Real-Time Violence Detection Using Audio and Video: A Review

Nisma Navas[1], Sreeram I S[2], Nezvi Hussain K H[3] , C A Jasna[4] ,Archana Madhusudhanan[5]

[1,2,3,4]*U.G. Student, Department of Computer Science and Engineering, Universal Engineering College*

[5]*Assistant Professor, Department of Computer Science and Engineering, Universal Engineering College*

---***---

Abstract - In response to the urgent demand for increased public safety and security measures, this study presents a novel method for real-time violence identification in surveillance footage. This paper studies audio and video modalities, leveraging deep learning algorithms for comprehensive analysis. Discusses YOLOv4, a cutting-edge object recognition model, to accurately detect pertinent objects like individuals and weapons within video frames. Subsequently, employment of MobileNet for feature extraction and classification. By integrating audio analysis with video, the system achieves enhanced accuracy and robustness. Discuss audio features from accompanying audio streams and fuse them with visual characteristics to provide a holistic understanding of the scenario and different approaches. With its real-time capability, the system holds significant promise for deployment in various security and surveillance applications, thereby assisting efforts related to public safety. This integrated approach marks a significant advancement in violence detection technology, offering valuable support in addressing contemporary security challenges.

***Key Words*:** audio, MobileNet V2, video, violence detection, YOLOv4.

1. INTRODUCTION

As the world grows more interconnected by the day, protecting public safety and security has become crucial. Modern technology that can quickly recognize and address such dangers is desperately needed, as monitoring systems become more and more widespread [1]. Presents a novel technique for real-time violence detection in surveillance footage in answer to this pressing need. Leverages the combination of audio and visual modalities, driven by cutting-edge deep learning algorithms, to provide a thorough analysis and improve security protocols [2]. In the past, surveillance systems relied heavily on human operators. However, with advancements in efficiency and reliability, automated systems have become the preferred choice. These automated systems are highly beneficial for detecting violence and enhancing security measures [3].

Various deep-learning approaches have been extensively reviewed and discussed in recent years. The remainder of this paper is organized as follows:

There are 3 stages: Object detection, feature extraction and classification, and alert system.

Object detection techniques identify several items inside an image and offer information about their locations by drawing bounding boxes around them [1]. This is in contrast to image classification, which classifies a whole image into a single class. In many applications, such as augmented reality, driverless cars, surveillance, and facial recognition, object detection is essential.

Feature extraction is the process of eliminating unnecessary information from raw data and highlighting key features to convert it into an analysis-ready format. Feature extraction is used in many domains, including signal processing, computer vision, and natural language processing, to represent data in a more understandable and useful manner. On the other hand, classification entails labeling or classifying incoming data according to predetermined features. Usually, this procedure entails using labeled data to train a model that will determine the correlation between input attributes and output classes. The model can predict the class of unseen data after it has been trained [4].

An alarm system for violence detection combines machine learning algorithms with a variety of technologies, like audio and video analysis, to identify and notify authorities or staff about potentially dangerous situations [5].
The process of identifying violent behaviour or events through the use of audio signals is known as audio detection of violence. To discriminate between violent and non-violent noises, this approach usually entails analysing several aspects retrieved from the audio data. Audio detection techniques and discussed [13,14]

The objectives of this paper put forward is:

1. Innovative method for detecting violence.
2. Detection is done by audio and video.
3. Training on complex backgrounds.
4. Alert is given after detection.
5. Use of YOLO V4, and MobileNetV2 algorithms.

Detecting violent interactions is crucial in several delicate settings, including prisons, mental health facilities, and train stations, where maintaining security and safety is of the utmost significance. Handcrafted characteristics are the mainstay of traditional methods for identifying violent encounters in video surveillance. These methods frequently include extracting statistical features from motion regions. Nevertheless, these techniques have poor flexibility when used with other datasets or contexts.

The drawbacks of manually created feature-based techniques originate from their dependence on pre-established features created by human specialists. These traits might not fully represent the variety and complexity of violent interactions in various contexts. Consequently, these techniques might have trouble adapting to different situations or generalizing well to new datasets, ultimately resulting in less-than-ideal performance [6].

2. Related Works

The increasing number of surveillance cameras to watch over human activities necessitates the use of automated systems that can identify violent and suspicious situations. The field of computer vision and image processing is actively researching abnormal and violent behavior identification to draw in new researchers.

Muhammad Ramzan, et al. depicted in [7], methods of detection in this work are categorized into three groups: classical machine learning, Support Vector Machine (SVM), and Deep Learning for the detection of violence. Every single method's feature extraction and object detection approaches are also provided. Data collection, selection, screening out studies that aren't relevant, and analysis of suggested methods for violence identification based on characteristics taken from films are all part of the process.

A. Object Detection.

Nadeem Ahmed, et al. proposed in [8], the You Only Look Once (YOLOv4) paradigm is employed in the real-time abysmal activity detection system's methodology to detect objects. A cutting-edge neural network called YOLOv4 approaches object recognition in photos and videos in a novel way. It is a very reliable real-time detection model that works well with both images and movies and has a high degree of accuracy. When faced with a variety of scenarios, including low light, fuzzy, pixelated, zoomed-in, and low-resolution images or videos, the YOLOv4 model can identify abnormal behavior.

Tufail Sajjad Shah Hashmi, et al. compared in [4] between YOLOv2 and YOLOv4. Batch Normalisation (BN) was introduced to scale and modify activation. 2% increase in accuracy while using BN. A competitor of the Single Shot MultiBox Detector (SSD) Utilized fully connected layers, max-pooling, and convolutional layers. In YOLOv4, Cross-stage partial connections (CSP) were put into practice to improve learning. Batch division was done using Cross mini-Batch Normalisation (CmBN). Utilizing Self Adversarial Training (SAT) to enhance data. Self-regularized neuronal activation via Mish-Activation. Augmenting mosaic data to increase accuracy. For CNN, remove the block regularisation. Bounding box regression problem with CIoUs loss. In contrast: Compared to YOLOV2, YOLOV4 exhibits improvements in architecture and methodology. To improve accuracy and speed in object detection, YOLOV4 concentrates on enhancing learning capacity, data augmentation, and regularisation.

Tahreem Tahir et al. proposed in [9] that the YOLOv4 model performs violence detection for weapon detection. The outcomes attained When compared to other algorithms, YOLOV4 has performed better while consuming less time overall. YOLOV2 achieved 47% accuracy after 7 hours of training on D4, YOLOV3 achieved 49% accuracy after 5 hours of training on D4, and YOLOv4 achieved 52% accuracy after 5 hours of training. After spending three hours on the first layer, CNN achieved 35% accuracy, 38% on the second layer, and 41% on the third layer. When trained on YOLOV4, it achieved the highest accuracy on D6, while when trained on YOLOV2, it achieved the lowest accuracy on D4.

A. Feature Extraction and Classification.

Manjit Kumar Gautam, et al. proposed in [5], that a strategy for identifying violent behavior in surveillance videos uses a three-step process with deep learning methods. To identify people in the video feed, a MobileNet CNN model is first used. From a series of frames, a 3D-CNN model then extracts spatiotemporal properties to identify violent highlights. The observed actions are then classified by feeding these features into a softmax classifier. When violence is detected, a notification is issued to the security troops in the area so they can respond. To improve deployment efficiency, the OPENVINO toolbox is used to optimize the model. The usefulness of the suggested approach in detecting violence and enhancing security measures is demonstrated by the high accuracy rate of 99.9% on the violent crowd dataset, 98% on the hockey fight dataset, and 96% on the violence in movies dataset. The model training is employed by the MobileNet v2 framework.

Howard, et al put forward in [10], to drastically reduce processing and model size, MobileNet uses a simplified architecture based on depthwise separable convolutions, which substitute ordinary convolutions with two layers: depthwise and pointwise convolutions. These convolutions make up the majority of its network structure, with batch normalisation and ReLU nonlinearity after each layer to facilitate the exploration of diverse network topologies. Models can be made smaller and more effective by adding width and resolution multipliers as global hyperparameters. A table describing layer kinds, strides, filter shapes, and input sizes outlines the architecture's main body. Overall, efficiency is the top priority for MobileNet, which maintains good performance in activities like ImageNet classification and on-device identification while optimizing for latency and compactness.

Mark Sandler, et al. depicted in [11], MobileNetV2 is designed as a mobile architecture tailored for resource-constrained environments, prioritizing reduction in operations and memory usage while maintaining high accuracy. In its layer module, it presents the inverted residual with linear bottleneck, which effectively extends, compresses, and filters features to improve performance. For non-linearity, the design makes use of thin bottleneck layers with shortcut connections and lightweight depthwise convolutions in convolutional blocks. In comparison to ShuffleNet, NASNet-A, and MobileNetV1 models, MobileNetV2 performs better on a variety of workloads and benchmarks. It can also be used as a feature extractor for object detection, providing competitive accuracy with lower computational complexity and parameter requirements. While integration with DeepLabv3 resulted in Mobile DeepLabv3, a simplified form for semantic segmentation tasks, variants such as SSDLite use separable convolutions for effective object detection. MobileNetV2 concludes by showcasing enhanced performance across various model sizes, highlighting accuracy and efficiency in real-time and mobile applications.

B. Alert system.

Alert generation depicted in [5] by Manjit Kumar Gautam, et al., when the warning generation system notices violent activity in video frames, it triggers a counter. An alert is set off when 30 consecutive frames are highlighted, promptly alerting security staff or other authorities. The alert, which includes information on the incident's location, date, and time, is sent via Telegram. For more insight, the notice also includes a screenshot and a brief video of the incident.

In [12], the authors introduced an alert system for police. The suggested solution uses MATLAB's image processing capabilities to identify aggressive behavior, which then sends out an alarm and uses GPS and GSM to pinpoint the location. The block diagram of the system includes parts including an Arduino Uno microcontroller, an LCD screen for information display, GSM for cellular connectivity, a buzzer for aural signaling, and GPS for exact location tracking. Embedded C programming is utilized to control the embedded system, and the Arduino IDE is used for coding and compilation. The central controller is the Arduino Uno, which communicates with the GPS and GSM modules and interfaces with other hardware parts. The system uses a combination of hardware and software components for efficient implementation to deliver effective real-time violent incident detection and alerts.

C. Audio Detection.

In [13], the authors put forward two approaches for identifying violence in speech data are described in the study's material and methods section: Shallow Networks (SNs) and Transfer Learning using YAMNet. With an input layer, one hidden layer, and an output layer for the binary categorization of violent and non-violent speech, SNs provide a more straightforward method for pattern recognition. Mel-frequency Cepstral Coefficients (MFCCs), Delta MFCCs, pitch, Harmonic Noise Ratio (HNR), Short-time energy (STE), Zero Cross Rate (ZCR), Spectral Rolloff (SR), Spectral Centroid (SC), and Spectral Flux (SF) are among the pertinent features that are used in feature extraction. These features are acquired using the Audio Processing Toolbox in MATLAB. With labelled datasets, supervised learning is used to train the SN. Grid search is used to optimise parameters for best results.

The authors proposed in [14], the audio data is converted into spectrograms in order to represent audio as interpretable visual images. These spectrograms can subsequently be used for feature extraction or fed straight into Deep Learning classifiers. There are several ways to create spectrograms, including as the Mel-Spectrogram, Chromagram, and Short-Time Fourier Transform. Although every method has pros and cons of its own, the Mel-Spectrogram method is the most commonly used since it is in sync with human perception of sound. Mel-spectrograms use the mel scale as the frequency axis rather than a linear scale to show the frequency content of an audio stream with time. Equal distances on the scale correlate to equal perceived variances in pitch since the Mel scale considers humans' non-linear perception of frequency shifts. Scaling the frequency axis using the Mel scale and calculating the audio signal's Short-Time Fourier Transform (STFT) are the two steps in converting audio to a Mel spectrogram. A Mel spectrogram is the output, showing successive frequencies over time. The amplitudes are represented by colors and expressed in decibels.

3. CONCLUSIONS

The review article provides a thorough analysis of the integration of audio and video modalities utilising cutting-edge deep learning algorithms for real-time violence detection in surveillance film. It explores the use of YOLOv4 for object detection, MobileNetV2 for feature extraction and classification, and the setup of an alert system for timely alerts. The use of spectrograms and other audio characteristics for auditory violence detection is also explored in this work. The suggested approach improves the system's accuracy and resilience by integrating feature extraction, classification, and object recognition to provide a comprehensive knowledge of violent scenes. The study emphasises how important automated technologies are to enhancing security procedures, especially in settings where public safety is given top priority. The review's interdisciplinary approach, which has promising applications in the security and surveillance domains, emphasises the significance of thorough investigation and integration of diverse modalities in violence detection technologies. In the future, utilising cutting-edge technologies like reinforcement learning and attention mechanisms as well as investigating multimodal fusion approaches could further improve violence detection systems' capabilities, increasing their adaptability and efficacy in a variety of contexts.

Furthermore, the deployment of decentralised surveillance networks with real-time analysis and response capabilities is possible due to improvements in edge computing and distributed processing architectures, which could increase the reach and usefulness of violence detection systems.

ACKNOWLEDGEMENT

The authors extend sincere gratitude to Mrs. Archana Madhusudhanan for her Guidance and project coordinators, Mrs. Najla Nazar and Mrs. Gishma K M for mentorship. Their expertise played a vital role in shaping the direction and focus of this research.

REFERENCES

1. F. Reinolds, C. Neto 2, and J. Machado, Deep Learning for Activity Recognition Using Audio and Video, Electronics 2022, 11, 782.
2. N. Mumtaz, N, Ejaz, S. Habib, S. M. Mohsin, P. Tiwari, S. S. Band, and Neeraj Kumar, An Overview of Violence Detection Techniques: Current Challenges and Future Directions, Computer Vision and Pattern Recognition (cs.CV), 2022.
3. V. Dandage, H. Gautam, A. Ghavale, R. Mahore, and Prof. P.A. Sonewar, Review of Violence Detection System using Deep Learning, International Research Journal of Engineering and Technology (IRJET), vol. 06, issue: 12, Dec 2019.
4. T. S. S. Hashmi, N. U. Haq, M. M. Fraz, and M. Shahzad, Application of Deep Learning for Weapons Detection in Surveillance Videos, Conference: 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2).
5. M. K. Gautam, P. K. Rajput, Y. Srivastava, and Dr. A. Kansal, Real-Time Violence Detection and Alert System, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 12, Issue: III Mar 2024.
6. P. Zhou, Q. Ding, H. Luo, and X. Hou, Violent Interaction Detection in Video Based on Deep Learning, IOP Conf. Series: Journal of Physics: Conf. Series 844, 2017.
7. M. Ramzan, A. Abid, H. U. Khan , S. M. Awan, A. Ismail, M. Ahmed , M. Ilyas , A. Mahmood, A Review on state-of-the-art Violence Detection Techniques, IEEE, 2019.
8. M. Rahman, R. Rahman, K. A. Supty, R. T. Sabah, Md. R. Islam, Md. R. Islam, and N. Ahmed, A Real Time Abysmal Activity Detection System Towards the Enhancement of Road Safety, Conference: 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET).
9. Tahreem Tahir, Performance Evaluation and Comparison of YOLOv4 and Multiple Layers of CNN for Weapon Detection , IEEE, 2023.
10. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, arXiv preprint arXiv:170404861. 2017.
11. Sandler M, Howard A, Zhu M, Zhmoginov A, and Chen LC., Mobilenetv2: Inverted residuals and linear bottlenecks., In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018.
12. L. A. Kumar, B. Prathiba, and S. R. Meyyammai, Violence Alert System to the Police using GSM and GPS With the Help of Matlab, International Research Journal of Engineering and Technology (IRJET)Vol: 08 Issue: 04, Apr 2021.
13. F. Z. Zhou, D. T. Berengué, R. G. Pita, M. U. Manso, and M. R. Zurera, Computationally constrained audio-based violence detection through transfer learning and data augmentation techniques, Applied Acoustics, Vol: 213, October 2023, 109638.
14. D. Dalila, V. Bruno, and N. Paulo, Violence Detection in Audio: Evaluating the Effectiveness of Deep Learning Models and Data Augmentation, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, nº 3, September 2023.