International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217111

Detecting Criminal Activities of Surveillance Videos using Deep Learning

Mrunal Malekar
Department of Electronics and Telecommunications, Vishwakarma Institute of Technology, Pune, Maharashtra, India

Article Info
Volume 7, Issue 1
Page Number: 188-193
Publication Issue :
January-February-2021

Article History
Accepted : 01 Feb 2021
Published : 05 Feb 2021

ABSTRACT

Videos generated by surveillance cameras inside the ATM were very long. In case, any robbery had taken place inside the ATM; it became time consuming to watch the entire long video. Hence, there was a need to process these surveillance videos by extracting the priority frames from it in which suspicious activities like robbery, murder, kidnap, etc. had taken place. The objective of this paper was to propose algorithm that would generate a detect the suspicious frames from that long surveillance video for the authorities which would consists of priority information. In this paper a novel approach dealing with Convolutional Neural Networks using Deep Learning was used to sample the priority information from the surveillance videos. The priority information was the suspicious activities like robbery, murder, etc. which take place inside the ATM. The results of the CNN model effectively were able to extract suspicious activity frames from a long video and thus extract suspicious frames and create a video from it.

Keywords : Convolutional Neural Network, Deep Learning , Max Pooling

I. INTRODUCTION

This “Robbers fled with entire ATM machine which contained Rs 30 lakh near Nawada metro station in Dwarka, police said on Tuesday”. We see such types of news daily on TV. In a day, at many places robbery takes place at ATM machine which arises the problem of security. To avoid this problem, watchmen are assigned to each ATM machine. CCTV cameras placed inside the ATM record many such videos every day. Recorded videos are too long and automated video analysis approaches [2] have not yet gave the desired results. As the videos are of very long duration it becomes tedious to watch the entire video [1].

There should be a system which extracts only the priority information from the long video and generates a summary. In case of surveillance videos, the priority information is the suspicious activity like robbery and murder. There is therefore a need to sample this priority information from long duration videos. This is where our problem statement comes. Sampling of priority information from these videos lead to video summarization. The primary objective of generating summaries of videos is to be able to extract the priority information from the video and thus reduce the time length of the video. By having captured the priority information in the summarized video, the time needed for browsing through the entire surveillance video is decreased.[3]

There has been a lot of research work in this video summarization domain. Video summarization finds its application in surveillance, news, sports , movie trailer and documentaries related areas. In [4] authors

have presented an approach that uses event detection and clustering which generates static and dynamic summaries. In this approach keyframes were detected using energy difference between frames and then the events were clustered based on their final appearance. And then finally based on the cluster structure the summary was built. In [5] the authors have researched on video summarization in surveillance and news domain where they have presented an approach which deals with spectral clustering methods for temporal segmentation of the video followed by key frame extraction which uses motion analysis of detected frames. In this approach they have assumed that scenes which are full of motion contain priority information than the static ones. In [6] video summarization for sports highlights generation has been discussed in great detail. An algorithm has been used in it which uses textual extraction method. In this the score bar is detected from each frame following which the text on the score is converted into a sentence using OCR. Once the textual information has been obtained, the difference in score and wickets is detected to obtain information about fours, sixes and fall of wickets. In literature [7] the authors have researched about movie summary generation. Important scenes, events, people and objects are included in the movie summary. A motion attention model has been which detects the importance of clips that by detecting degrees of motion in the pictures which depict motion. The video summary generation algorithm used in [7] is based on face detection and multi-feature fusion approaches. In [8] authors have discussed in great length about various video summarization techniques namely feature based, cluster based, event, shot selection and trajectory-based techniques. In [9] the authors have used a method for generating summary of surveillance videos by obtaining spatio-temporal trajectories of static objects along with motion acceleration by using background subtraction, optical flow and pixel clustering. Video summarization approach that uses temporal reduction and tetris-like strategy has been discussed in [9]. Spatio-Temporal redundancy can be used to generate video summaries of surveillance cameras [10]. The primary goal is to reduce the spatio-temporal redundancy from surveillance videos. It is removed by extracting frames which have redundant information and removing those frames from the video. In [11] key frames are extracted from surveillance videos using greedy search algorithm which forms the summary. Techniques like background estimation, feature extraction methods, similarity measurements and object extraction have been used [11]. One of the main issues while extracting priority information is to efficiently extract the detected object. This issue has been tackled in [12] using a technique called as background cut.

In this paper an approach has been presented for the problem statement. We have devised an algorithm that uses Convolutional Neural Networks to sample priority information and generate video summaries. We have made a dataset of frames in which suspicious and non-suspicious movements are seen. Using Keras model and Tensor flow, we have trained our machine with this training dataset. We have used convolutional neural networks to train our dataset.

Then we have tested our model on a testing video. If the suspicious movement is present then that frame is extracted accordingly and a summary of the video is formed which has the priority information frames (robbery, kidnapping, theft, etc.) and this short-summarized video will be mailed immediately to the host using SMTP protocol. Thus, the priority information is effectively sampled from a long surveillance video.

This project is an attempt to generate short summarized video of the long duration surveillance video inside the ATM. The project is implemented using Deep learning in Python.

The structure of this paper is as mentioned : Section 2 deals with the technical specifications of Convolutional neural network. Section 3 explains the working of our CNN model for detection of suspicious activities in videos. Section 4 contains a discussion of the experiments and evaluations of our implemented model for suspicious activity detection and priority information sampling. Conclusion is presented in the final section.

II. METHODOLOGY

A. Generation of Dataset

The dataset used for training consists of many surveillance videos taken inside the ATM. Video is converted to frames using frame rate of 1 frame per second. Keyframes are taken out of those videos and stored in Suspicious-Violent and non-suspicious category folders. This forms the training dataset for the model. There are total 251 frames in the Training dataset. Similarly, Validation dataset is created which also has Suspicious-Violent and non-suspicious category frames. Validation set consists of 109 images. This dataset comprising of Training (70%) and Validation (30%) is used for training of our CNN model.

(a)

(b)

Fig. 1 (a) Suspicious activity frame inside ATM. (b) Non-suspicious activity frame inside ATM

B. Building the convolutional Neural Network

The initial stage of CNN model deals with the convolutive part. The convolution layer is the main part of the CNN. Convolution is used to extract features from the images. In this first stage, input image is passed to several filters/kernel of 3 x3 size to produce convolution maps. In this approach,32 filters have been used. Thus 32 feature maps are produced for one image. The objective of convolution is to extract features from images. The input to the convolutional layer is an input image of dimensions[W1*H1] and a filter preferably [3*3] which extracts certain features from images. The output of the convolution is obtained by applying product between an array of input data which is the image and a two-dimensional array of weights, called filter or kernel. RELU , an activation function is applied in the convolutional layer such as the max(0,x) function, to product elementwise non-linearity.

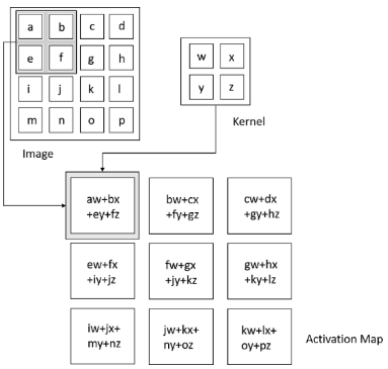

Fig. 2 Convolution Operation

Next comes the max pooling layer which is inserted between successive convolutional layers to down sample the image and to reduce the computation load in the network. The image would contain a lot of pixel values and it is typically easy for the network to learn the features if image size is progressively reduced. In the end the feature vector is flattened as a single vector which forms the input layer to the CNN. Then the fully connected layer is added which has full connection to all the activations in the previous layer. Fully connected layer is applied to generate the final output equal to the number of classes we need. Dense layers are used to predict the classes (output) for our test images. The model is compiled with a categorical cross entropy function, Adam Optimizer and accuracy metric. Once the CNN model is ready it is fit on the training dataset with epoch size of 5 and batch size of 5.

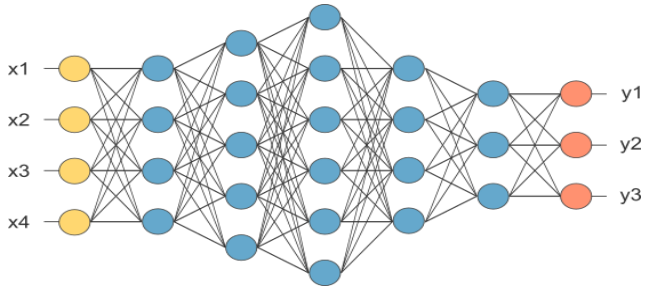

Fig. 3 After pooling, fully connected layers are added

C . Prediction on Testing Video Frames

Once the model is trained it is used for making predictions on the testing videos. The video is converted into frames and the CNN model predicts the category(Suspicious or non-suspicious) for each frame. The frame predicted as Suspicious contain activities/actions like robbery, stealing, etc and those frames are saved in a directory. These suspicious frames contain priority information and they are played at frames per second of 2 to generate a video summary

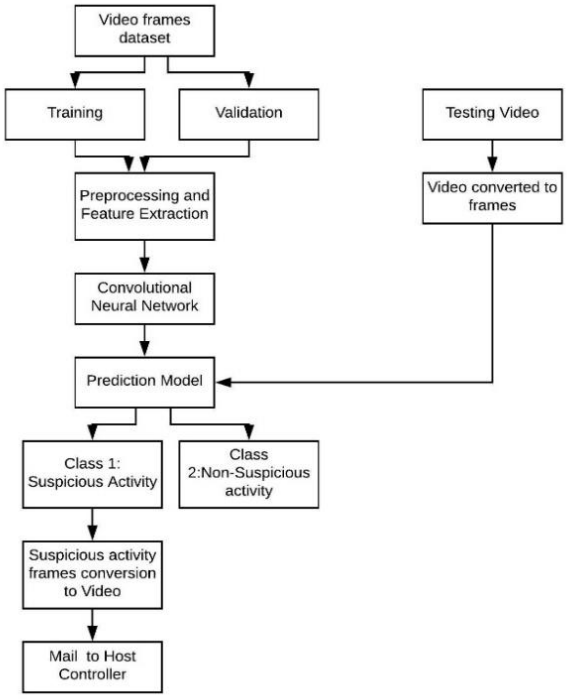

Fig. 4 Flowchart - Process

D . Mailing the suspicious part of video

The generated short video is mailed to the host using the SMTP protocol. Python has smtplib module which can be used to establish SMTP session. SMTP object is defined which can be used to send mails to any host. Figure 5 shows how the summarized video is mailed to a host.

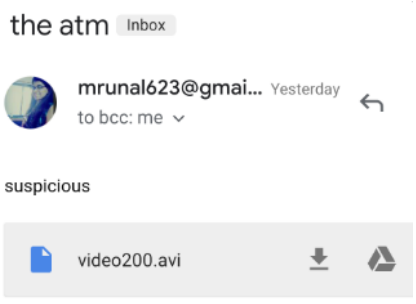

Fig. 5. Video mailed over to Gmail Id using SMTP Protocol

III.EXPERIMENTAL ANALYSIS AND RESULTS

A. CNN MODEL

The CNN model was built using 2 convolutional layer:C1,C2 ,one pooling layer and one fully connected layer :FC3. The output layer consists of 2

neurons hence 2 classes. The model is compiled with a categorical cross entropy function, Adam Optimizer and accuracy metric. The SoftMax activation is normally used in the last layer of CNN because it converts the output of your neural network into probability distribution.

The input of our convolutional layer takes images of size [64×64×3]. On the first Convolutional Layer, we have applied 32 convolution kernels with size of [3x3] by striding it with a stride of 1. The output volume had for size [62×62×3] where height and width are 62 and 62 respectively and depth D is 3. Once the CNN model is ready it is fit on the training dataset with epoch size of 5 and batch size of 5. Once the model is trained, the validation accuracy and loss is found out on the test data.

B. EXPERIMENTS AND RESULTS

The model is trained with stochastic gradient training experiments with different epoch sizes, batch size of 5 and samples per epoch of 1097. We evaluated our model on two testing videos of different scenarios (ATM surveillance) with epoch size of 4 and 5 respectively. We observed an increase in the value of the validation accuracy when the number of epochs was increased. A maximum accuracy of 95.2% was achieved with epoch size of 5.

TABLE 1: VALIDATION ACCURACY WITH EPOCHS

Number of epochs	Validation Accuracy of CNN (%)	Steps
4	91.9	219
5	95.2	219

Below are some experimental graphs for number of epochs equal to 2.

Fig. 5 Validation accuracy vs Time duration

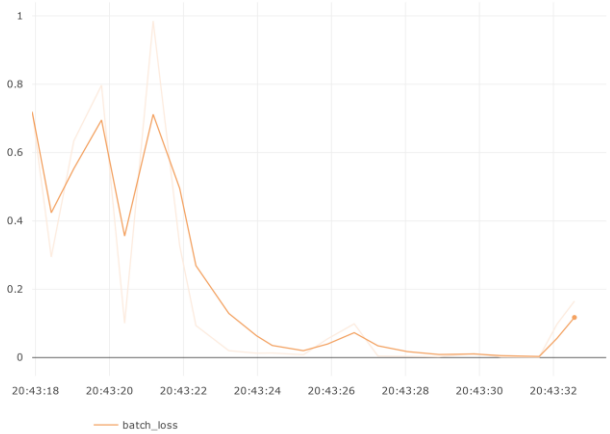

Fig. 6 Batch Loss vs Time duration

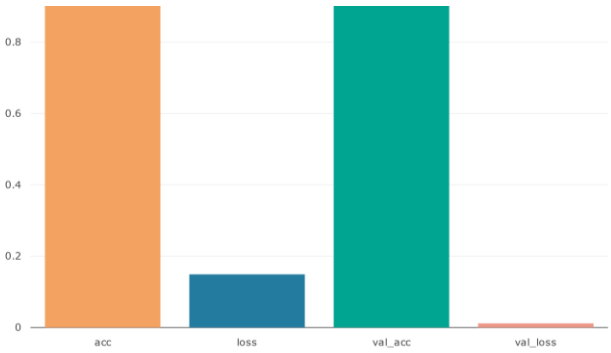

Fig. 7 Training accuracy,Validation accuracy,Training loss,Validation loss

The table below presents in detail as to how many suspicous frames were detected from a given test video.

The test video which has Man being beaten up had 80 suspicious frames out of which during testing the CNN Model predicted 78 frames as suspicious which were true positive. Capturing the detected suspicious

frames a summary video was created with which was of 32 seconds.

TABLE 2 : RESULTS OF EXTRACTION OF SUSPICIOUS FRAMES

Duration of Test Video	Duration of Summary video	Descriptive Features in summary video	Total number of frames in Test video	Total no. of suspicious frames in Test video	Captured no. of priority frames in summary video
1 minute 25 seconds	32 seconds	Man being beaten up in ATM	85	80	78
4 minutes 50 seconds	1 minute 40 seconds	Woman being mobbed inside the ATM	173	133	131
7 minutes	2 minutes 13 seconds	Man stealing cash from ATM	439	190	185

IV. CONCLUSION

The proposed approach aims to reduce the time needed to process and watch the entire long duration CCTV video captured in the ATM. In this paper we demonstrated the use of Convolutional Neural Networks to extract priority information from videos. Using CNN, priority information was extracted from the long video and a short summary was generated which consists of activities like robbery, stealing, etc. The generated summary of the long video is mailed to the host controller via SMTP protocol over the Internet. The proposed approach thus makes the process of browsing surveillance videos less time consuming. Our proposed algorithm gives remarkable accuracy when it is tested on different videos and compresses the long video by huge extent giving us the summarized video as output. As surveillance cameras are used at all places, video synopsis generation will prove to be helpful to the authorities.

V. REFERENCES

[1]. Bo Luo and Xiaoou Tang, "Video Caption Detection And Extraction Using Temporal Information",2003 IEEE Conference .

[2]. Po Kang Lai, Kelvin Moutet,Marc Decombus and Robert Laganiere , "Video Summarization of Surveillance Cameras", 2016 IEEE Conference .

[3]. Uros Damnjanovic, Virginia Fernandez, Ebroul Izquierdo, José María Martinez, "Event Detection and Clustering for Surveillance Video Summarization", 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services.

[4]. Uros Damnjanovic, Virginia Fernandez, Ebroul Izquierdo, José María Martinez , "Event Detection and Clustering for Surveillance Video Summarization",2008 IEEE Conference .

[5]. Muhammad Ehsan Anjum, Syed Farooq Ali, Malik Tahir Hassan, Muhammad Adnan, "Video Summarization Sports Highlights generation", 2013 International IEEE Conference.

[6]. Yuxiang Xie, Jingmeng He, Lili Zhang, Xin,Xiduo Luan, "A Movie Summary Generation System", 2017 IEEE Second International Conference on Data Science in Cyberspace

[7]. Vikas Choudhary, A.K Tiwari, "Surveillance Video Synopsis", 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing

[8]. Sinnu Susan Thoma, Sumana Gupta,Venkatesh K. Subramanian , "Smart Surveillance Based on Video Summarization", 2017 IEEE Region 10 Symposium (TENSYMP).

[9]. W. Zhou, D. Saha, and S. Rangarajan, "A System Architecture to Aggregate Video Surveillance Data in Smart Cities," 2018 in Proc. IEEE GLOBECOM.

Cite this article as :

Mrunal Malekar, "Detecting Criminal Activities of Surveillance Videos using Deep Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 1, pp. 188-193, January-February 2021.

Available at doi : https://doi.org/10.32628/CSEIT217111

Journal URL : https://ijsrcseit.com/CSEIT217111