REAL TIME VIOLENCE ALERT SYSTEM

PROJECT REPORT

Submitted to the APJ Abdul Kalam Technological University

in partial fulfillment of requirements for the award of degree

Bachelor of Technology

in

Computer Science and Engineering

by

BEN SAM SABU*(MBT18CS036)*
C J PETER*(MBT18CS039)*
GOVIND B CHANDRAN*(MBT18CS052)*
MANNU THOMAS*(MBT18CS071)*

Department of Computer Science and Engineering
Mar Baselios College of Engineering and Technology
(Autonomous)
Mar Ivanios Vidya Nagar, Nalanchira
Thiruvananthapuram- 695015

May 2022

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAR BASELIOS COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
MAR IVANIOS VIDYA NAGAR, NALANCHIRA
THIRUVANANTHAPURAM-695015

CERTIFICATE

This is to certify that the report entitled **Real Time Violence Alert System** submitted by **Ben Sam Sabu** (MBT18CS036), **C J Peter** (MBT18CS039), **Govind B Chandran** (MBT18CS052), **Mannu Thomas** (MBT18CS071), to the APJ Abdul Kalam Technological University in partial fulfillment of the B.Tech. degree in Computer Science and Engineering is a bonafide record of the project preliminary work carried out by them under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Ms. Jayalekshmi J
(Project Guide)
Assistant Professor
Department of CSE

Dr. Shini Renjith
(Project Co-ordinator)
Assistant Professor
Department of CSE

Dr. Tessy Mathew
(Head of the Department)
Associate Professor
Department of CSE

Place: Thiruvananthapuram

Date:

ACKNOWLEDGEMENT

First of all, we thank God almighty for His blessings that enabled us to do this work. We convey our gratitude to our Bursar, **Rev. Fr. John Varghese Palanikunnathil** for presenting this opportunity. We express our sincere gratitude to our Principal, **Dr. Abraham T Mathew** for giving us this opportunity to do this project. We also extend our gratitude to the Head of the Department, **Dr. Tessy Mathew** for all the support. We extend our thanks to the project coordinator, **Dr. Shini Renjith** for arranging the everything in a successful manner. We are immensely thankful to our project guide, **Ms. Jayalekshmi J**, for all her guidance and support. We also thank all the staff members of Computer Science Department for their help. Last, but not the least, we thank all our friends and family members for their care and encouragement.

Ben Sam Sabu
C J Peter
Govind B Chandran
Mannu Thomas

ABSTRACT

Hearing about the violent activities that occur on a daily basis around the world is quite overwhelming. Personal safety and social stability are seriously threatened by the violent activities. A variety of methods have been tried to curb the violent activities which includes installing of surveillance systems. It will be of great significance if the surveillance systems can automatically detect violent activities and give warning or alert signals.

The whole system can be implemented with a sequence of procedures. Firstly, the system has to identify the presence of human beings in a video frame. Then, the frames which are predicted to contain violent activities has to be extracted. The irrelevant frames are to be dropped at this stage. Finally, the trained model detects violent behaviour and these frames are separately saved as images. These images are enhanced to detect faces of people involved in the activity, if possible. The enhanced images along with other necessary details such as time and location is sent as an alert to the concerned authority.

The proposed method is a deep learning based automatic detection approach that uses Convolutional Neural Network to detect violence present in a video. But, the disadvantage of using just CNN is that, it requires a lot of time for computation and is less accurate. Hence, a pre-trained model, MobileNet, which provides higher accuracy and acts as a starting point for the building of the entire model. An alert message is given to the concerned authorities using telegram application.

Table of Contents

List of Figures

Chapter 1

INTRODUCTION

Violent behavior in public places is an issue that has to be addressed. Communities are also eroded by violence, which reduces productivity, lowers property values, and disrupts social services. Across the world, violence is a severe public health issue. It affects people at various phases of life, from infants to the elderly.

Recognizing violence is challenging since it must be done on real-time videos captured by a large number of surveillance cameras at any time and in any location. It should be able to make reliable real-time detection and alert corresponding authorities as soon as violent activities occur.

Public video surveillance systems are widespread around the world and can provide accurate and complete information in many security applications. However, having to watch videos for hours reduces your ability to make quick decisions. Video surveillance is essential to prevent crime and violence. In this regard, several studies have been published on the automatic detection of scenes of violence in video. This is so that authorities do not have to watch videos for hours to identify events that only last a few seconds. Recent studies have highlighted the accuracy of deep learning approaches to violence detection.

Indeed, deep learning methods have proven effective for extracting spatiotemporal features from videos. A function that represents the motion information contained in a series of frames in addition to the spatial information contained in a single frame. In this work we will be discussing about the implementation of a Real-Time violence alert system using MobileNetv2. The frames obtained as output from the model is enhanced. Then these frames along with time and location of the recorded incident are send to the nearby police station as an alert via the alert module of the proposed system.

The remaining report is categorized as follows, Section II presents related studies and their detailed comparisons. Section III discusses the selected approach in a detailed manner explaining the proposed architecture and the dataset, whereas section IV provides details of experimentation and discusses the evaluation of the approach. Finally, section V concludes the report and discusses future innovations that could be brought into our project.

Chapter 2

LITERATURE REVIEW

Recently proposed methods for violence detection can be roughly classified into three categories visual based approach, audio-based approach and hybrid approach

Visual Based Approach : Visual information is retrieved and represented as relevant features in this approach. Local features and global features are two types of features. Position, velocity, form, and color are examples of local features, while average speed, region occupancy, relative positional fluctuations, and the interactions between objects and backdrop are examples of global features.

Audio Based Approach : Audio data is used to classify violence in this approach. It uses a hierarchical technique based on Gaussian mixture models and Hidden Markov models to distinguish gunshots, explosions, and automobile braking in audio.

Hybrid Approach : The emphasis in the hybrid method is on merging visual and audio characteristics. Some techniques recognize violent incidents in videos utilizing flame and blood detection and recording the degree of motion, as well as the typical sounds of violent occurrences. The CASSANDRA system, detects aggression in surveillance videos using motion features associated with articulation in video and scream-like cues in audio.

2.1 SPATIO-TEMPORAL MODELLING METHOD WITH 2D CNN

This approach mainly consists of 3 modules, Motion Saliency Map(MSM),2D CNN's with frame-grouping, Temporal Squeeze and Excitation Block. MSM can effectively highlight moving objects, to magnify prominent regions connected to violence in videos. Attention maps are generated by dilation of the motion boundaries, which can represent the boundaries of moving objects. Then, using the 2D CNN backbones, convert each three-channel frame into a single-channel frame in the middle of the forward path and group three consecutive frames to learn Spatio-temporal representation in a video. This technique is known as frame-grouping. Temporal Squeeze and Excitation Module is a module that is used to classify a target event with less computational cost.

This approach cannot be used in moving cameras as a sequence of frames in a fixed position is necessary to recognize violence. The computational cost of this approach is high as it requires lots of calculations using 3 different modules with a predefined set of tasks.

2.2 SPACE TIME INTEREST POINT (STIP)

Interest points are detected using the Space Time Interest Point (STIP) approach, which analyses spatial and temporal differences. A 3 dimensional volume is extracted around each space time interest point; the volume depicts how a 2D image segment evolves over time. The scale of the identified characteristics determines the size of the 3D volume. The discovered interest points have a significant degree of intensity variation in space and non-constant motion in time. These important points can be found on a number of spatial and temporal scales.A Histogram of Oriented Gradients is used to identify the appearance of the 3D cube, and a Histogram of Oriented Flows is used to define the motion. These traits may be utilised to recognise motion events with high accuracy, and they are resistant to changes in pattern scale, frequency, and velocity.

Despite promising results with a high accuracy rate for this task, the computational cost of extracting such features is extremely high for practical applications, such as surveillance and media rating systems.

2.3 VIOLENCE DETECTION USING LOW-LEVEL FEATURES

The motion region is segmented according to the distribution of the optical flow field. In the motion domain, two types of low-level traits are extracted to express the appearance and dynamics of aggressive behavior. The low-level functions are the Oriented Gradient Local Histogram (LHOG) descriptor extracted from the RGB image and the Optical Flow Local Histogram (LHOF) descriptor extracted from the optical flow image. The extracted features are encoded using a Bag of Words (BoW) model to remove redundant information and obtain a constant-length vector for each video clip. Finally, we use a support vector machine (SVM) to classify the video layer vectors.

Detection of violence in a crowded scene was a challenge due to the serious occlusion and moving crowd. It has low detection rate and high false alarm.

2.4 SPATIO-TEMPORAL FEATURES WITH 3D CNN

For this purpose, the framework of triple staged end to end deep learning violence detection is proposed. Firstly in the surveillance video streams, persons are detected using lightweight CNN model to overcome and reduce the huge processing of unusable frames. Secondly, an order of 16 frames with detected individuals is passed to 3D CNN, where the spatiotemporal features of these sequences are extracted and fed to the Softmax classifier. Then, the 3D CNN model is optimized using a neural networks optimization toolkit and an open visual inference developed by Intel. Trained model is converted into an intermediate illustration and changes it for execution at the end platform for the final detection of violence

This process is really computationally expensive and depends on a lot of parameters, hence is not suitable for real-time environments and for daily use.

2.5 SENSOR-NETWORK APPROACH

Deep neural network images are extracted from an MPEG encoded video input stream. Then the DNN is trained to recognise frames which correspond to violent behavior.

Similar to other methods, lots of computational power is required which is not suitable for real -time usage / environments or for daily usage.

Chapter 3

METHODOLOGY

3.1 PROBLEM STATEMENT

The aim is to develop a real time surveillance system that can recognize violence and give alert to notify the concerned authorities.

3.2 USE CASE DIAGRAM

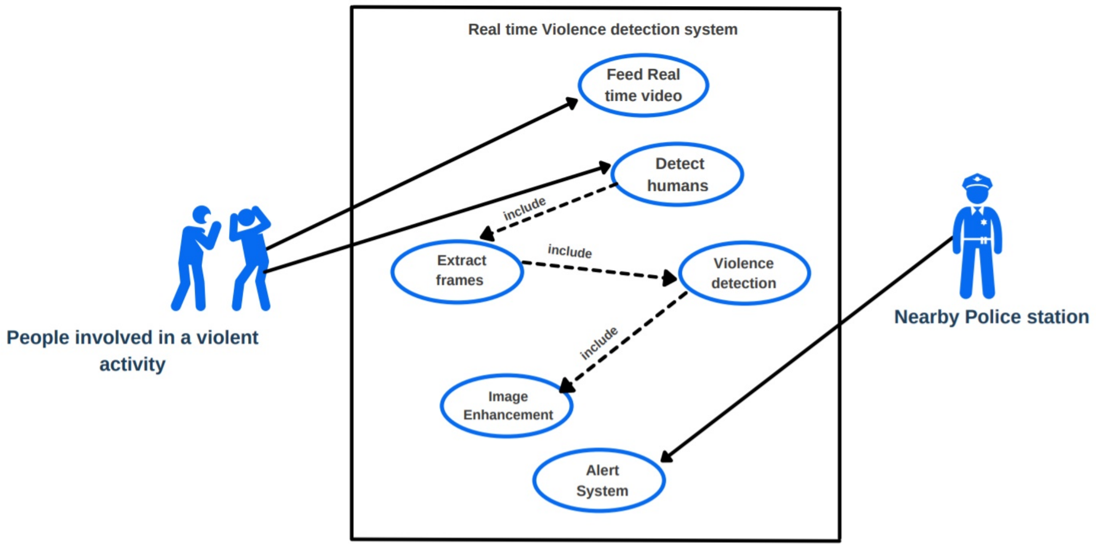


Figure 3.1: Use Case diagram

The rectangle helps to define the scope of the proposed architecture. Anything happens within the rectangle happens within the system. There are 2 actors in this scenario, People involved in a violent activity and a nearby police station. People involved in a violent activity are primary actors as their actions will be detected by a real-time system. Use cases are represented by oval shape and they represent an action that accomplishes some sort of task within the system. Whenever a real-time video is given as input, the system will try to detect humans. Every time humans are detected the given included use cases are executed as well. They are Frames Extraction, Violence Detection and Image Enhancement. After these three steps alert system will send an alert to a nearby police station.

3.3 ARCHITECTURE

Footage from the surveillance camera is broken down into frames. The frames are given as input to MobileNet v2 classifier for detecting violent activities in the given sequence of input frames. If no violent activity is recognized the respective frames are discarded. The violence detected frame is obtained and it is enhanced for better clarity. That frame, along with the location are sent to the nearest authorities using Telegram bot.

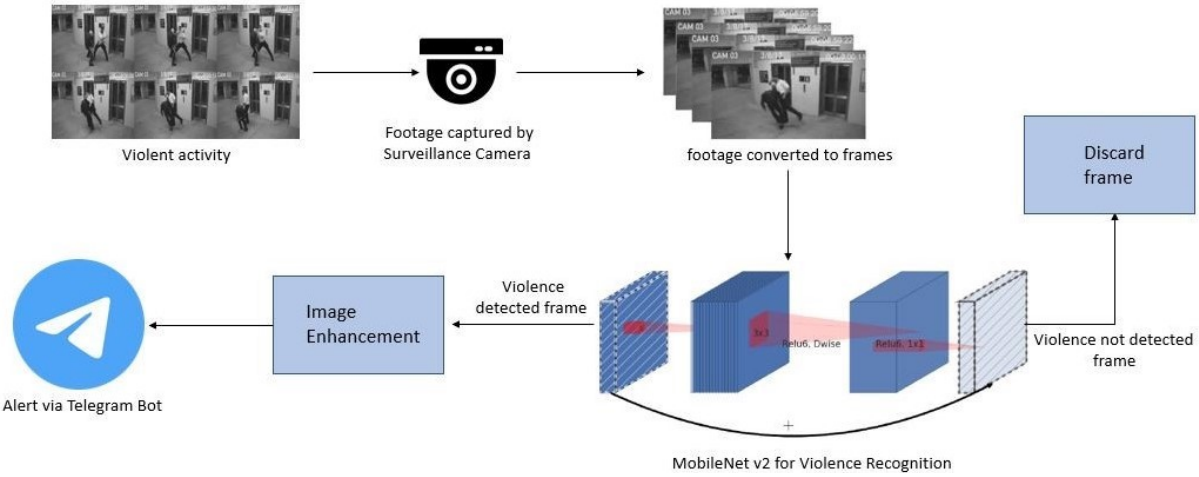

Violent activity
Footage captured by Surveillance Camera
footage converted to frames
Discard frame
Violence detected frame
Image Enhancement
3x3
Relu6, Dwise
Relu6, 1x1
+
Violence not detected frame
Alert via Telegram Bot
MobileNet v2 for Violence Recognition

Figure 3.2: High level architecture diagram

Dataset: The dataset contains 1000 video clips which belongs to two classes, violence and non-violence respectively. The average duration of the video clips is 5 seconds and majority of those videos are from CCTV footages. For training, 350 videos each from the violent and non-violent classes are taken at each epoch.

Figure 3.3: Video clips from the violence dataset

MobileNet V2: The MobileNet architecture is primarily based on depth wise separable convolution, in which factors a traditional convolution into a depth wise convolution followed by a pointwise convolution.

The module presents a residual cell (has a residual/identity connection) with stride of 1, and a resizing cell with a stride of 2. From Figure 3.3, "conv" is a normal convolution, "dwese" is a depth wise separable convolution, "Relu6" is a ReLu activation function with a magnitude limitation, and "Linear" is the use of the linear function.

The main strategies introduced in MobileNetV2 were linear bottleneck and inverted residual blocks. In the linear bottleneck layer, the channel dimension of input is expanded to reduce the risk of information loss by nonlinear functions such as ReLU. It stems from the fact that information lost in some channels might be preserved in other channels. The inverted residual block has a ("narrow" -"wide"-""narrow") structure in the channel dimension whereas a conventional residual block has a ("wide" - "narrow"- "wide") one. Since skip connections are between narrow layers instead of wider ones, the memory footprint can be reduced.

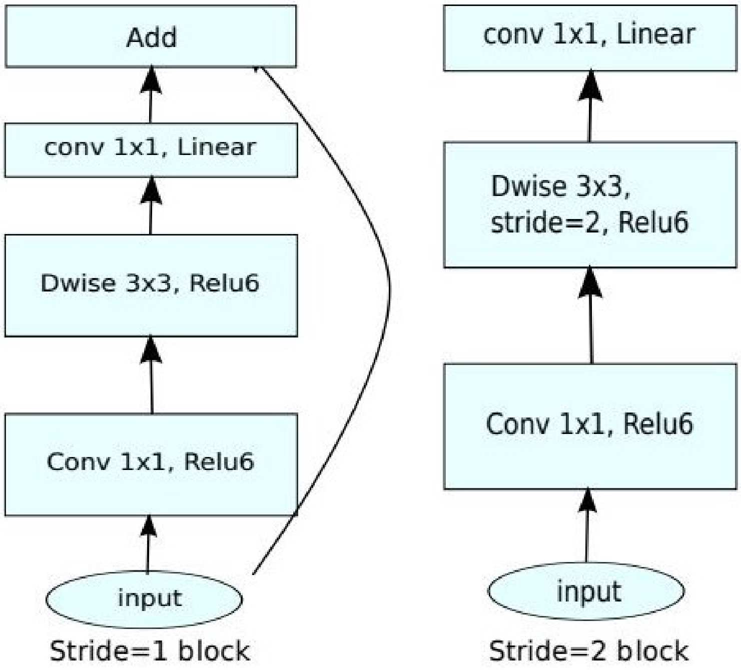


Figure 3.4: MobileNet v2 Architecture

Image Enhancement: Image Enhancement is performed on the frames that are obtained as output. This is performed using the inbuilt functions provided by the Python Imaging Library(PIL). PIL offers extensive file format support, efficient presentation, and fairly powerful image processing capabilities. The Core Image Library is designed to provide quick access to data stored in several major pixel formats. It provides a solid foundation for common image processing tools. The brightness and colour of the obtained output frames is increased by a factor of 2.

Alert Module: The alert module sends alert message to the specified authority. Figure 3.4 describes the architecture of the implemented alert system. When a frame is detected true for violence, the system initialises a counter variable to one. Then it checks the subsequent 30 frames, whether if they too have violence detected true. The counter is incremented at each consecutive frame that is true for violence. If a frame is false for violence, the counter variable is set to 0 and starts checking the consecutive frame respectively checking whether violence is recognized. On the other hand, if the violence is detected true for the 30 consecutive frames, the current time is obtained using an inbuilt python function and an alert is sent to a Telegram group that consists officials of higher authorities. The Alert message comprises of an image of the detected violent activity, current timestamp and the location where the camera is placed.

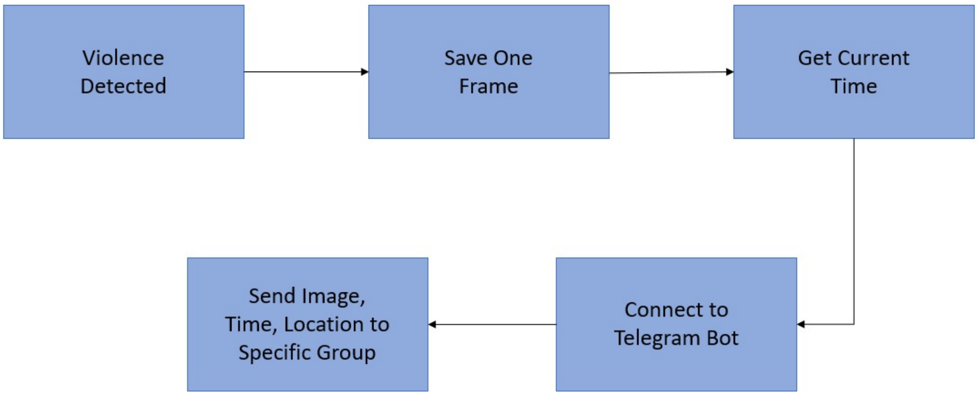
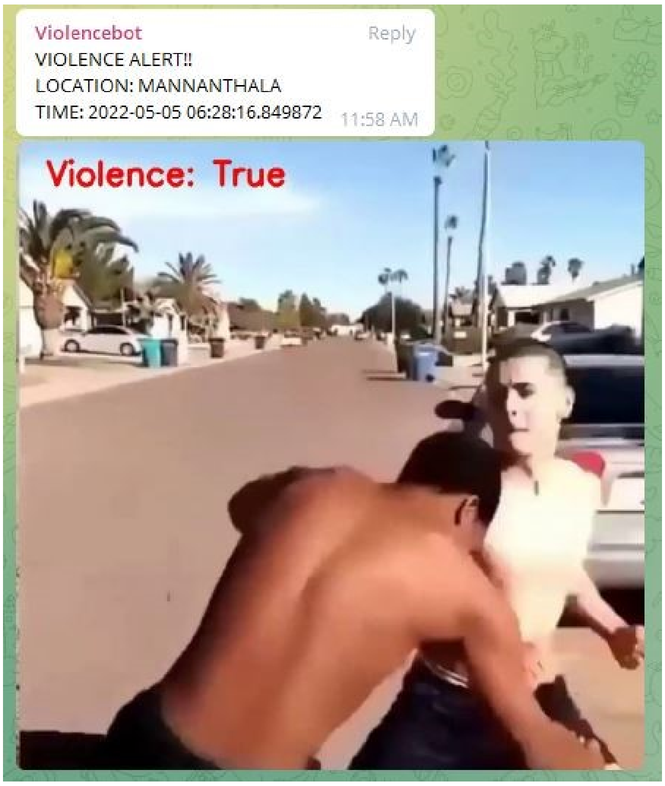


Figure 3.5: Architecture diagram of the Alert System

Figure 3.5 shows the alert message that is sent to the telegram group by the telegram bot. The concerned authorities can view the alert and take necessary actions.



Violencebot

VIOLENCE ALERT!!
LOCATION: MANNANTHALA
TIME: 2022-05-05 06:28:16.849872

11:58 AM

Figure 3.6: Screenshot of the alert message

3.3.1 Operating Environment

Python - The language used here is a popular programming language called Python which due to its huge number of pre-built libraries, is quite suitable for doing Machine Learning and Deep Learning projects. Python has libraries like Tensorflow and Open-CV which we have used in our project. Due to its easy to learn syntax, it is used world-wide for different purposes like web development etc.

Google Colaboratory - It is an environment or a web application developed by Google for helping people do python related projects like Machine Learning and Deep Learning based. To help us with the tasks that require extreme amount of computational power, Google allows us to use their Graphical Processing Unit or their Tensor Processing Unit for free if we ever need them. This project does need a decent amount of computational power, hence is used here.

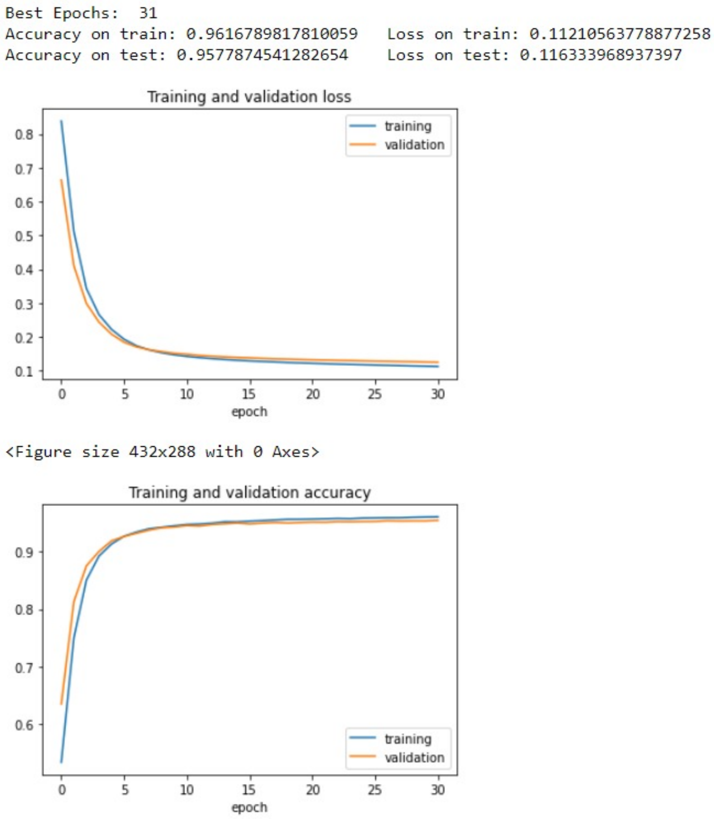
Chapter 4

RESULTS AND DISCUSSION

In this section testing and training accuracy are displayed in the below given graphical representation. Fig. 4.1 displays the training and testing accuracy and loss for the MobileNet v2 model when a dataset containing 1000 videos of average duration 7 seconds is given as input. For each epoch 350 videos from the violence class and 350 videos from the non-violence are trained. 96% accuracy was obtained on training and a respective accuracy of 95% was obtained when a CCTV footage that was not included in the dataset was given for testing. The obtained output video frames are shown in Figure 4.3

In the graph in Figure 4.1 the accuracy and loss comes to a constant level of increment and decrement after approximately 5 epochs. The obtained confusion matrix and other evaluation parameters are shown in Fig. 4.1

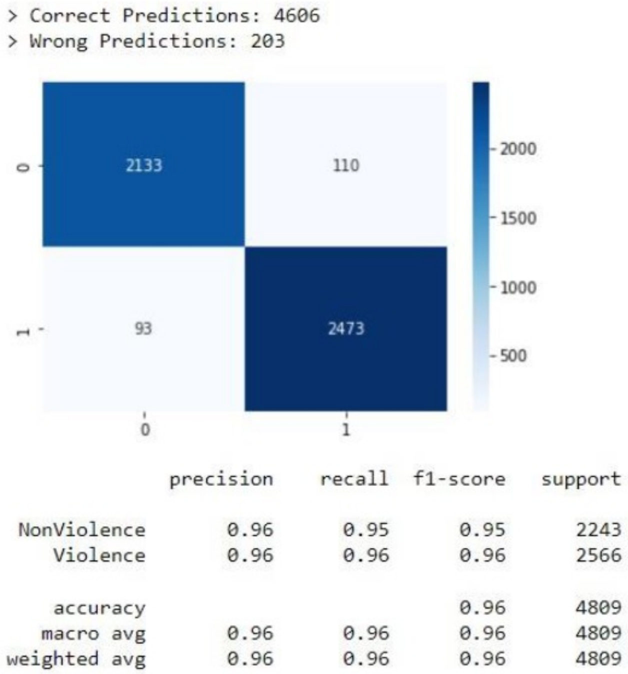
A video with violence is given as input to the system. Figure 4.3 shows one frame in the video that was labeled to have violent activity. Another video clip without violent activity was given as input. Figure 4.4 shows one frame of that video which is rightly labelled as false or violence.



```
Best Epochs:  31
Accuracy on train: 0.9616789817810059    Loss on train: 0.11210563778877258
Accuracy on test: 0.9577874541282654     Loss on test: 0.116333968937397
```

```
<Figure size 432x288 with 0 Axes>
```

Figure 4.1: Accuracy and Error of the training set



```
> Correct Predictions: 4606
> Wrong Predictions: 203
```

```
              precision    recall  f1-score   support

 NonViolence       0.96      0.95      0.95      2243
    Violence       0.96      0.96      0.96      2566

    accuracy                           0.96      4809
   macro avg       0.96      0.96      0.96      4809
weighted avg       0.96      0.96      0.96      4809
```

Figure 4.2: Confusion matrix of the trained model

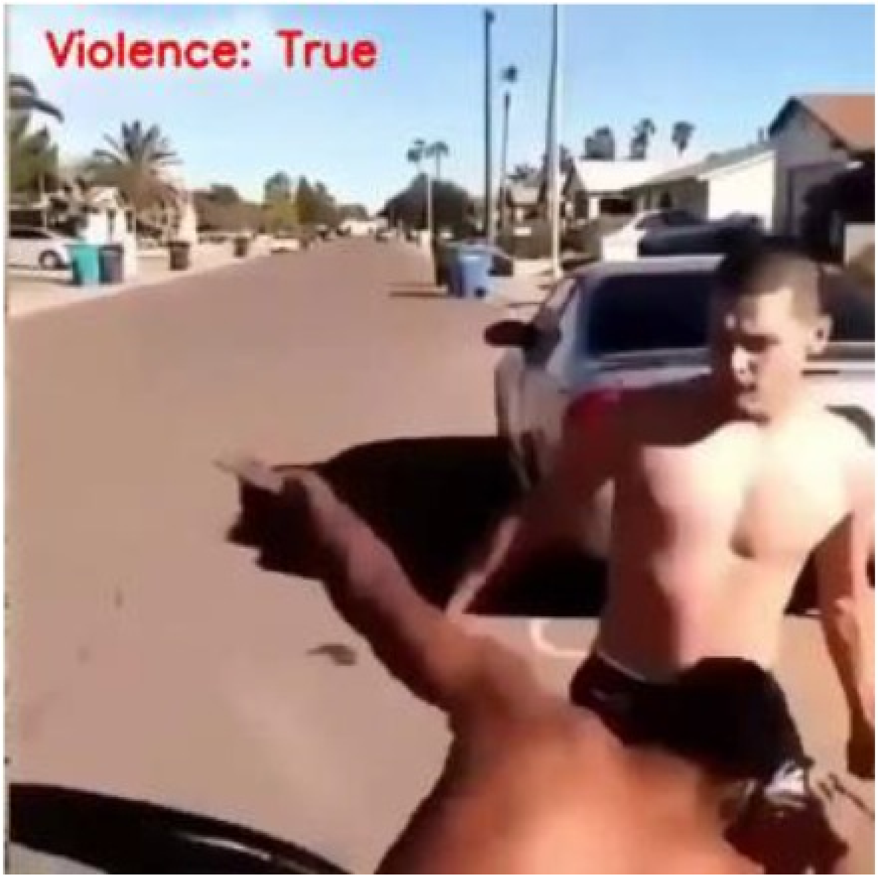


Figure 4.3: Output frame that recognized violence

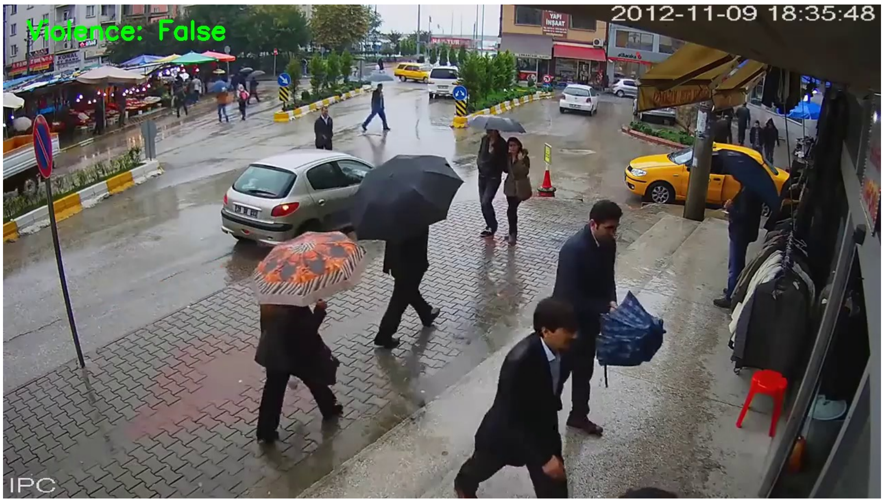


Figure 4.4: Output frame that did not recognize violence

4.1 COMPARISON WITH CNN-LSTM ARCHITECTURE

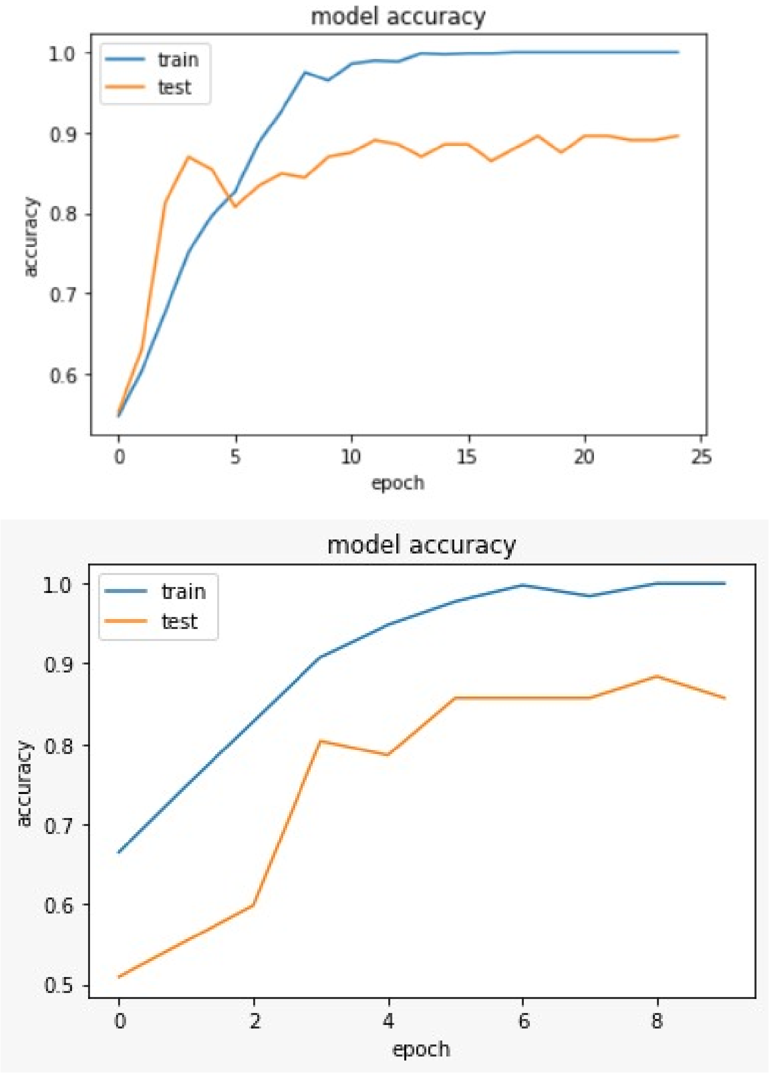
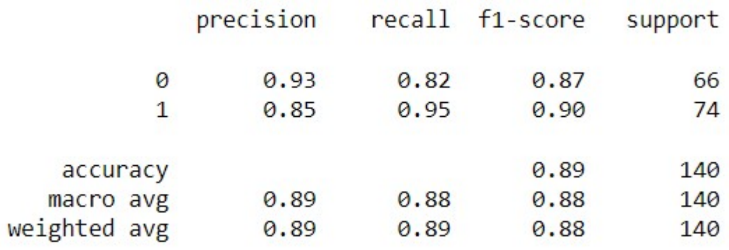


Figure 4.5: Comparison of Training and Testing accuracy of MobileNet v2 and CNN-LSTM Models



```
              precision    recall  f1-score   support

           0       0.93      0.82      0.87        66
           1       0.85      0.95      0.90        74

    accuracy                           0.89       140
   macro avg       0.89      0.88      0.88       140
weighted avg       0.89      0.89      0.88       140
```

Figure 4.6: Evaluation metrics of the CNN-LSTM Model

As the comparison shown in the Fig 4.1 MobileNet v2 has shown improvement than CNN-LSTM in the violence detection task. The above shown graphs has shown that MobileNet v2 is capable of performing better than model trained using CNN-LSTM. This proves that MobileNet v2 too can become the state of the art model for Real-time Violence detection.

Chapter 5

CONCLUSION

Violence scene detection in real-time is a challenging problem due to the diverse content and large variations quality. In this research, we use the MobileNet v2 model to offer an innovative and efficient technique for identifying violent events in real-time surveillance footage. The proposed network has a good recognition accuracy in typical benchmark datasets, indicating that it can learn discriminative motion saliency maps successfully. It's also computationally efficient, making it ideal for use in time-critical applications and low-end devices. Here, we had also shown the working of an Alert system that is integrated with the pretrained model. In comparison to other state-of-the-art approaches, this methodology will give a far superior option.

Future Scope: This model could be upgraded to work in multiple cameras connected by a single network in a concurrent fashion. A short video of the violent activity could be incorporated along with the alert message.

References

[1] Khan SU, Haq IU, Rho S, Baik SW, Lee MY. Cover the Violence: A Novel Deep-Learning-Based Approach Towards Violence-Detection in Movies. Applied Sciences. 2019; 9(22):4963. https://doi.org/10.3390/app9224963J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Gopalakrishna, MT. (2016). Violence Detection in Surveillance Video-A survey. International Journal of Latest Research in Engineering and Technology (IJLRET). NC3PS - 2016. 11-17. K. Elissa, “Title of paper if known,” unpublished.

[3] M. -S. Kang, R. -H. Park and H. -M. Park, "Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition," in IEEE Access, vol. 9, pp. 76270-76285, 2021, doi: 10.1109/ACCESS.2021.3083273.

[4] Ullah FUM, Ullah A, Muhammad K, Haq IU, Baik SW. Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network. Sensors (Basel). 2019 May 30;19(11):2472. doi: 10.3390/s19112472.

[5] A. -M. R. Abdali and R. F. Al-Tuma, "Robust Real-Time Violence Detection in Video Using CNN And LSTM," 2019 2nd Scientific Conference of Computer Sciences (SCCS), 2019, pp. 104-108, doi: 10.1109/SCCS.2019.8852616.

[6] J. C. Vieira, A. Sartori, S. F. Stefenon, F. L. Perez, G. S. de Jesus and V. R. Q. Leithardt, "Low-Cost CNN for Automatic Violence Recognition on Embedded

System," in IEEE Access, vol. 10, pp. 25190-25202, 2022, doi: 10.1109/ACCESS.2022.3155123.

[7] Sandler, Mark Howard, Andrew Zhu, Menglong Zhmoginov, Andrey Chen, Liang-Chieh. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. 4510-4520. 10.1109/CVPR.2018.00474.

[8] J. Wang and Z. Xu, "Crowd anomaly detection for automated video surveillance," 6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15), 2015, pp. 1-6, doi: 10.1049/ic.2015.0102.

[9] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo and A. F. Dragoni, "Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset," in IEEE Access, vol. 9, pp. 160580-160595, 2021, doi: 10.1109/ACCESS.2021.3131315.

[10] M. Ramzan et al., "A Review on State-of-the-Art Violence Detection Techniques," in IEEE Access, vol. 7, pp. 107560-107575, 2019, doi: 10.1109/ACCESS.2019.2932114.