Real-time CCTV Surveillance Footage Violence Detection using Deep Learning

Rohan Pawar
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
rohan.pawar2020@vitstudent.ac.in

Abstract—**With the rapid growth of surveillance cameras to monitor the human activity demands such system which recognize the violence and suspicious events automatically. In smart buildings, this technology is mandatory. Violence Detection in real time will prevent any mishappenings from taking place by raising alerts to the security guards. Our model also has YOLO object detection inculcated in it in order to recognize any weapons or tools that an attacker may be using against people and the security guards will act accordingly.**

Keywords—Violence Detection, CCTV surveillance technique, Deep Learning, MobileNet, Transfer Learning

I. INTRODUCTION (*HEADING 1*)

Nowadays, there has been a rise in the amount of disruptive and offensive activities that have been happening. Due to this, security has been given principal significance. Public places like shopping centers, avenues, banks, etc. are increasingly being equipped with CCTVs to guarantee the security of individuals. Subsequently, this inconvenience is making a need to computerize this system with high accuracy. Since constant observation of these surveillance cameras by humans is a near-impossible task. It requires workforces and their constant attention to judge if the captured activities are anomalous or suspicious. Hence, this drawback is creating a need to automate this process with high accuracy. Moreover, there is a need to display which frame and which parts of the recording contain the uncommon activity which helps the quicker judgment of that unordinary action being unusual or suspicious. Therefore, to reduce the wastage of time and labor, we are utilizing deep learning algorithms for automated violence detection system. Its goal is to automatically identify signs of aggression and violence in real- time, which filters out irregularities from normal patterns. We intend to utilize different Deep Learning models (CNN) to identify and classify levels of high movement in the frame. From there, we can raise a detection alert for the situation of a threat, indicating the suspicious activities at an instance of time.

II. WORKING

A. Objective

The key objective of CCTV footage violence detection system is to identify suspicious events and send an alert message to the authorized person. So that he can immediately take the action on them without affecting anyone.

B. Proposed Methodology

MobileNet is a streamlined architecture that uses depth wise separable convolutions to construct lightweight deep convolutional neural networks and provides an efficient model for mobile and embedded vision applications. Transfer learning will be applied to this to train the model for Violent activities detection as well as object detection.

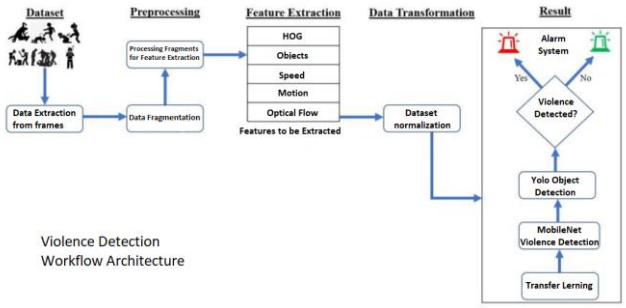

Figure.1- Violence Detection Workflow

III. BACKGROUND STUDY

A. Literature Review

In today's modern world of 24/7 surveillance, vision sensory data are widely used to monitor activities automatically and report them to connected departments for counter actions. Automated surveillance is a major problem and concern of computer vision experts, where deep models have achieved tremendously precise results in many computer vision problems, ranging from object detection to complex multiple activities prediction and perception. Video data (VD) based on video data is applicable to a large number of real-world scenarios. It is evident that the occurrence of violence is extremely rare when compared to the normal pattern of continuous video surveillance. [3].

In paper [4], the proposed violence detection system uses the Histogram of Oriented Gradients (HOG) method in big data Spark framework for better extraction of features in real time. Since the input to the system is a sequential video streaming, Long-Short Term Memory (LSTM) is used for recognizing the violence actions. The case study has been done on developing a violence detection system for a football

stadium, where there is a possible occurrence of violence between the players as well as the audience.

Deep learning-based models have achieved encouraging results for fight activity recognition on benchmark data sets such as hockey and movies[1]. However, these models have limitations in learning discriminating features for violence activity classification with abrupt camera motion. This research work investigated deep representation models using transfer learning for handling the issue of abrupt camera motion. Consequently, a novel deep multi-net (DMN) architecture based on AlexNet and GoogleNet is proposed for violence detection in videos. AlexNet and GoogleNet are top-ranked pre-trained models for image classification with distinct pre-learnt potential features. The fusion of these models can yield superior performance. The proposed DMN unleashed the integrated potential by concurrently coalescing both networks. The results confirmed that DMN outperformed state-of-the-art methods by learning finest discriminating features and achieved 99.82% and 100% accuracy on hockey and movies data sets, respectively. Moreover, DMN has faster learning capability i.e. 1.33 and 2.28 times faster than AlexNet and GoogleNet[1].

Detecting any activity from video requires some elements. The video feature is one of the basic elements for detecting any activity from that video. The detecting process and methodology directly depend on the video features that are extracted from a video. Those features are used to analyze the pattern of the activity. In a video that contains a fight scene, the movement of the objects is faster than the normal video. Analyzing those extracted features classifies the activity of the video.[2]

The proposed architecture uses convolution neural networks as the spatial feature extractor followed by an LSTM network to perform sequence prediction on the feature vectors. For the spatial feature extraction, we have employed a transfer learning approach with CNN[5]. It is observed the models trained on the benchmark datasets do not work accurately with the real- time CCTV footage. It is mainly due to the fact that the videos are unrealistic and do not aptly depict the real world scenarios. These videos differ a lot from the actual CCTV ones in terms of the camera angle too[6].

A new architectural model has been proposed in paper [7] which uses convolutional long short-term Memory (convLSTM). This model has been constructed using a series of layers (convolutional and pooling) to extract features. A total of 256 ˙letters has been used with Recti˝ed Linear Unit (ReLU) being the activation function. Difference between adjoining frames in the input layer has been taken to identify changes in videos. The model has been trained and tested using three different open datasets namely movie dataset, hockey fight dataset and violence flow dataset. Several pre-processing techniques has been applied on the datasets before training. Using the same data, an accuracy of 94.6% was achieved by LSTM while an appreciably higher accuracy of 97.1% has been accomplished using convLSTM[7].

In continuation of previous work in paper [7],another research based their approach on a completely new algorithm known as the Motion Weber Local Descriptor (MoWLD). This is capable of finding both temporal and spatial information from the interest points of the videos. The researchers have rebuilt the WLD histograms and oriented them by collecting the Wild histograms from the adjacent regions. Multi-scale optical flow has been used to adopt WLD features while reduction of data dimension has been obtained using the Kernel Density Estimation (KDE). Furthermore, max-pooling technique has been applied to get more compact features representation. The model has been trained and tested using three different sources of datasets. The BEHAVE dataset has obtained the highest accuracy rate of 94.9% followed by the hockey got dataset Violence Detection with Deep Learning Approaches with an appreciable accuracy of 91.9% and finally the crowd violence dataset with a satisfactory accuracy rate of 89.78%[7].

Some authors focused on the real-time violent detection problems, worked on that and obtained some fascinating results. Surveillance video cameras are used to get real-time unique data. Depending on the changes of magnitudes of row vector, they made some statistic by using descriptor named violent flow. Then, they collected the statistics from their dataset for short frames. By using this technique, they got accuracy of 82.9%. Other researchers of same fields tried to build a model in a different way. To classify important context, like violence, they used CENTRIST-based features. The whole process starts with pre-processing followed by feature extraction. After that, for classification, they normalized the data and then applied feature reduction. They used two different datasets and those are violent flow dataset and Hockey Fights Dataset. From the first dataset, they obtained an accuracy of 91.46% and from the second dataset they obtained an accuracy of 92.79%.

The paper "Detecting Violence in Video Based on Deep Features Fusion Technique" proposes a method for detecting violence in videos using a fusion technique that combines the features extracted from different deep learning models. The proposed approach aims to address the limitations of previous methods that rely on a single feature extraction technique and suffer from poor performance when dealing with complex and diverse visual content.[8]

The proposed method consists of three stages: feature extraction, feature fusion, and classification. In the first stage, the authors use three pre-trained deep learning models (VGG16, ResNet50, and InceptionV3) to extract features from video frames. In the feature fusion stage, the extracted features are combined using a weighted average fusion technique. Finally, in the classification stage, the fused features are fed into a support vector machine (SVM) classifier to detect violence in videos.[8]

In conclusion, the paper presents a novel approach for detecting violence in videos using a fusion technique that combines deep features extracted from different models. The proposed method achieved better performance than previous methods, demonstrating the potential of using multiple deep learning models for video analysis tasks.[8]

In the paper [9], the authors present a deep learning-based approach for detecting criminal activities in surveillance videos. The proposed system aims to assist law enforcement agencies in identifying criminal behaviour and potentially preventing crime before it occurs.

The proposed system consists of three main stages: pre-processing, feature extraction, and classification. In the pre-processing stage, the authors use a Gaussian filter to remove

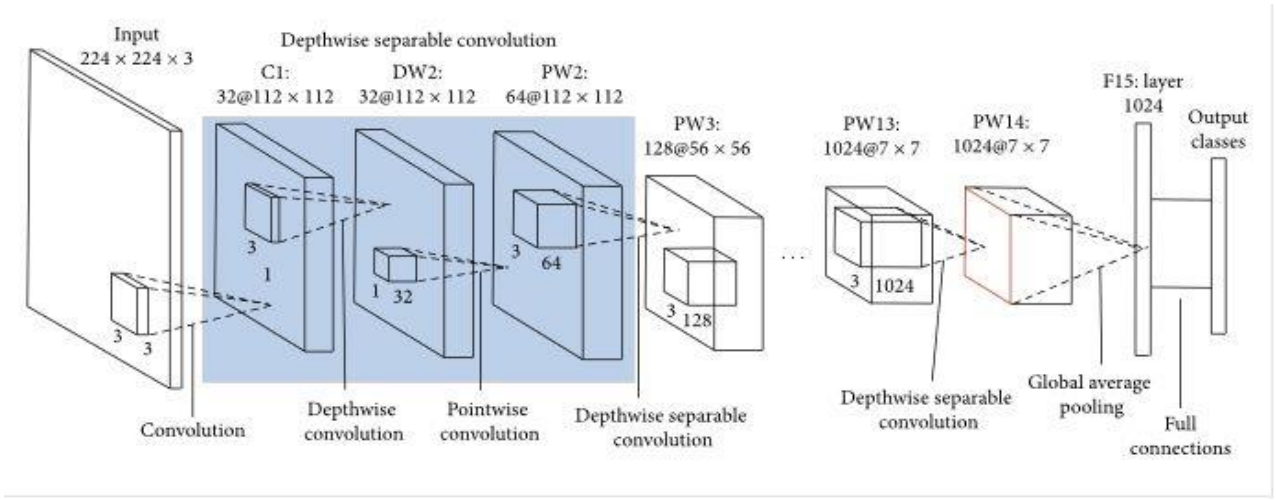

Figure 2- MobileNet Architecture

noise from the video frames. In the feature extraction stage, they use the convolutional neural network (CNN) architecture ResNet50 to extract features from the video frames. Finally, in the classification stage, the authors use a long short-term memory (LSTM) network to classify the extracted features into normal or criminal activity.[9]

In continuation of the previous context of paper [9], the proposed approach was evaluated on a dataset of surveillance videos that contained criminal and non-criminal activities. The results showed that the proposed system achieved an accuracy of 92.8%, outperforming previous methods that relied on handcrafted features. The authors also conducted an ablation study to evaluate the contribution of each stage to the performance of the proposed system. In conclusion, the paper presents a deep learning-based approach for detecting criminal activities in surveillance videos. The proposed system achieved high accuracy and demonstrated the potential of using deep learning for crime prevention and law enforcement.

In paper [10], authors propose an approach for violence detection in videos using pretrained deep learning models with different architectures. The authors argue that using pretrained models can reduce the need for large training datasets and improve the performance of violence detection systems.The proposed approach consists of three stages: feature extraction, feature fusion, and classification. In the feature extraction stage, the authors use three pretrained deep learning models, VGG16, InceptionV3, and MobileNetV2, to extract features from video frames. In the feature fusion stage, the extracted features are concatenated and fed into a dense layer to generate a fused feature representation. Finally, in the classification stage, the fused features are fed into a support vector machine (SVM) classifier to detect violence in videos.

B. About Dataset

Dataset has been taken from Kaggle.com. It is called real-life-violence-situation-dataset by Mohamed Mustaf. It contains 2000 videos in total (1000 violence and 1000 Non-violence). Due to hardware limitations, only 700 videos were taken to train the model.

C. Model Architecture

CNN- CNN stands for Convolutional Neural Network, which is a type of neural network commonly used in deep learning tasks such as image recognition, object detection, and segmentation. CNNs are designed o automatically learn spatial hierarchies of features from raw input data. The main building blocks of a CNN are convolutional layers, pooling layers, and fully connected layers. After several convolutional and pooling layers, the output is passed through one or more fully connected layers, which perform classification or regression based on the learned features. These layers are similar to those in a traditional feedforward neural network, where each neuron in the layer is connected to all the neurons in the previous layer.

- Convolutional Layers-

  Convolutional layers use filters or kernels to perform a convolution operation on the input data and extract features such as edges, corners, and textures.For advanced feature extraction, some people also prefer to use HOGs, SIFT and SURF methods.
- Pooling Layers-

  Pooling layers reduce the spatial dimensions of the output by downsampling the input, which helps to reduce the number of parameters and computational cost of the network.
- Fully Connected Layers/Dense Layers-

  Fully connected layers are used to make predictions based on the learned features and are typically placed at the end of the network. CNNs are trained using backpropagation with gradient descent to minimize a loss function and optimize the network's parameters.

MobileNet- MobileNet is a convolutional neural network architecture that is designed for efficient and lightweight image classification tasks on mobile and embedded devices. It was first introduced in a 2017 research paper by Google researchers Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNet is designed to be highly efficient by using depthwise separable

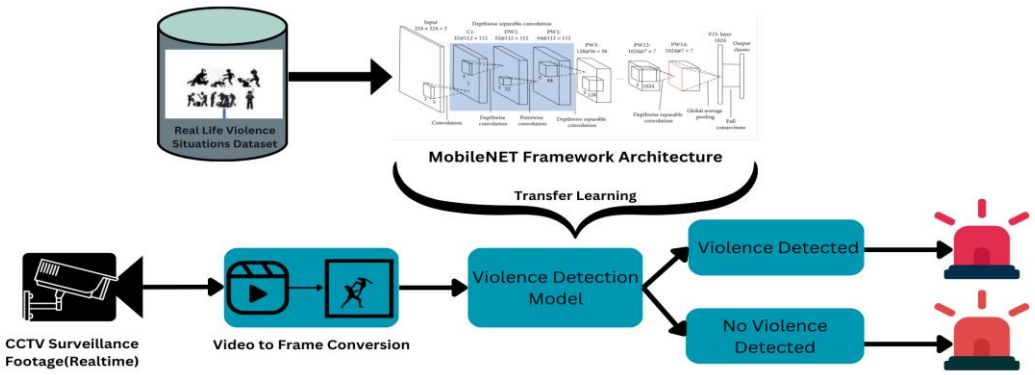

Figure 3- Violence Detection Model Architecture

convolutions, which are a type of convolutional operation that splits a standard convolution into two separate operations: a depthwise convolution and a pointwise convolution. In the depthwise convolution, a single filter is applied to each input channel separately, while in the pointwise convolution, a 1x1 convolution is applied to combine the output of the depthwise convolution.

By using depthwise separable convolutions, MobileNet is able to significantly reduce the number of parameters and computations required for violence and non-violence classification, while still maintaining a high level of accuracy.

IV. RESULTS AND DISCUSSION

Because of limited space allotted on Google Collab, I was only able to train the model with 700 videos (350 violence, 350 non-violence). So, the accuracy of the model is 89% but it can be 95-99%. If we are able to train it with 2000 videos.

Best Epoch	9
Accuracy on train	0.888836801
Accuracy on test	0.87644875
Loss on train	0.26514703
Loss on test	0.284462363

Table 1- Evaluation Matrix

The training loss indicates how well the model is fitting the training data, while the validation loss indicates how well the model fits new data. The training and validation loss and training and validation accuracy curves are given below:

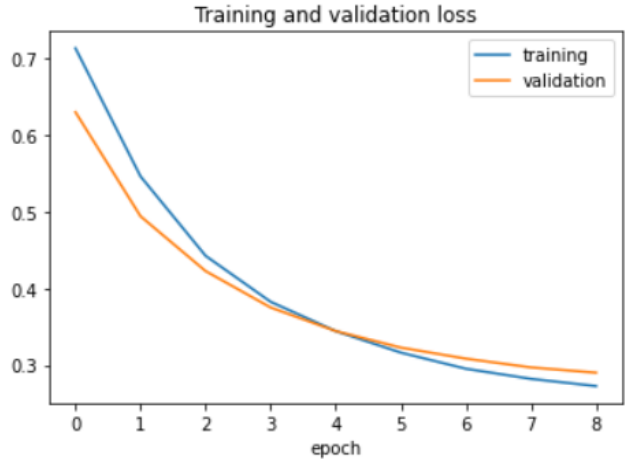

Figure 4- Training and Validation Loss Curve

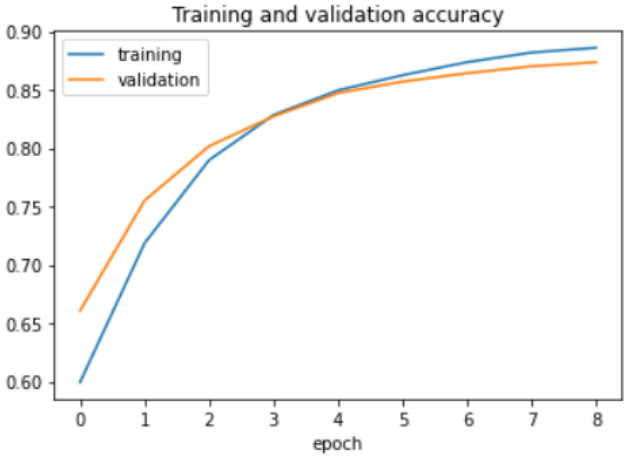

Figure 5- Training and Validation Accuracy Curve

Confusion Matrix below depicts the different outcomes of the prediction and results of a classification problem and helps visualize its outcomes. The confusion matrix of the violence detection model prediction is given below:

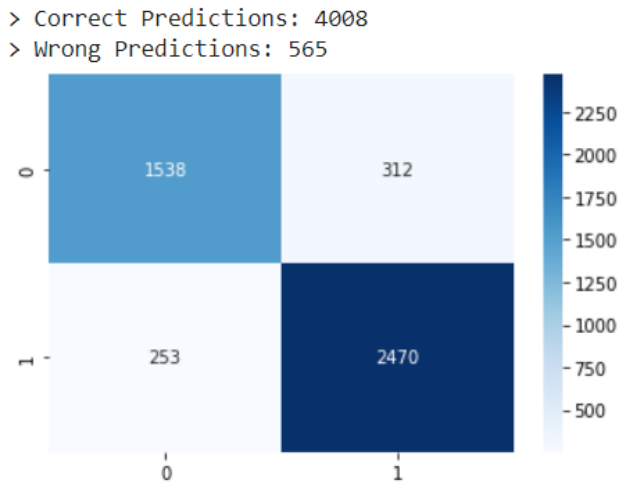

Figure 6- Confusion Matrix

Confusion Matrix		
	Actually Positive	**Actually Negative**
Predicted Positive	True Positive = 1538	False Positive = 312
Predicted Negative	False Negative = 253	True Negative = 2470

Table 2- Confusion Matrix Values

Clasiffication Report				
	Precision	**Recall**	**f1-score**	**Support**
NonViolence	0.86	0.83	0.84	1850
Violence	0.89	0.91	0.9	2723
Accuracy			0.9	4573
Macro avg	0.87	0.87	0.87	4573
Weighted avg	0.88	0.88	0.88	4573

Table 3 – Classification Report

V. CONCLUSION

Artificial intelligence is crucial for spotting violence in video footage. Violence detection is still an issue even though it appears to be complicated at the moment.

With this paper, we aimed to demonstrate how violence detection is possible and how it can be easily implemented using the most basic techniques currently available, all due to the ongoing developments in deep learning and AI. These models have learned how to extract dynamic and practical features from everyday images and use them as a jumping-off point for learning new tasks and also alerts when violence is detected and produce the beep sound, so that the nearest authority can take the action according to the situation. After being trained on more than a million images, these networks can categorize images into violent and non violent classes. Transfer learning was used with pre-trained networks because it is frequently simpler and faster than creating or training a network from inception.

REFERENCES

[1] Aqib Mumtaz, Allah Bux Sargano, Zulfiqar Habib, “Fast Learning Through Deep Multi-Net CNN Model For Violence Recognition In Video Surveillance”, The Computer Journal, Volume 65, Issue 3, March 2022, Pages 457–472

[2] “State-of-the-Art Violence Detection Techniques: A review”, DOI: 10.9734/AJRCOS/2022/v13i130305

[3] An Overview of Violence Detection Techniques: Current Challenges and Future Directions, Nadia Mumtaz, Naveed Ejaz, Shabana Habib, Prayag Tiwari Shahab S. Band

[4] Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM, Dinesh Jackson Samuel R., Fenil E, Jeeva S

[5] A fully integrated violence detection system using CNN and LSTM, Sarthak Sharma, B Sudharsan, Saamaja Naraharisetti, Vimarsh Trehan, Kayalvizhi Jayavel, ISSN: 2088-8708, DOI: 10.11591/ijece.v11i4.pp3374-3380

[6] Human Violence Detection Using Deep Learning Techniques, S A Arun Akash, Conf. Ser. 2318 012003

[7] Violence Detection by Pretrained Modules with Di®erentDeep Learning Approaches, Shakil Ahmed Sumon, Raihan Goni, Niyaz Bin Hashem ,Tanzil Shahria and Rashedur M. Rahman

[8] Detecting Violence in Video Based on Deep Features Fusion Technique, Heyam M. Bin Jahlan and Lamiaa A. Elrefaei

[9] Deep Neural Network for Gender-Based Violence Detection, Carlos M. Castorena, Itzel M. Abundez, Roberto Alejo, Everardo E. Granda-Gutiérrez, Eréndira Rendón and Octavio Villegas

[10] Design of a Deep Learning-based Detection System for Criminal Activities, Verlyn V. Nojor; Jarod Augustus C. Austria; Adrian A.