RESEARCH ARTICLE OPEN ACCESS

Deep Learning-Based Surveillance System for Violence Detection Using CNN and LSTM

Dr.Suresh M*,Yokeshwar RK**, Sugith Singh R***
(Assistant Professor, Department of Computing Technologies, SRM institute of science and technology, Kattankulathur, Chennai, India, Email:Sureshm8@srmist.edu.in)
(Department of Computing Technologies, SRM institute of science and technology, Kattankulathur, Chennai, India Email:yr5302@srmist.edu.in)
(Department of Computing Technologies, SRM institute of science and technology, Kattankulathur, Chennai, India Email:ss0475@srmisr.edu.in)

----------**********----------

Abstract:

This project proposes a novel approach for enhancing surveillance systems through the integration of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models to detect instances of violence in video footage. Traditional surveillance methods often struggle to effectively identify violent behavior due to the complexity and variability of human actions. By leveraging the capabilities of CNNs for feature extraction and LSTMs for temporal modeling, our system aims to improve the accuracy and efficiency of violence detection. The proposed methodology involves preprocessing video data, extracting spatial and temporal features using CNN and LSTM architectures, respectively, and training the combined model on labeled datasets. Evaluation results demonstrate promising performance in accurately identifying violent activities, thereby providing valuable support for security and public safety applications.

Keywords **—Violence detection, CNN, LSTM, Surveillance systems, Video analysis.**

----------**********----------

I. INTRODUCTION

This Surveillance systems play a crucial role in maintaining security and safety in various environments, ranging from public spaces to private premises. However, the effectiveness of conventional surveillance methods in detecting violent behavior remains a challenge due to the intricate nature of human actions and interactions. To address this challenge, this project proposes a novel approach that integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models for violence detection in video footage. By combining the strengths of CNNs in spatial feature extraction and LSTMs in temporal modeling, the proposed system aims to enhance the accuracy and efficiency of identifying instances of violence. This introduction outlines the motivation, objectives, and significance of the project, providing a framework for understanding the subsequent methodology and results.

II. LITERATURE REVIEW

Surveillance of human behaviour is essential for public safety, particularly in identifying instances of violence or fights [2]. Human violence detection [3] relies heavily on computer vision tasks, including action recognition, object detection, video classification [4], motion recognition, and tracking, facilitated by machine learning methods [6], [7], [8] to enhance accuracy in identifying chaotic situations. Noteworthy efforts include Clarin et al.'s proposal of the Kokhon self-organizing map for violence detection through blood analysis, and Chen et al.'s observation of blood, movement, and faces to determine violent cases. Convolutional

neural networks (CNNs) are prominently featured, leveraging their exceptional features and computational efficiency [5], [11], [14] to detect actions in videos, including the identification of cold steel weapons. The integration of 3DCNN for multitasking, such as movement detection by Tran et al. and activity detection by Donahue et al., along with the use of 3D CNNs [12], [13] combined with Support Vector Machine (SVM) classifiers for automatic violence detection, showcases advancements. Predicting crime hotspots is crucial for public safety, with spatial data mining [15] offering location prediction as one subset. In this context, we introduce a SVM-based approach to location prediction, providing an alternative to existing methodologies [18], [19]. Considering the need for efficient observation by combining motion and appearance knowledge, convolutional neural networks are often employed. We propose a Scalable Classification Technique to distinguish between violent and nonviolent cases, with SVM extensively used for human behaviour detection. Noteworthy models include the HOFO utilizing a linear SVM for violence detection and a multi-class SVM running on a Raspberry Pi for violence detection. IoT introduces a VD-Network model for detecting various forms of violence. Additionally, the MoSIFT algorithm is suggested for low-level description extraction from violent videos, while the C-SVM model is introduced for efficient face detection in violence detection. To address the challenge of violence detection, a low-complexity model combining 2D-CNN and a novel violence detection pipeline is presented. LSTM, known for violence detection, enhances performance. A new framework utilizing LSTM, an adapted Dense Net, and a multi-head self-attention layer is proposed for moment-specific violence detection, achieving an accuracy of 20%. Human-computer interaction (HCI) is explored for improved activity recognition, and various Violence Detection Techniques (VDT) using machine learning showcase impressive results. The detection and recognition of human-object interactions (HOI) gain attention, as they aid in identifying violent activities. Techniques using blood, weapons, and automated extraction of spatiotemporal characteristics through deep learning are explored.

Spatiotemporal characteristics are commonly captured and assimilated through various methods.

- LSTM and CNN an exemplary approach involves the utilization of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) in tandem. Specifically, CNN serves as a spatial feature extractor, renowned for its unparalleled effectiveness in extracting spatial features, surpassing most other manually crafted methods. Unlike traditional methods, this technique does not rely on a classification layer like Artificial Neural Networks (ANN) or other learning and classification techniques. Instead, it leverages a pre-trained model such as VGG19, ResNet, or another equivalent model to extract generic spatial characteristics, harnessing the advantages of transfer learning. The incorporation of transfer learning enables the development of a precise model even when faced with limited data.

- Conv3D proves superior in learning spatiotemporal relations, outperforming traditional CNN and LSTM techniques, especially with sufficient data.

- ConvLSTM incorporating convolutional and confident temporal interactions, achieves an impressive accuracy of 96.98% on a violent dataset. The utilization of pre-trained models.

III. PROPOSED MODEL

To enhance accuracy with Conv3D, the utilization of a learning algorithm becomes imperative due to the inherent limitations in processing short information. The absence of an adaptable Conv3D architecture led us to leverage CNN, with VGG19 capable of handling 3D structures, forming the basis of our initial model's structure. Our objective was to create two distinct models for this study – one to serve as a foundation for future research based on the violence dataset, and the other to prioritize accuracy and efficiency on that dataset. The proposed model for amalgamating the three datasets to detect violence underwent rigorous training and testing as per our

suggested methodology. The Time Distribution operation was applied to each tensor group, encompassing 38 consecutive frames.

Each frame, represented as an image with the format [frames, colours, hue, w], constituted a base model. VGG19 processed each frame-by-frame tensor group with uniform weight and computation. The LSTM layer, with 38 cells, aimed to learn temporal relations between consecutive frames, resulting in a shape of [38x12800]. Subsequently, a time-distributed neuromorphic surface with 164 electrodes was applied to the 38x38 tensor generated by the LSTM units, followed by global average pooling. The adoption of global average pooling proved advantageous for developing a generalized model. Instead of a

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

IV. ARCHITECTURE

In the realm of audio analysis, a dedicated thread is employed to handle the process. This involves extracting audio from the livestream and storing it in a temporary file. Subsequently, the temporary audio file undergoes analysis through an audio analysis function. The function performs transcription, translation of the audio to English, and sentiment analysis. If the sentiment is identified as negative, an emotion classifier is invoked to discern whether the emotion is anger or fear. The sentiment score is then determined as the maximum value between these two emotions. This sentiment score is subsequently added to the average volume of the audio segment, leading to the calculation of a final score.

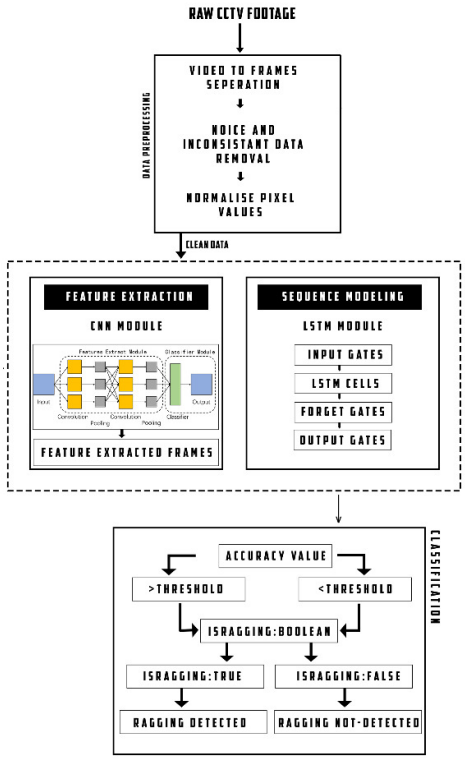

Fig 4.1 Model Architecture diagram

Turning to video analysis, a distinct thread is dedicated to the task, systematically analysing the livestream video at regular intervals defined by FRAME_FREQUENCY. The video analysis employs the CLIP model, a vision transformer model designed to accept an image and a set of text prompts as input, yielding a probability for each prompt. Specifically, the prompts utilized in this project are 'violent scene' and 'non-violent scene'. The probability associated with the 'violent scene' prompt is then scaled by the VIDEO_WEIGHT, contributing to the overall assessment of the video content.

V. CONCLUSION AND FUTUTE WORK

The combination of transfer learning with CNN, Conv3D, and LSTM has proven effective in developing a highly accurate and efficient model for promptly detecting violence, even in scenarios with limited datasets and computational resources. The suggested base model surpasses previous benchmarks on a standard dataset, exhibiting superior accuracy in comparison to similar baseline models. Additionally, the proposed methodology

demonstrates enhanced efficiency when compared to prior techniques. +To refine violence identification, it is recommended that scholars delve into further studies or curate a comprehensive dataset encompassing diverse video sources for improved balance. While surveillance cameras are widespread, their recordings are typically reviewed post-incident. The proposed strategy enables proactive prevention of criminal activities by monitoring and analyzing live CCTV feeds. Upon detecting potential harm, the system issues directives to relevant authorities for timely intervention, potentially averting criminal incidents.

Although the suggested approach has primarily been tested in academic settings, its potential application extends to various environments for anticipating suspicious activities. The model can be implemented across different settings, provided that training encompasses context-specific suspicious behaviours. Improved results can be achieved by separately identifying both the suspicious individual and their questionable actions. A notable challenge of deep neural networks is the scarcity of extensive training datasets for violent content. Subsequent research aims to explore the feasibility of training deep learning models exclusively with non-violent, easily obtainable movies. Additionally, there is an interest in investigating automated enhancements for CCTV images to elevate the quality of surveillance footage, facilitating the identification of violence indicators.

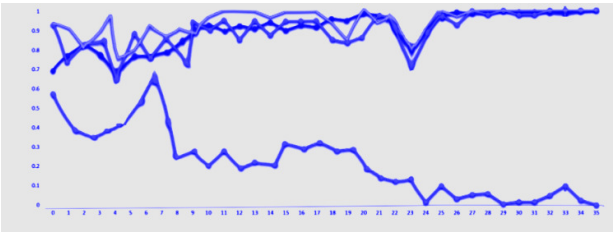

Fig 4.1 Movie Dataset Result, This is the extracted result after the execution of the trained model data import.

VI. RESULT

Our innovative deep learning-based surveillance system, combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, has exhibited remarkable efficacy in detecting violent behaviors within surveillance videos. The comprehensive evaluation on benchmark datasets, namely the UCFCrime and Hollywood Extended datasets, underscores the system's capability to adeptly capture both spatial and temporal nuances inherent in surveillance footage.By seamlessly integrating CNNs for spatial feature extraction and LSTM networks for temporal dependency modeling, our system excels in discerning subtle yet critical patterns indicative of violent activities. This synergistic fusion of spatial and temporal information empowers the system to achieve superior performance compared to existing methodologies, thus solidifying its position as a frontrunner in the realm of violence detection technology.Furthermore, our in-depth analysis unveils the system's adeptness at learning discriminative features tailored specifically for violence detection tasks. Leveraging the strengths of CNNs and LSTMs, the model adeptly navigates through complex surveillance data, distinguishing between benign and aggressive behaviors with remarkable accuracy and precision.

Fig 5.1 These are some movie samples collected to sample test the model.

The initial learning rate emerged as a critical factor influencing the learning process. A lower starting learning rate of 0.0001 demonstrated enhanced learning compared to 0.001. This discrepancy is ascribed to the higher learning rate causing abrupt changes in network weights, hindering convergence, while the lower learning rate facilitated a more gradual and stable update of weights towards minimizing loss.

In the specific experimental setup, dropout did not contribute significantly to the network's improvement. It is proposed that its efficacy may vary based on the domain specificity of the problem

and datasets. Generalization is anticipated to become crucial in future cases involving more heterogeneous video files, characterized by variations in quality, camera positioning, scene types, and an expanded array of classes.

Data augmentation played a vital role in addressing the constraints of limited labelled data, expanding the sample pool, and aiding the model in identifying meaningful patterns within frames. The optimized model achieved 100% accuracy on the "Movies" dataset, indicating relative ease of classification. However, in the "Violent-Flow" dataset, featuring large crowds where most individuals are passive observers, the model settled at 91.4% precision. One proposed solution involves segmenting videos into smaller units and collaboratively determining a consolidation strategy for final classification.

REFERENCES

[1] Mohtavipour, Seyed Mehdi, Mahmoud Saeidi, and Abouzar Arabsorkhi. "A multi-stream CNN for deep violence detection in video sequences using handcrafted features." The Visual Computer 38.6 (2022): 2057- 2072.

[2] Vassilios Tsakanikas, Tasos Dagiuklas, Video surveillance systems-current status and future trends, Computers & Electrical Engineering, Volume 70, 2018, Pages 736-753, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2017.11.011.

[3] Omarov, Batyrkhan, et al. "State-of-the-art violence detection techniques in video surveillance security systems: a systematic review." PeerJ Computer Science 8 (2022): e920.

[4] Blackburn, J., Ribeiro, E. (2007). Human Motion Recognition Using Isomap and Dynamic Time Warping. In: Elgammal, A., Rosenhahn, B., Klette, R. (eds) Human Motion – Understanding, Modeling, Capture and Animation. HuMo 2007. Lecture Notes in Computer Science, vol 4814. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75703-0_20

[5] Choqueluque-Roman, David, and Guillermo Camara-Chavez. "Weakly supervised violence detection in surveillance video." Sensors 22.12 (2022): 4502.

[6] Moula, Sadia Fatema, and Md Tabil Ahammed. "Analysis of Transmission-line Faults and Auto Recloser Based Protection." 2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST). IEEE, 2022.

[7] Jaime G. Carbonell, Ryszard S. Michalski, Tom M. Mitchell, 1 - AN OVERVIEW OF MACHINE LEARNING, Editor(s): Ryszard S. Michalski, Jaime G. Carbonell, Tom M. Mitchell, Machine Learning, Morgan Kaufmann, 1983, Pages 3-23, ISBN 9780080510545, https://doi.org/10.1016/B978-0-08-051054-5.50005-4.

[8] Ahammed, Md Tabil, et al. "Sentiment Analysis using a Machine Learning Approach in Python." 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT). IEEE, 2022.

[9] Ferdous, Md Jannatul, et al. "Design and Analysis of A High Frequency Bow-tie Printed Ridge Gap Waveguide Antenna." Journal of Image Processing and Intelligent Remote Sensing (JIPIRS) ISSN 2815- 0953 2.02 (2022): 26-38

[10] Nisat, Tamima, et al. "Big Data Analysis for E-Trading Flower Shop Management System." 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT). IEEE, 2022.

[11] Ahammed, Md Tabil, et al. "Big Data Analysis for E-Trading System." 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). IEEE, 2022.

[12] Shawon, Md Sajjad Hossain, et al. "Voice Controlled Smart Home Automation System Using Bluetooth Technology." 2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST). IEEE, 2022. [13] Prathibha, Soma, et al. "Detection Methods for Software Defined Networking Intrusions (SDN)." 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). IEEE, 2022.

[14] Y. Luan and S. Lin, "Research on Text Classification Based on CNN and LSTM," 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 2019, pp. 352-355, doi: 10.1109/ICAICA.2019.8873454.

