JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Violence Detection using Human Action Recognition

Chaitali Nimse
Department of Electronics and Telecommunication
SIES Graduate School of Technology

Shubhangi Kharche
(Associate Professor)
Department of Electronics & Telecommunication
SIES Graduate School of Technology

Pooja Guchait
Department of Electronics and Telecommunication
SIES Graduate School of Technology

Chetna Pradhan
Department of Electronics and Telecommunication
SIES Graduate School of Technology

Pavlin Fernandes
Department of Electronics and Telecommunication
SIES Graduate School of Technology

Abstract

Violence detection using human action recognition is a research area that aims to automatically recognize patterns of human actions that are indicative of violent behaviour. The goal of this approach is to develop computer vision and machine learning techniques that can analyse video footage and identify instances of violent behaviour in real-time. A violence detection project using human action recognition is needed to address the increasing need for improved safety and security in various settings. By developing accurate and reliable methods for detecting violent behaviour in real-time, Additionally, this technology can be used in surveillance and law enforcement to identify and apprehend individuals who engage in violent behaviour. It typically involves using traditional computer vision techniques, such as feature extraction and machine learning algorithms, to analyse video footage and detect patterns of human behaviour associated with violent actions. In contrast, the proposed system for violence detection using human action recognition involves using deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to automatically extract features from video frames and identify patterns of human movement associated with violent behaviour. The goal of the violence detection project is to develop a machine learning model that can automatically detect instances of violence in video footage. The input data for the model consists of video footage captured by surveillance cameras or other sources, and the output will be a binary classification indicating whether the footage contains violence or not. The model is trained on a large and diverse dataset of video footage that contains examples of both violent and non-violent behaviour. The success of the model is measured using standard evaluation metrics such as accuracy, precision, recall, and F1 score. The frames individually are given to the model to predict if the frames contain any action of violence or not. The model has a high accuracy in detecting violence while minimizing false positives and false negatives. The model is trained using the BidirectionalLSTM and an accuracy score of 0.92 (92%) is obtained.

1. INTRODUCTION

Violence detection using human action recognition is a research area that aims to automatically recognize patterns of human actions that are indicative of violent behaviour. The goal of this approach is to develop computer vision and machine learning techniques that can analyse video footage and identify instances of violent behaviour in real-time. A violence detection project using human action recognition is needed to address the increasing need for improved safety and security in various settings. By developing accurate and reliable methods for detecting violent behaviour in real-time, this paper work has the potential to prevent violence and improve response times in emergency situations. Additionally, this technology can be used in surveillance and law enforcement to identify and apprehend individuals who engage in violent behaviour. It typically involves using traditional computer vision techniques, such as feature extraction and machine learning algorithms, to analyse video footage and detect patterns of human behaviour associated with violent actions. These systems often rely on hand-crafted features that require a significant amount of manual tuning to achieve accurate results. In contrast, the proposed system for violence detection using human action recognition involves using deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to automatically extract features from video frames and identify patterns of human movement associated with violent behaviour. These systems can be trained on large datasets to improve their accuracy and can adapt to different environments and contexts.

The paper is organised to cover literature survey in section 2, methodology in section 3 and Result and Discussion in section 4 followed by Conclusion and References in section 5 and 6 respectively.

2. LITERATURE SURVEY

The main aim of this research paper is to create a complete system that can perform real- time video analysis which will help recognize the presence of any violent activities and notify the same to the concerned authority, such as the police department of the corresponding area. Using the deep lea rning networks CNN and LSTM along with a well-defined system architecture, The author [1] has achieved an efficient solution that can be used for real-time analysis of video footage so that the concerned authority can monitor the situation through a mobile application that can notify about an occurrence of a violent event immediately. Convolution neural networks are used as the spatial feature extractor in the proposed architecture, followed by an LSTM network to perform sequence prediction on the feature vectors. A transfer learning strategy with CNN for spatial feature extraction is used. The architecture of the Xception network was examined using a pre-trained model on the ImageNet dataset. Rather than training from scratch, a pre-trained Xception model is employed as a feature extractor since it outperformed other pre-trained CNN models such as VGG, LeNet and ResNet. And then authors have fine-tuned it by leaving the first foulayers alone and retraining the last four layers on new datasets. Hockey, Cinema, and the UCF Crime dataset has been evaluated for the datasets.

In [2], the authors have proposed a deep learning model based on 3D convolutional neural networks, without using hand-crafted features or RNN architectures exclusively for encoding temporal information. The improved internal designs adopt compact but effective bottleneck units for learning motion patterns and leverage the DenseNet architecture to promote feature reusing and channel interaction, which is proved to be more capable of capturing spatiotemporal features and requires relatively fewer parameters. The performance of the proposed model is validated on three standard datasets in terms of recognition accuracy compared to other advanced approaches. Meanwhile, supplementary experiments are carried out to evaluate its effectiveness and efficiency. The final results demonstrate the advantages of the proposed model over the state-of-the-art methods in both recognition accuracy and computational efficiency.

The main aim of this research paper [3] is to Identify human activity involving finding human actions in sensor data.It has become more crucial in various industries, such as security, sports, and healthcare, because of the quick development of smart gadgets. Convolutional neural networks (CNN) and long short-term memory (LSTM) networks are two deep learning methods that are used to demonstrate significant promise for properly identifying human actions. It combines CNN and LSTM to achieve high recognition accuracy. In order to capture the temporal dynamics of the activity, It first extracts features from the raw sensor data using a CNN and then feeds these features into an LSTM network.

Detection of human activity is a constant problem for computer vision-based systems. [4] Due to the effective creation of artificial neural networks, such as the convolutional neural network CNN, activity recognition is more feasible and accurate. To extract spatial and temporal information from the input video, CNN and long short-term memory units (LSTM) -attention models in their recognition model architecture are used. It focuses on using Resnet and 3D CNN to identify human behaviour in real-time without the use of an LSTM attention model. In this paper the 2D Resnet is transformed to a 3D CNN to enhance the accuracy in human activity recognition.

The main aim of this paper [5] is to be automated in the identification and categorization of human action in films. The widely used 2D convolution neural network (CNN), incapable of temporal modeling, is not appropriate for this reason. Its solution to this problem is presented in this paper as a unique 2D CNN with an inter frame information extraction module based on bilinear operation. By using the parameter decomposition method, this model can substantially increase the 2D CNN's capacity for temporal modeling while requiring only a modest amount of storage and computation. It also features a flexible shape that makes balancing performance and complexity simple. Additionally, two different benchmarks, including both temporal-related benchmarks, are used to validate the performance of this novel network.

The main aim of this paper [6] is to identify and understand the actions of individuals in videos and export relevant tags. Actions in a video also possess characteristics in the temporal domain in addition to the spatial correlation seen in 2D images. The recognition will be impacted by the complexity of human behaviour, such as shifting views, background noises, and other factors. In this research paper, three approaches are developed and put into action for tackling those difficult problems. Two-Stream CNN, CNN+LSTM, and 3D CNN, which is based on convolutional neural networks (CNN), are used for recognizing human actions in videos.

This paper [7] focuses on overview of deep sequence learning approaches along with localization strategies of the detected violence. This overview also dives into the initial image processing and machine learning-based VD literature and their possible advantages such as efficiency against the current complex models. Furthermore, the datasets are discussed, to provide an analysis of the current models, explaining their pros and cons with future directions in VD domain derived from an in-depth analysis of the previous methods. In this review, the baseline research contributions in VD domain are discussed, leading to the current state-of-the-art approaches and advanced deep models. The research articles are retrieved from several famous databases such as Web of Science and Google Scholar. The overall PRISMA flow diagram for the research articles retrieval process is visualized in Figure 5. Overall, there are lots of articles retrieved during the initial search, that are first scrutinized using the title. For instance, the research contributions mentioning non-surveillance domains violence are not included in our review. Similarly, lots of low-level features-based articles are available in the early VD literature, but the ones with comparatively higher number of citations are discussed in the proposed survey.

In this paper [8], the authors have proposed a triple-staged end-to-end deep learning violence detection framework. First, persons are detected in the surveillance video stream using a light-weight convolutional neural network (CNN) model to reduce and overcome the voluminous processing of useless frames. Second, a sequence of 16 frames with detected persons is passed to 3D CNN, where the spatiotemporal features of these sequences are extracted and fed to the SoftMax classifier. Furthermore, they have optimized the 3D CNN model using an open visual inference and neural networks optimization toolkit developed by Intel, which converts the trained model into intermediate representation and adjusts it for optimal execution at the end platform for the final prediction of violent activity. After detection of a violent activity, an alert is transmitted to the nearest police station or security department to take prompt preventive actions. Authors then found that our proposed method outperforms the existing state-of-the-art methods for different benchmark datasets.

In this paper [9], authors have proposed a novel violence detection pipeline that can be combined with the conventional 2-dimensional Convolutional Neural Networks (2D CNNs). Frame-grouping is proposed to give the 2D CNNs the ability to learn spatio-temporal representations in videos. It is a simple processing method to average the channels of input frames and group three consecutive channel-averaged frames as an input of the 2D CNNs. Furthermore, it was presented that spatial and temporal attention modules that are lightweight but consistently improve the performance of violence recognition. The spatial attention module named Motion Saliency Map (MSM) can capture salient regions of feature maps derived from the motion boundaries using the difference between consecutive frames. The temporal attention module called Temporal Squeeze-and-Excitation (T-SE) block can inherently highlight the time periods that are correlated with a target event. Our proposed pipeline brings significant performance improvements compared to the 2D CNNs followed by the Long Short-Term Memory (LSTM) and much less computational complexity than existing 3D-CNN-based methods. MobileNetV3 and EfficientNet-B0 with our proposed modules achieved state-of-the-art performance on six different violence datasets.

In this paper [10], a proposed two-layer deep model is built for classifying video sequences into violent and non-violent actions. The first layer extracts the space features of video frames using the pre-trained model DenseNet-121. Then, the extracted features are fed to a long short-term memory (LSTM) network. LSTM captures the temporal features by learning the dependencies between frames, which links all frames of a video as one action. The proposed model is experimentally evaluated on two datasets. The recognition rate has improved up to 96%, which is better than those of most existing similar models over the open HOCKEY dataset and up to 92% over the real-live violence situations (RLVS) dataset. A simple, efficient model for detecting violence has been introduced. It is a combination of two deep learning models; DenseNet121 and LSTM.

In this work [11], authors proposed a deep learning architecture for violence detection, which combines both recurrent neural networks (RNNs) and 2-dimensional convolutional neural networks (2D CNN). In addition to video frames, optical flow computed using the captured sequences is used. CNN extracts spatial characteristics in each frame, while RNN extracts temporal characteristics. The use of optical flow allows to encode the movements in the scenes. The proposed approaches reach the same level as state-of-the-art techniques and sometimes surpass them. The techniques were validated on three databases achieving very interesting results. An end-to-end deep learning methods using RGB frames and an optical flow with a CNN-LSTM network to detect violent scenes in videos is introduced. The architectures presented reach the same level as modern techniques and sometimes surpassed them. These approaches have been tested on three public databases to validate their performances.

In this work [12], authors propose a novel 3D ConvNet along with a technique for extracting interest frames. The Structural Similarity Index Measure (SSIM) is exploited to extract interest frames as significant temporal information. Indeed, the SSIM uses the statistical features of two consecutive frames for this reason. In this way, sixteen video frames with the smallest SSIM are considered as dominant motion frames, which are then sent to a 3D CNN for classification.

METHODOLOGY

The goal of the violence detection is to develop a machine learning model that can automatically detect instances of violence in video footage. The input data for the model consists of video footage captured by surveillance cameras or other sources, and the output will be a binary classification indicating whether the footage contains violence or not. The model is trained on a large and diverse dataset of video footage that contains examples of both violent and non-violent behaviour. The success of the model is measured using standard evaluation metrics such as accuracy, precision, recall, and F1 score. The model has a high accuracy in detecting violence while minimizing false positives and false negatives. The dataset used is a collection of various YouTube videos containing both violence and non-violence videos. The code is then constructed in order to achieve the goals. At first the modules and libraries required are imported. Thereafter, the dataset is pre-processed and is prepared for further model training.

i) Data Pre-processing: The data is pre-processed and made sure it is fit for the proper training. The height and width of the video frames to be extracted are specified. Also, the sequence length is specified.

ii) Feature Extraction: Spatial and spatio-temporal features from the video frames are extracted using deep learning models and saved.

iii) Model Training: The dataset is split into training set (90%) and testing dataset (10%). Henceforth layers are formed and activation function is defined (here we used the Relu activation function). The model is then trained on the training dataset. In this step the model will learn all the features of a violent and non-violent video. Also, the learning rate is reduced once the model's accuracy becomes stagnant or starts decreasing. To maintain the accuracy. Early Stopping is used which avoids the model overfitting.

iv) Model Evaluation: The trained model is then evaluated on a separate dataset to assess its accuracy and generalization ability. This step includes assessing the model's performance through various performance metrics like Accuracy, F1 score, classification report, etc. This step involves testing the model's performance under different environmental and contextual conditions.

v) Model Improvement: The model is continuously improved by updating it with new data and adapting it to new environments and contexts.

Fig. 1 depicts the system architecture of the model. It consists of the several layers used to construct the model like time distributed layer, dropout layers, dense layer etc. We can also see the sequence of the layers in the system architecture.

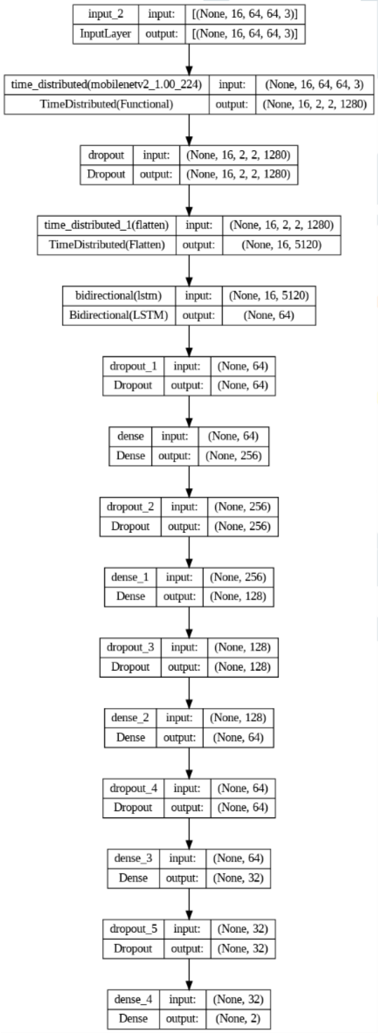

input_2 | input: | [(None, 16, 64, 64, 3)]
InputLayer | output: | [(None, 16, 64, 64, 3)]
time_distributed(mobilenetv2_1.00_224) | input: | (None, 16, 64, 64, 3)
TimeDistributed(Functional) | output: | (None, 16, 2, 2, 1280)
dropout | input: | (None, 16, 2, 2, 1280)
Dropout | output: | (None, 16, 2, 2, 1280)
time_distributed_1(flatten) | input: | (None, 16, 2, 2, 1280)
TimeDistributed(Flatten) | output: | (None, 16, 5120)
bidirectional(lstm) | input: | (None, 16, 5120)
Bidirectional(LSTM) | output: | (None, 64)
dropout_1 | input: | (None, 64)
Dropout | output: | (None, 64)
dense | input: | (None, 64)
Dense | output: | (None, 256)
dropout_2 | input: | (None, 256)
Dropout | output: | (None, 256)
dense_1 | input: | (None, 256)
Dense | output: | (None, 128)
dropout_3 | input: | (None, 128)
Dropout | output: | (None, 128)
dense_2 | input: | (None, 128)
Dense | output: | (None, 64)
dropout_4 | input: | (None, 64)
Dropout | output: | (None, 64)
dense_3 | input: | (None, 64)
Dense | output: | (None, 32)
dropout_5 | input: | (None, 32)
Dropout | output: | (None, 32)
dense_4 | input: | (None, 32)
Dense | output: | (None, 2)

Figure 1:System Architecture

3. Results And Discussions

A violence detection system was developed and successfully implemented. Fig. 2 and Fig. 3 shows that the model successfully detected the violence taking place in the given input video. The model is trained using the BidirectionalLSTM and an accuracy score of 0.92 (92%) is obtained. The given input video is divided into 16 frames as seen in Fig. 2. The frames individually are given to the model to predict if the frames contain any action of violence or not. In the result (Fig. 2- Fig. 3), all the frames show violence. In BidirectionalLSTMs, the frames are processed in both directions within two hidden layers, pushed toward the same output layer. This is how the model predicts the final output of the video (i.e., if the video contains violence or not) by relating the output of each frame.

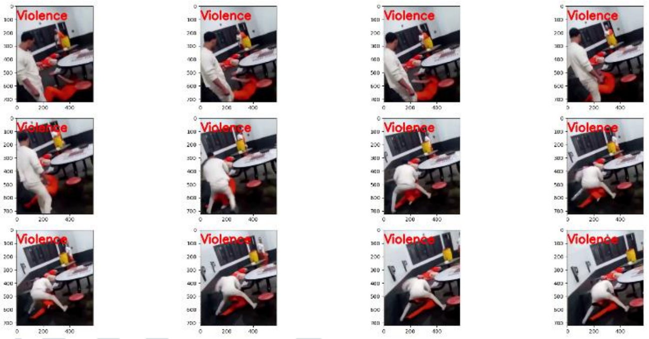


Figure 2: Frame wise output Prediction

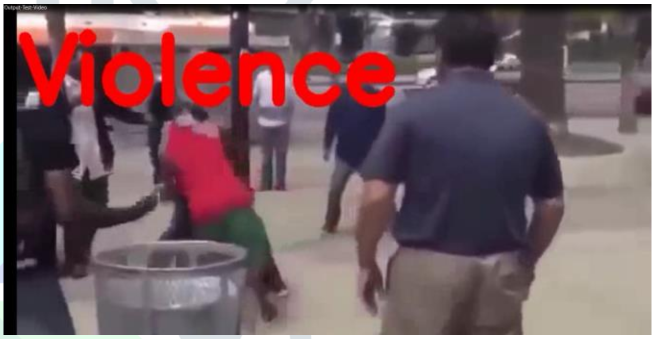


Figure 3: Final Output-Violence Detected

Accuracy

The Fig.4 shows the confusion matrix obtained for our model. The non-violence is given the value 1 and the violence is given the value 0. It is the graph constructed for Predicted labels against the true labels. The confusion matrix obtained showed the following result: i) True Positive: 90 i.e., the model correctly predicted non-violence videos as non-violence. ii) True Negative: 7 i.e., the model incorrectly predicted the non-violence videos as violence videos. iii) False Positive: 94 i.e., the model correctly predicted the violence videos as violence videos. iv) False Negative: 9 i.e., the model incorrectly predicted the violence videos as non-violence videos.

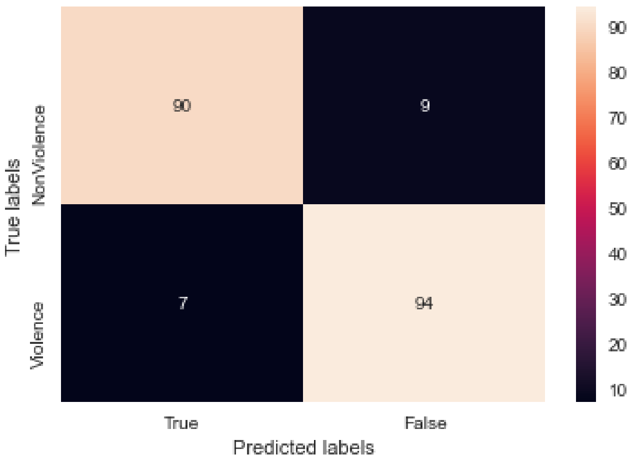


Figure 4: Confusion matrix

Fig. 5 shows the Total Accuracy Vs Validation Accuracy graph. We can see that the lines of both accuracy and validation have increasing accuracy with the successive number of epochs. But between the 15th and the 35th epoch there is not much increase in the graph of both validation accuracy and training accuracy as the model has reached accuracy of 1 and is not increasing after that. If we continue the epoch after this point the model will become overfit hence the epochs are stopped at the 39th epoch. Hence we can draw a conclusion that by continuing the epochs even after sufficient learning, the model becomes overfit.

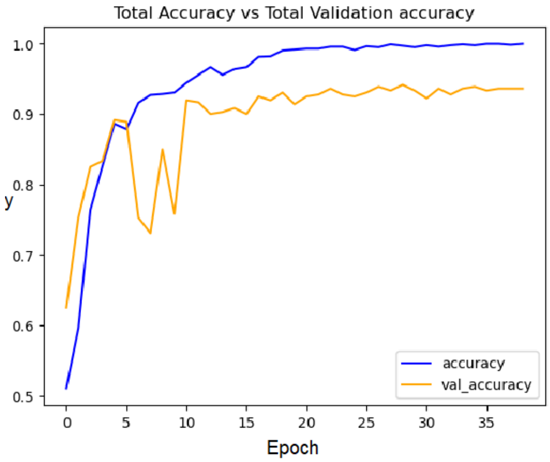


Figure 5: Total Accuracy Vs Validation Accuracy

Fig. 6 shows the Total Loss Vs Validation Loss graph. The graph consists of number of epochs on x-axis and the loss on the y-axis.

It is observed that the training loss decreases with the increasing number of epochs and the validation loss decreases at the start i.e., from epoch 0 to epoch 5 but after that it increases until it reaches the 10th epoch. There after there is a sudden decrease in the validation loss and it is stagnant after that for epoch 10 to epoch 35.

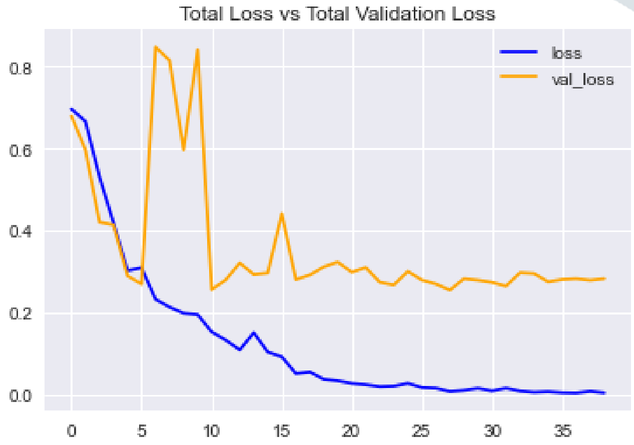
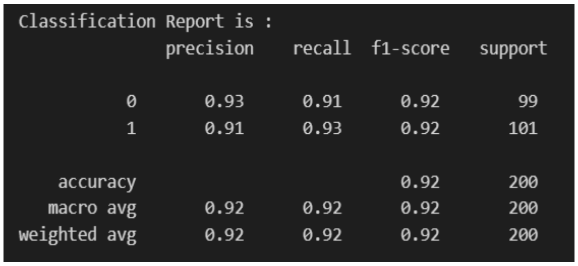


Figure 6: Total Loss Vs Validation Loss

Fig. 7 shows the classification report of the model. Accuracy of 0.92 is obtained. The macro average is 0.92 and weighted average is 0.92. The precision for violence is 0.93 and non-violence is 0.91. The recall value obtained for violence is 0.91 and non-violence is 0.93. F1 score obtained is 0.92 for violence and 0.92 non-violence. Support value is 99 for violence and 101 for non-violence.


```
Classification Report is :
              precision    recall  f1-score   support

           0       0.93      0.91      0.92        99
           1       0.91      0.93      0.92       101

    accuracy                           0.92       200
   macro avg       0.92      0.92      0.92       200
weighted avg       0.92      0.92      0.92       200
```

Figure 7:Classification Report

4. Conclusion & Future Scope

By developing accurate and reliable methods for detecting violent behaviour in real-time, Additionally, this technology can be used in surveillance and law enforcement to identify and apprehend individuals who engage in violent behaviour. The goal of this approach is to develop computer vision and machine learning techniques that can analyse video footage and identify instances of violent behaviour. By developing accurate and reliable methods for detecting violent behaviour in real-time, this paper work has the potential to prevent violence and improve response times in emergency situations. It typically involves using traditional computer vision techniques, such as feature extraction and machine learning algorithms, to analyse video footage and detect patterns of human behaviour associated with violent actions. These systems can be trained on large datasets to improve their accuracy and can adapt to different environments and contexts. The model is trained on a large and diverse dataset of video footage that contains examples of both violent and non-violent behaviour. The model has a high accuracy in detecting violence while minimizing false positives and false negatives. The frames individually are given to the model to predict if the frames contain any action of violence or not. The precision for violence is 0.93 and non-violence is 0.91. The recall value obtained for violence is 0.91 and non-violence is 0.93. However, there are still some challenges associated with violence detection using human action recognition, such as data availability and privacy concerns. The model will be able to recognise violence in busy areas in the future when we increase the number of layers, and we can further enhance the accuracy by training the model on different types of datasets.

5. References

[1] Sarthak Sharma, B Sudharsan, Saamaja Naraharisetti, Vimarsh Trehan, Kayalvizhi Jayavel,”A fully integrated violence detection system using CNN and LSTM” (2021)

[2]Ji Li Xinghao Jiang Tanfeng Sun Ke Xu,”Efficient Violence Detection Using 3D Convolutional Neural Networks” (2020)

[3]Chamani Shiranthika; Nilantha Premakumara; Huei-Ling Chiu; Hooman Samani; Chathurangi Shyalika; Chan-Yun Yang,”Human Activity Recognition Using CNN & LSTM” (2020)

[4] N. Archana; K. Hareesh,”Real-time Human Activity Recognition Using ResNet and 3D Convolutional Neural Networks” (2021)

[5] Jue Wang; Huanzhang Lu; Yao Zhang; Feng Ma; Moufa Hu,”Temporal Factorized Bilinear Modules with 2D CNN for Action Recognition in Videos”,(2022)

[6] Zeqi Yu; Wei Qi Yan,”Human Action Recognition Using Deep Learning Methods”(2020)

[7] Nadia Mumtaz, Naveed Ejaz, Shabana Habib, Syed Muhammad Mohsin, Prayag Tiwari, Shahab S. Band , Neeraj Kumar, ”An Overview of Violence Detection Techniques: Current Challenges and Future Directions” (2022)

[8]Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq,Sung Wook Baik, “Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network”, (2020)

[9] Min-Seok Kang; Rae-Hong Park; Hyung-Min Park, “Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition”, (2021)

[10] Yahia Reda Elkhashab1, Wessam H. El-Behaidy, “Violence Detection Enhancement in Video Sequences Based on Pre-trained Deep Models”, (2023)

[11] Abdarahmane Traoré; Moulay A. Akhloufi, “Violence Detection in Videos using Deep Recurrent and Convolutional Neural Networks”, (2020)

[12] Javad Mahmoodi, Hossein Nezamabadi-pour,Dariush Abbasi-Moghadam, “Violence detection in videos using interest frame extraction and 3D convolutional neural network”, (2022)