iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52182

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Real-Time Video Violence Detection Using CNN

Rizana Shaheer[1], Malu U[2]
[1]*Dept. of Electronics and Communication Engineering, Mahaguru Institute of Technology, Kerala*
[2]*Dept. of Electronics and Communication Engineering, Mahaguru Institute of Technology, Kerala*

Abstract: ***In order to effectively enforce the law and keep cities secure, monitoring technologies that detect violent events are becoming increasingly important. In computer vision, the practice of action recognition has gained popularity. In the field of computer vision, action recognition has gained popularity. The action recognition group, however, has mainly concentrated on straightforward activities like clapping, walking, jogging, etc. Comparatively little study has been done on identifying specific occurrences that have immediate practical applications, like fighting or violent behaviors in general. The responsiveness, precision, and flexibility of violent event detectors are indicators of their effectiveness across a range of video sources. This capacity might be helpful in specific video surveillance situations. Several research focused on violence identification with an emphasis on speed, accuracy, or both while ignoring the generalizability of various video source types. In this paper, a deep-learning-based real-time violence detector has been proposed. CNN serves as an extractor of spatial features in the suggested model. Here, a convolutional neural network (CNN) architecture called MobileNet V2 is utilized to extract frame-level information from a video, and LSTM, which focuses on the three factors (overall generality, accuracy, and quick reaction) as a temporal relation learning approach.***

Keywords: Machine learning, Deep learning, Convolutional Neural Network (CNN), Support Vector Machine (SVM), Particle Swarm Optimization (PSO)

I. INTRODUCTION

Recent advancements in the study of human action recognition have rekindled interest in the detection of certain actions, such as acts of aggression. A particular aberrant incident that includes one or more people using physical force to harm a person or damage property is known as a violent act[1].To differentiate between typical human behaviour and violent behaviour, violence recognition is a crucial first step in the development of automatic security monitoring systems. Until now, most camera surveillance systems are supervised by the human to manually analyze visual information and detect violent behavior. This manual supervision is practically infeasible and inefficient which leads to strong demand for automated violence detection systems.

Detecting violence in a video is basically a matter of splitting the video into frames to extract frame-level information, followed by detecting the movement of people in those frames. Then, spatial features are retrieved using either manual or machine learning methods. At last, classification algorithms are used to classify the retrieved features. The area of computer vision has grown rapidly due to the rapid rise of deep learning algorithms and the availability of vast amounts of data and computational resources. A Real-time violence detection from video using MobileNet architecture is presented in this paper. In Chapter 3, the system description is explained. The experimental result of the system is presented in Chapter 4. The conclusion of the work is given in Chapter 5.

II. LITERATURE REVIEW

A fast fight detection system has been suggested in [2]. Blobs of movement are initially discovered, and then they are characterized using a variety of features. The suggested approach has no assumptions about the number of people, body part recognition, or prominent point tracking. This technique's classification accuracy falls short of the best state-of-the-art fight detection techniques. The suggested method has trouble grouping videos with constant movements, such as those of moving clouds or tree branches in a windy environment.

Extreme acceleration patterns are presented as the primary distinguishing feature of a novel violence detection approach in [3]. By the application of the Radon transform to the power spectrum of successive frames, these high accelerations are calculated. Compared to state-of-the-art recognition techniques, accuracy gains of up to 12% are made. The technique could potentially function as the first attentional stage in a cascade framework that additionally incorporates STIP or MoSIFT features when the highest level of accuracy is required. The Violent Flows (ViF) descriptor and SVM classifier have been suggested in [4] as a unique method for the real-time identification of emerging violence in crowded environments.

This approach takes into account the statistics of how flow-vector values alter over time. ViF descriptor is used to represent these data points, which have been collected for brief frame sequences. The linear SVM is then used to categorize ViF features as violent or non-violent.

An end-to-end trainable deep neural network model has been put forth in [5] to solve the issue of violence detection in videos. In this approach, frame-level attributes are extracted using a convolutional neural network (CNN), and then features are aggregated in the temporal domain using a convolutional long short-term memory (convLSTM). In comparison to state-of-the-art approaches, this strategy performed better when tested on three separate datasets. A 3D convolutional neural network-based deep learning model with better internal architecture is provided in [6]. The suggested model may efficiently learn the temporal and spatial features of aggressive behaviors despite having relatively few parameters. This model can process data in real-time and is extremely effective at conserving computing resources.

Gated Recurrent Neural Networks are used in [7] to develop a novel hybrid deep learning model for the recognition of human action. To move around the subject in each frame of the movie, GMM, and a Kalman filter are utilized. Gated Recurrent Neural Networks only use data on the bounding boxes. For detecting human motion against a background, the Gaussian mixture model is used. The evaluation and feature extraction from every frame of the video, in time, was the key benefit of this novel method.

III.METHODOLOGY

The proposed system is a deep-learning-based violence detector. The temporal relation learning technique LSTM is combined with CNN, a spatial feature extractor, in the suggested model. The real-world dataset is used to record spatial and temporal characteristics for real-time implementation. The deep learning algorithm effectively distinguishes between violent and nonviolent content in the input video. Figure 1 shows the architecture of the proposed system.

A. Data Collection

The practice of acquiring and analysing information from a wide variety of sources is known as data collection. The real-life violence dataset was used for this implementation. The dataset is made up of videos that are divided into two folders: violence and no violence.

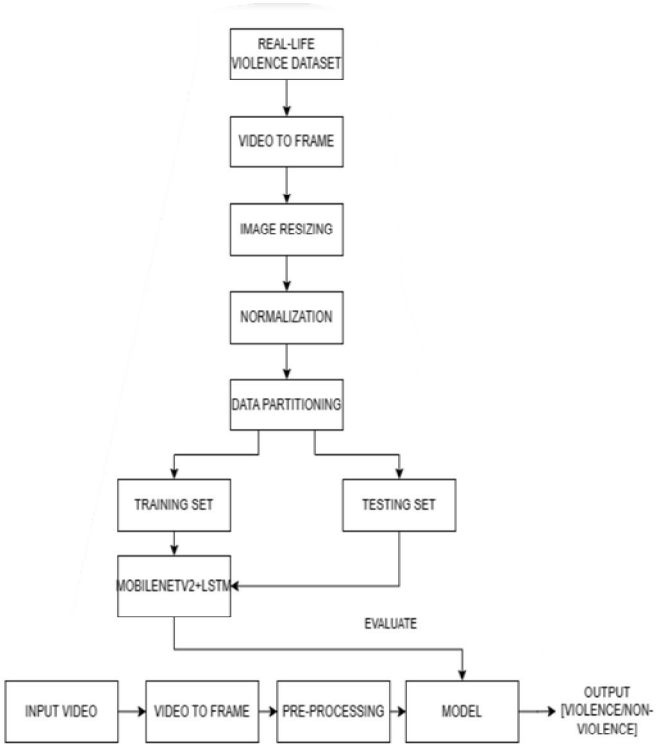


Fig 3.1 Flowchart

B. Data pre-processing

Real-world data often include errors and missing values, and it could even be in an inappropriate format that prevents it from being used directly by machine learning models. Pre-processing data can increase the correctness and quality of a dataset, making it more dependable by removing missing or inconsistent data values brought on by human or computer mistakes. It ensures consistency in data [7]. Frame extraction from videos, image resizing, and image normalization is the pre-processing steps employed in this work.

C. Data Partitioning

Data for training and testing were initially separated from the dataset. 20% of the dataset's images are for the testing set, while the other 80% are for the training set. The primary distinction between the two is that while the testing images lacked the ground truth labels for each image, the training data did. This enables us to assess the model, test it using data that hasn't been seen before, and use our own testing procedures.

D. Training

The neural network is trained using a training set. A MobilenetV2 and LSTM combo network has been employed in this study. LSTM network is utilized to capture variable dependencies, while MobilenetV2 gathers spatial characteristics. The process of developing a neural network involves constructing a sequential model and including the necessary network layers with the necessary parameters.

After adding MobilenetV2 to the network's topmost layer, further layers with the appropriate parameters, such as the Time Distributed layer, Flatten layer, Dropout, Dense layer, etc., are included as the remaining layers.

One AvgPool and 53 convolution layers make up the MobileNet V2 model. Inverted Residual Block and Bottleneck Residual Block are its two primary parts[8].

In the MobileNet V2 design, there exist two different types of convolution layers: 1x1 convolution and 3x3 depth-wise convolution. It creates a lightweight deep neural network by using depth-wise convolutions to drastically lower the number of parameters when compared to other networks. Only one filter is applied to each input channel using the depth-wise convolutions. Figure 3 shows the MobileNet architecture.

A long short-term memory network enables information to be retained. It is a unique variety of recurrent neural networks that has the capacity to address the vanishing gradient issue that RNNs are afflicted with. Three components—referred to as gates—make up an LSTM. They manage how data enters and exits the memory cell or lstm cell. The Forget gate, Input gate, and Output gate are the names of the first, second, and third gates respectively. For large datasets, LSTMs are favored.

E. Model Creation and Prediction

Model testing is conducted following training. After each iteration is finished, the model is tested with a test dataset. Testing is conducted following training. After each iteration is finished, the model is tested with a test dataset. Lastly, the model generated is used for future predictions. By using a testing set, evaluation is carried out for several classifiers. The performance of the model is evaluated using various parameters like classification accuracy, F1 score, precision, etc.

F. Experimental results

The model's performance across all classes is often described by its accuracy metric. When every class is equally important, it is helpful. It is determined by dividing the total number of predictions by the number of predictions that were correct. MobileNet architecture is used for violent and nonviolent activity detection from real-time video.1000s of real-time videos with an average runtime of 7 seconds are provided as the input dataset.

Videos of violence and nonviolence lessons are taught during each era. On training, the accuracy of more than 91% and 90%, respectively from real-time videos that weren't part of the testing dataset.

Confusion Matrix, a table that displays the evaluation results where the Y-axis corresponds to True Labels and the X-axis to Predicted Labels is shown in Figure 6. A heat map is a style of visualization in which each column is represented by a different color.

Figure 7 shows one frame in the video that was labeled to have violent activity. Figure 8 shows one frame in the video that was labeled to have non-violent activity.

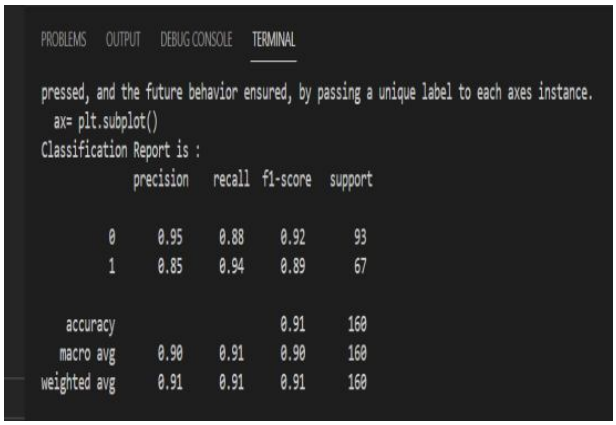


Fig 2: Classification report

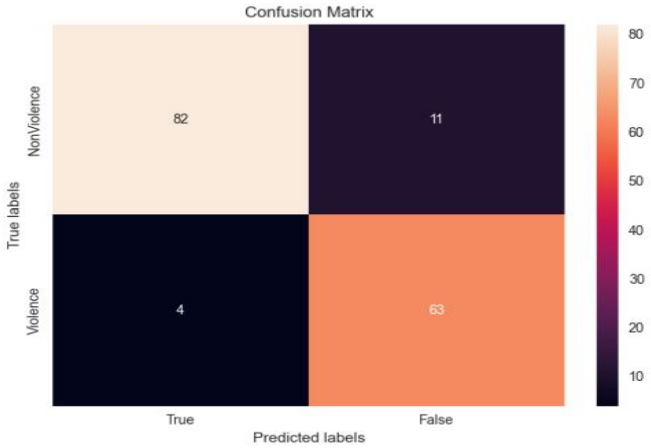


Fig 3: Confusion matrix

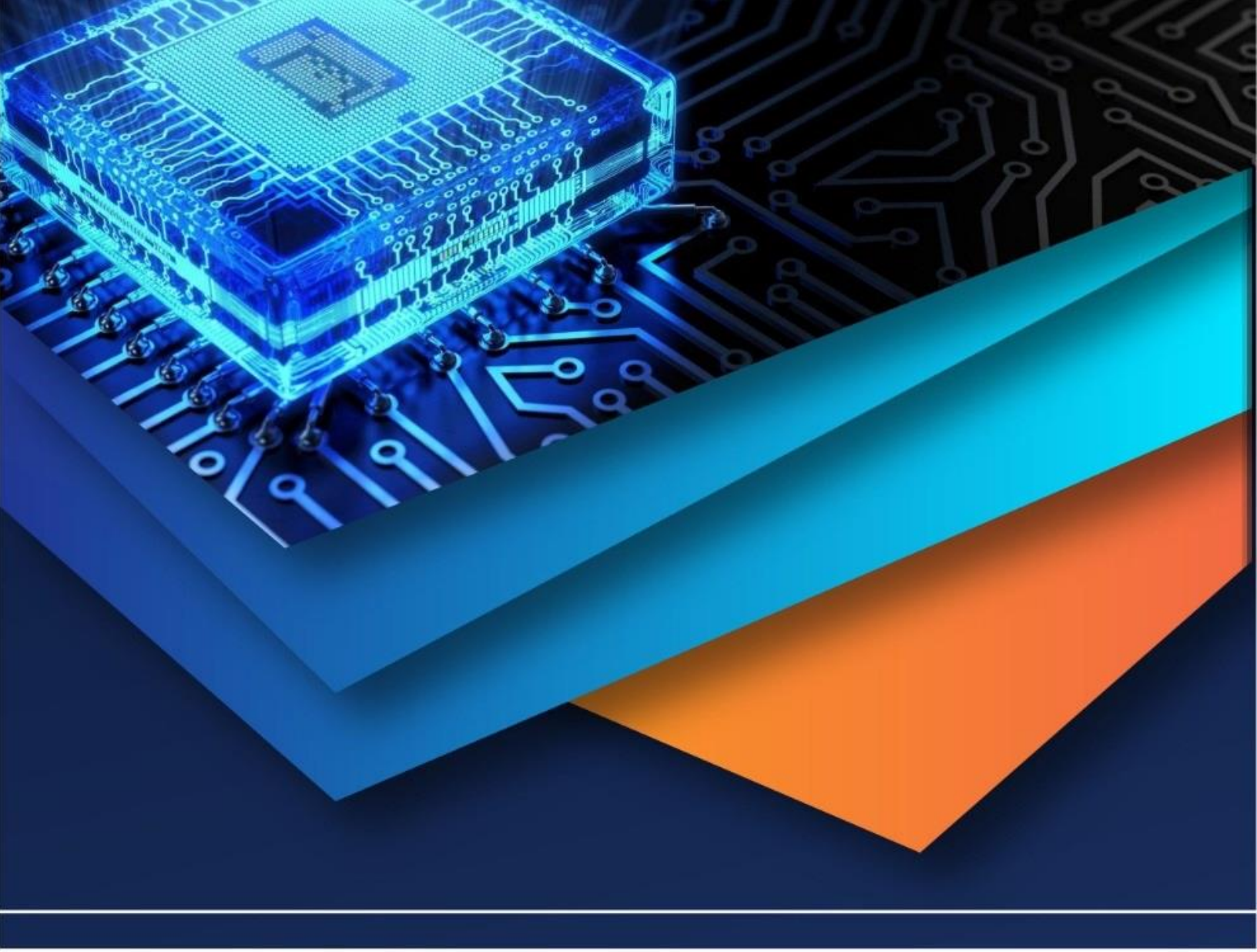
Fig 4: Violence and nonviolence detection from Video

IV. CONCLUSIONS

This study shows how to achieve significant accuracy and robustness in violence detection from real-time videos with limited data and computing resources using transfer learning. The system makes use of MobileNet architecture for violence detection. The proposed base model is tested on a benchmark dataset, and the results show that it performs better than the prior research for the same dataset with 91% accuracy. The suggested approach was also faster than earlier works. There is still scope for advancement in the creation of a fresh, well-balanced, substantial data set with numerous video sources for the identification of violence with greater sophistication to detect the violent action itself rather than merely the presence or absence of the violence.

REFERENCES

[1] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In CVPR Workshops, June 2012.
[2] Serrano Gracia I, Deniz Suarez O, Bueno Garcia G, Kim TK. Fast fight detection. PLoS One. 2015 Apr 10;10(4):e0120448.doi:10.1371/journal.pone.0120448. PMID: 25860667; PMCID: PMC4393294.
[3] Deniz, Oscar & Serrano Gracia, Ismael & Bueno, Gloria & Kim, Tae-Tyun. (2014). Fast violence detection in video. VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications. 2.
[4] Hassner, T., Itcher, Y., & Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 1-6.
[5] Sudhakaran, Swathikiran & Lanz, Oswald. (2017). Learning to detect violent videos using convolutional long short-term memory. 1-6. 10.1109/AVSS.2017.8078468.
[6] Li, Ji & Jiang, Xinghao & Sun, Tanfeng & xu, ke. (2019). Efficient Violence Detection Using 3D Convolutional Neural Networks. 1-8. 10.1109/AVSS.2019.8909883.
[7] https://ca.indeed.com/career-advice/career-development/datapreprocessing#:~:text=Importance%20of%20data%20preprocessing&text=It%20improves%20accuracy%20and%20reliability,It%20makes%20data%20consistent.
[8] https://medium.com/@godeep48/an-overview-on-mobilenet-an-efficient-mobile-vision-cnnf301141db94d
[9] Howard, Andrew & Zhu, Menglong & Chen, Bo & Kalenichenko, Dmitry & Wang, Weijun & Weyand, Tobias & Andreetto, Marco & Adam, Hartwig. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.
[10] Benjamin Lindemann, Timo Müller, Hannes Vietz, Nasser Jazdi, Michael Weyrich,A survey on long short-term memory networks for time series prediction, Procedia CIRP,Volume 99, 2021, Pages 650-655, ISSN 22128271,https://doi.org/10.1016/j.procir.2021.03.088.
[11] A. A. Eitta, T. Barghash, Y. Nafea and W. Gomaa, "Automatic Detection of Violence in Video Scenes," 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9533669.
[12] E. Fenil, Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm, 2019.
[13] K. Lloyd, A. D. Marshall, S. C. Moore and P. L. Rosin, Detecting violent crowds using temporal analysis of GLCM texture, vol. abs/1605.05106, 2016, [online] Available: http://arxiv.org/abs/1605.05106.
[14] S. Akti, G. A. Tataroglu and H. K. Ekenel, "Vision-based fight detection from surveillance cameras", 2019 Ninth International Conference on Image Processing Theory Tools and Applications (IPTA), Nov 2019.
[15] E. Y. Fu, H. Va Leong, G. Ngai and S. Chan, "Automatic fight detection in surveillance videos", Proceedings of the 14th International Conference on Advances in Mobile Computing and Multi Media ser. MoMM '16, pp. 225-234, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY