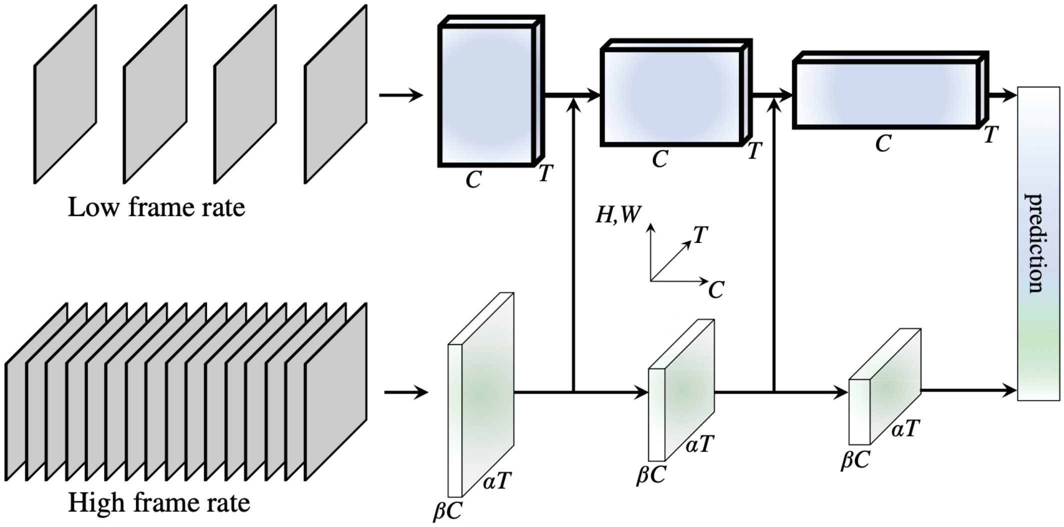

Figure 2: SlowFast Model[6]

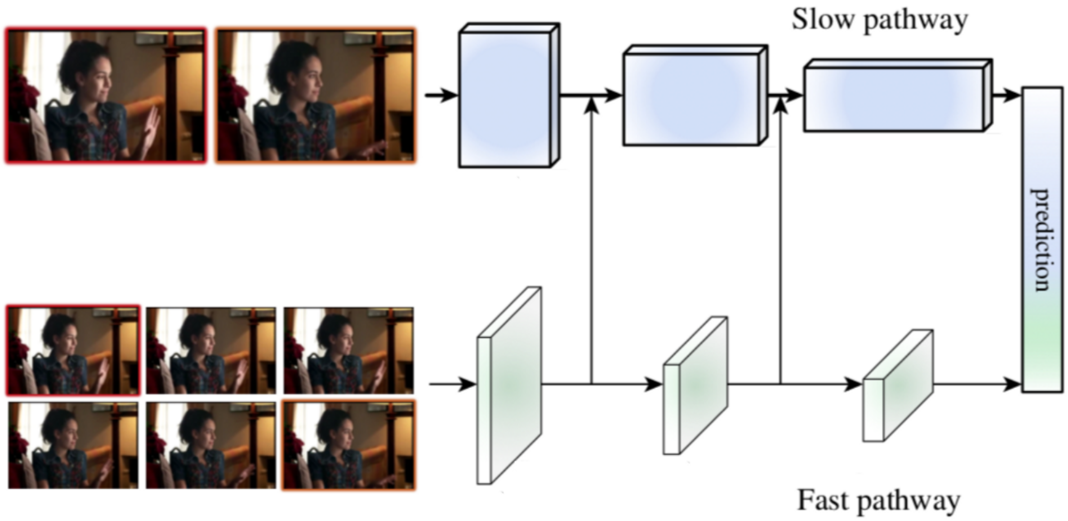

Figure 3: SlowFast Model[6]

key concept in our Slow pathway is a large temporal stride τ on input frames.i.e., it processes only one out of τ frames. A typical value of τ we studied is 16—this refreshing speed is roughly 2 frames sampled per second for 30-fps videos. Denoting the number of frames sampled by the Slow pathway as T, the raw clip length is $T \times \tau$ frames

3.2 Fastpathway

In parallel to the Slow pathway, the Fast pathway is another convolution model with the following properties.It has the high frame rate than the slow model. But it has the lower temporal stride than slow pathway. Fast model works with a smaller temporal stride $\tau \div \alpha$ where $\alpha > 1$ is the frame rate ratio between the Fast and Slow pathways. The Fast pathway is kept lightweight by using a significantly smaller channel size (i.e. convolution width; number of filters used), typically set at $\frac{1}{8}$ of the Slow channel size upto res_3 and for the last layer we have set the Fast channel at $\frac{1}{4}$ of the Slow channel size.

3.3 Lateral Connections

The lateral connections here means information of one pathway is passed on to the other pathway. The two pathway are always aware of each other and are also aware of what representations the pathways are learning. We have also implemented the lateral connections in our network which is used to fuse optical flow-based, two stream networks.

As seen in the visual representation, the information is being fed to the slow pathway from the fast pathway making sure that the slow pathway is getting the information about the fast pathway regularly. The shape of single data point is different between the two pathways, for the fast pathway it is Fast(αT,S^2,βC) and for the slow it is Slow(T,S^2,$\alpha\beta$C), which makes the Slow pathway to perform data transformation on the results we are getting from the fast pathway through the lateral connections and the transformed data is then fused into slow pathway with help of performing concatenation and summation.

stage	Slow Pathway	Fast Pathway
data layer	stride 16, 1^2	stride 2, 1^2
$conv_1$	1×7^2, 64 stride 1, 2^2	5×7^2, 8 8 stride 1, 2^2
$pool_1$	1×3^2 max stride 1, 2^2	1×3^2 max stride 1, 2^2
res_2	$\begin{bmatrix} 1\times1^2, 64 \\ 1\times3^2, 64 \\ 1\times1^2, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 3\times1^2, 8 \\ 1\times3^2, 8 \\ 1\times1^2, 32 \end{bmatrix} \times 3$
res_3	$\begin{bmatrix} 1\times1^2, 128 \\ 1\times3^2, 128 \\ 1\times1^2, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times1^2, 16 \\ 1\times3^2, 16 \\ 1\times1^2, 64 \end{bmatrix} \times 4$
res_4	$\begin{bmatrix} 3\times1^2, 256 \\ 1\times3^2, 256 \\ 1\times1^2, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 3\times1^2, 64 \\ 1\times3^2, 64 \\ 1\times1^2, 256 \end{bmatrix} \times 6$

Table 1. SlowFast Architecture

4 Experimental result

Training Our models on DCSASS are trained from random initialization ("from scratch"), without using ImageNet or any pre-training. We use Adam optimizer for training. At first we transformed the total training video dataset into certain frames. Then we classify each of the frames as 'Normal' class and rest of them we put into certain classes.i.e,'Abuse','Arrest', 'Arson','Assault'.'Burglary','Explosion','Fighting','RoadAccidents','Robbery','Shooting', 'Shoplifting','Stealing','Vandalism'. For the temporal domain, we randomly sample a clip (of $\alpha T \times \alpha$ frames) from the full-length video, and the input to the Slow and Fast pathways are respectively T and αT frames; for the spatial domain, we randomly crop 224 × 224 pixels from a video,

4.1 Dataset

For our project we use DCSASS Dataset available in kaggle.DCSASS dataset consists of 16853 videos. we take 15% as our test data and rest of them we use for train data .In the train data set we again split it into two parts ,for training and validation,in validation also we use total 15% of our train dataset.There are 14 labels in the dataset. Among them 13 is labelled as abnormal and one of the label is labelled as 'Normal'.

4.2 Main Result

In our model we obtained 80.729% accuracy,better than previous model(79.8%),the plots for accuracy 4 and loss 5 are shown

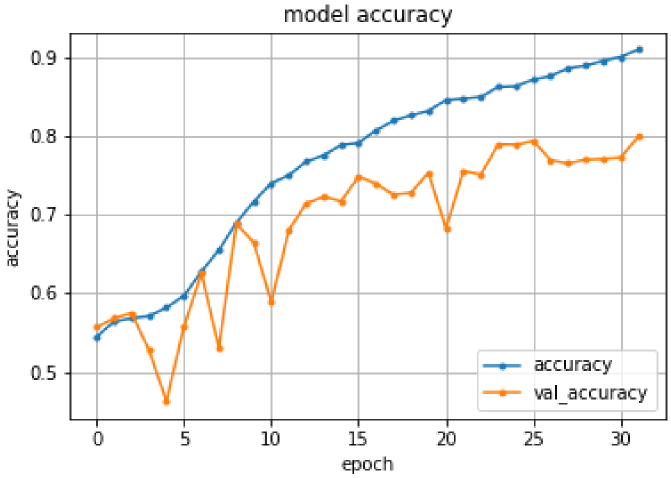

Figure 4: Accuracy obtained for the model

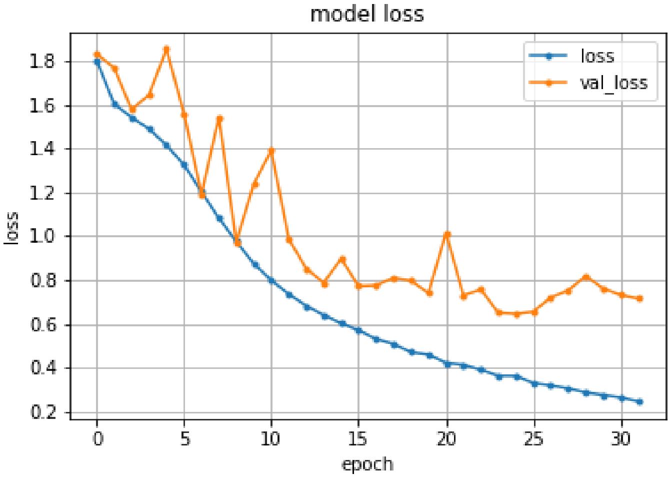

Figure 5: loss obtained for the model

5 Summary

In general, activity recognition considers the complete video observation sequences. Provided the suitable features to the system in real-time. We proposed techniques to analyse the surveillance footage considering 13 cases ranging from abuse, arrest to vandalism along with 14$^{(}th)$ cases i.e. normal. To detect different actions taking place in the surveillance footage we have used a deep neural network model namely SlowFast to deal with this problem. It is fascinating to notice that the our method is able to predict the ongoing activity at a early stage(approximately within 7 seconds). The proposed method is a machine approach to detect real-world unusual activities identification in surveillance videos. The necessity to develop such a security system is increasing number of crimes that are happening everyday.The result of the proposed system will be able to detect whether any anomaly action is taking place or not. Our proposed model has higher acccruacy than the model that were used previously for the action detection in video dataset. We used the 'SlowFast' model which is very new and still work is going on this model. The model was originally developed by Facebook researchers and it was tested on hand action recognition dataset and the dataset which we used is completely new for this model and to attain a high accuracy for the given dataset is great achievement.

6 References

1. Recognizing Human Actions by a Bag of Visual Words
2.ImageNet Classification with Deep Convolutional Neural Networks
3.Two-Stream Convolutional Networks for Action Recognition in Videos ,Karen Simonyan, Andrew Zisserman
4. Temporal Segment Networks for Action Recognition in Videos Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool,2017
5.P-CNN: Pose-based CNN Features for Action Recognition Guilhem Cheron,Ivan Laptev,Cordelia Schmid,2015
6. Christoph Feichtenhofer,Haoqi Fan,Jitendra Malik, SlowFast Networks for Video Recognition,Kaiming He,Facebook AI Research (FAIR)
7. Human Activity Classification Using the 3DCNN Architecture