Review of Violence Detection and Alert System in Video using Deep learning

Tejas Bose[1], Omkar Gaikwad[2], Prathamesh Chavan[3]

[1] *U.G Student, Department of Information Technology, AISSMS IOIT, Shivajinagar, Pune, India*
[2] *U.G Student, Department of Information Technology, AISSMS IOIT, Shivajinagar, Pune, India*
[3] *U.G Student, Department of Information Technology, AISSMS IOIT, Shivajinagar, Pune, India*

---***---

Abstract - The increasing prevalence of surveillance cameras for monitoring human activities necessitates automated systems capable of detecting violence and suspicious events. The detection of abnormal and violent actions has emerged as a significant area of research in computer vision and image processing, attracting considerable interest from researchers. This paper provides a comprehensive review of recent advancements in violence detection techniques. The methods reviewed are categorized based on their classification approaches, including traditional machine learning, Support Vector Machine (SVM)-based methods, and deep learning techniques. Additionally, this study highlights the feature extraction and object detection methods employed in each category. The datasets and video features that contribute significantly to the recognition process are also analyzed. To enhance understanding, an architectural diagram is presented to illustrate the key steps involved in the reviewed approaches. The findings of this review aim to guide future research by identifying gaps and opportunities for advancements in the field of violence detection**.**

1.INTRODUCTION

In an increasingly interconnected world, the demand for sophisticated security systems has become more critical than ever. Traditional surveillance technologies, which often rely on manual oversight or basic motion detection, are no longer adequate to address the intricate challenges posed by modern security needs. With the proliferation of public and private surveillance systems, the volume of video data generated has surged dramatically, making it exceedingly challenging to detect critical incidents, such as violent behaviour, in real-time. Consequently, there is a growing necessity for automated systems capable of identifying violent events with speed and precision, particularly in high-risk environments such as airports, shopping centres, industrial facilities, and public spaces.

Deep learning techniques have emerged as a promising solution to these challenges in the context of video surveillance. Harnessing the power of deep learning models like Convolutional Neural Networks (CNNs) has transformed multiple fields, including image and video analysis. These models excel at capturing both spatial and temporal patterns within video sequences, providing a significant edge over conventional approaches that typically rely on single-frame analysis. Among these advanced models, Three-Dimensional Convolutional Neural Networks (3D CNNs), especially those built on architectures like ResNet 3D, have demonstrated exceptional effectiveness in handling complex tasks such as activity recognition and violence detection in video data.

This system aims to overcome the limitations of traditional video surveillance methods by offering a real-time, automated solution for detecting violent actions within video streams. By leveraging deep learning, particularly the ResNet 3D architecture, the system not only analyses individual frames but also processes sequences of frames to capture the temporal context of violent behaviours, such as physical altercations or aggressive gestures. This ability to interpret patterns across multiple frames is vital for accurately identifying incidents, even in fast-moving and chaotic scenarios where conventional systems might falter.

Efficiency is another critical aspect of the system's design, given the need to process real-time video streams with minimal latency. To achieve this, video frames undergo pre-processing before being analyzed by the ResNet 3D model. This pre-processing step enhances the system's ability to quickly establish connections between frames and reliably detect violent events. When a potential incident is identified, the system promptly issues alerts to security personnel or relevant authorities, enabling immediate intervention. Such rapid responses are essential for mitigating harm, especially in situations where swift action could prevent escalation or save lives.

Additionally, the system is optimized for large-scale deployments, capable of managing multiple video streams concurrently without compromising performance. It is supported by a user-friendly interface that allows security teams to monitor live feeds, assess detected events, and manage alerts with ease. These features ensure that the system not only facilitates real-time monitoring but also promotes effective decision-making during critical incidents. By combining the strengths of deep learning with real-time video analytics, this system represents a significant step forward in automated violence detection, offering a scalable and practical solution for enhancing safety across diverse settings.

The primary objective of this system is to deliver actionable insights from video feeds promptly, allowing organizations to detect and address violent events before they escalate. Its capacity to analyse both spatial and temporal dimensions of video data marks a substantial advancement in violence detection automation. By improving the accuracy and efficiency of surveillance systems, this innovative approach empowers organizations to act decisively, ensuring that both public and private spaces remain secure and protected.

2. Related Work

In recent years, the need for advanced video surveillance systems capable of detecting violence in real-time has escalated due to security concerns across various domains, including public spaces, industrial environments, and urban areas. These systems often rely on sophisticated deep learning models to accurately identify violent actions from video footage, offering automated solutions to reduce human intervention and improve security monitoring. This literature review examines several recent advancements in the field of violence detection, focusing on methods utilizing convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid architectures combining multiple models to enhance detection performance.

Huszár et al. [1] introduced a framework for fast and accurate violence detection in automated video surveillance applications. Their work leveraged the power of CNNs for feature extraction and combined it with temporal models to process sequential data, allowing the system to detect violence in real-time. The primary challenge in this work was to reduce both the computational cost and the latency involved in processing high-definition video streams, which often require significant resources. To address this challenge, Huszár et al. proposed using lightweight models that balance speed and accuracy, making it suitable for deployment in surveillance applications. By using pre-trained CNNs such as VGG16, they were able to detect violent activities, such as fights, assaults, and abnormal behavior, even in crowded environments. The use of CNNs in this context was essential for understanding both spatial and temporal features, enabling the system to correctly classify video frames and detect violent incidents.

Shen Jianjie [2] further advanced the concept of violence detection by employing three-dimensional convolutional neural networks (3D CNNs). Unlike traditional 2D CNNs that focus on spatial feature extraction, 3D CNNs add a temporal dimension, making them particularly effective for dynamic events like violence. The paper introduced an architecture combining 3D convolutions with InceptionResNet, a variant of the Inception network known for its ability to learn features at multiple scales. This novel design allowed for better feature extraction from video sequences, significantly improving the model's ability to detect violent actions across time.

The 3D CNN approach excels in learning both spatial and temporal information, crucial for recognizing movements that are indicative of violent behavior, such as hitting, pushing, or aggressive gestures. By incorporating the InceptionResNet architecture, Shen Jianjie achieved improved accuracy compared to previous models, especially in cases where the violence was subtle or involved quick movements. The incorporation of InceptionResNet ensured that the model could efficiently capture high-level features while maintaining computational efficiency, making it a practical solution for real-time violence detection in surveillance systems.

Qiuhong Tian [3] proposed an innovative method for keyframe extraction, which is integral to improving the performance of 3D CNNs in action recognition tasks. Keyframe extraction refers to the process of selecting representative frames from a video sequence to reduce the amount of data that needs to be processed. By selecting keyframes that represent significant actions or changes in the scene, the model can focus on the most relevant information, speeding up computation and improving detection accuracy.

Tian's method of keyframe extraction relies on a novel algorithm designed to identify frames that capture critical moments in a video, such as the beginning or peak of violent actions. This technique is particularly useful in surveillance systems where large volumes of video data need to be analysed quickly. When combined with an enhanced 3D CNN, the system was able to detect violent actions more accurately by focusing on the most informative frames while reducing the noise introduced by irrelevant background information.

The enhanced 3D CNN model used by Tian was optimized for violence detection, focusing on learning both spatial and temporal features. This model was successful in improving detection rates for challenging cases such as low-light or low-resolution video, where traditional 3D CNNs might struggle.

Dong et al. [4] introduced a multi-stream deep learning approach to person-to-person violence detection. This method uses multiple streams of information, each designed to capture different aspects of a video sequence. For example, one stream might focus on motion information, while another might focus on object detection or background features. By combining these streams, the system can gain a more holistic understanding of the scene, which improves its ability to detect violence.

In this study, the authors used a combination of CNNs for spatial feature extraction and long short-term memory networks (LSTMs) for temporal processing. The CNNs extracted frame-level features, while the LSTMs processed these features over time, allowing the system to recognize violent actions that unfold over several frames. This hybrid architecture improved the model's performance in detecting complex behaviors, such as assaults or fights, where violent actions are often spread across multiple frames and can involve fast movements.

Dong et al.'s multi-stream architecture also benefited from the incorporation of human pose estimation, a technique that identifies and tracks human body parts across video frames. This addition made the model more robust in detecting violent actions even in scenes with multiple people or occlusions. The combination of CNNs, LSTMs, and pose estimation allowed for more precise violence detection, contributing to the advancement of person-to-person violence detection systems in surveillance applications.

As surveillance systems become more decentralized, the need for edge computing and IoT-based solutions has grown. A. Albunni et al. [5] explored the integration of convolutional neural networks with long short-term memory networks (CNN-LSTM) in an IoT environment for real-time violence detection. This model was designed to operate on IoT nodes deployed at the edge of the network, reducing the need for centralized servers and enabling faster response times. The combination of CNNs for spatial feature extraction and LSTMs for temporal processing made the system effective at detecting violent events in video streams.

By integrating violence detection models into edge devices, the authors were able to deploy real-time surveillance systems in industrial environments, reducing the latency typically associated with cloud-based solutions. Their system's ability to process video at the edge also minimized bandwidth usage, as only relevant events needed to be transmitted to central servers. This decentralized approach makes it feasible to deploy violence detection systems in large-scale surveillance networks with limited infrastructure.

F. U. M. Ullah and K. Muhammad [6] expanded on this concept by introducing AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks. Their model leverages AI to enhance the capabilities of edge devices, allowing them to perform real-time video analysis without relying on cloud resources. By embedding deep learning models directly into IoT nodes, the system could detect violent actions locally and take immediate action, such as triggering an alarm or notifying security personnel.

Khan et al. [7] proposed a deep learning approach to violence detection from industrial surveillance videos. Their system used deep convolutional neural networks (DCNNs) for detecting various types of violent actions, such as physical fights or aggressive gestures. This approach is particularly useful in industrial settings, where surveillance footage is often captured from a fixed camera angle and may involve low-resolution video.

The system demonstrated impressive performance in identifying violence even in challenging conditions, such as low-resolution or noisy video, highlighting the robustness of deep learning models in real-world applications. Their deep learning model incorporated temporal information, allowing it to recognize the progression of violent actions across video frames. This model provided a scalable solution for industrial surveillance, enhancing safety and security by detecting potential threats in real-time.

3. Proposed Methodology

The approach involves training the R3D ResNet model using a dataset of video clips labeled as violent or non-violent. Each video is divided into smaller clips, which are processed by the model through 3D convolutions to identify key features at different levels of the network. Residual connections help maintain effective learning in deeper layers by addressing the vanishing gradient issue. The model uses pooling and fully connected layers to classify the clips based on the spatiotemporal features it learns. To improve performance and adaptability, data augmentation methods like random cropping, flipping, and

temporal jittering is applied during training, alongside dropout regularization. The system begins by accepting video input, which goes through preprocessing. In this step, the video is cleaned, frames are extracted, and data augmentation is applied to expand the variety within the dataset. The frames are also normalized to ensure consistency in scale. After preprocessing, the system extracts spatial features, which capture details like objects and their positions in individual frames, and temporal features, which track motion and activity across frames. These spatial and temporal features are then combined to create a detailed representation of the video content. This combined data is fed into the R3D ResNet model, which is specifically designed for efficient and accurate video analysis. The model processes this information to determine whether the video contains violent content or not.

The system begins by accepting a video input, which goes through preprocessing. In this step, the video is cleaned, frames are extracted, and data augmentation is applied to expand the variety within the dataset. The frames are also normalized to ensure consistency in scale. After preprocessing, the system extracts spatial features, which capture details like objects and their positions in individual frames, and temporal features, which track motion and activity across frames. These spatial and temporal features are then combined to create a detailed representation of the video content. This combined data is fed into the R3D ResNet model, which is specifically designed for efficient and accurate video analysis. The model processes this information to determine whether the video contains violent content or not.

Once the prediction is made, post-processing techniques are employed to refine the results. Thresholding is applied to set a minimum confidence level for predictions, ensuring that only results meeting a certain standard are considered valid. Smoothing is used to reduce noise and improve the consistency of detection results. These steps enhance the reliability of the system. The final output indicates whether the video contains violent content, which can be utilized for various applications such as flagging inappropriate content, issuing alerts, or triggering further actions.

Algorithms used for Proposed Model:

The core operation in the R3D ResNet architecture is the 3D convolution, which captures both spatial and temporal features. The operation is mathematically expressed as:

$$Y(t,x,y)=\sum_{c=0}^{C-1}\sum_{k_t=0}^{K_t-1}\sum_{k_x=0}^{K_x-1}\sum_{k_y=0}^{K_y-1}W(k_t,k_x,k_y,c)\cdot X(t+k_t,x+k_x,y+k_y,c)$$

Where::

- Y(t,x,y) represents the output feature map at time t and spatial coordinates (x, y).
- X(t,x,y,c) is the input video clip at time t, spatial coordinates (x,y), and channel c.
- W(kt,kx,ky,c) denotes the 3D convolutional kernel with dimensions (Kt,Kx,Ky).
- C is the number of input channels.

Model: The R3D ResNet (ResNet-3D) is a deep learning model specifically designed for analyzing video data. It uses three-dimensional convolutions to process both spatial and temporal information in video clips. This model is based on the traditional ResNet architecture but has been extended to handle the unique challenges of video analysis. A key feature of R3D ResNet is residual learning, which helps the model learn differences (residuals) instead of direct mappings. This makes training deeper networks easier and prevents issues like vanishing gradients. The model extracts information in a

hierarchical way: Early layers focus on simple details like textures and edges, Deeper layers identify more complex patterns, such as object interactions and movements.

This step-by-step feature extraction is essential for understanding the sequence of actions in a video. By analyzing entire video clips, R3D ResNet effectively captures relationships between frames, making it well-suited for tasks like detecting violent actions or recognizing specific activities.

In the current work, we have utilized the Movies Fight Detection Dataset, a benchmark dataset sourced from Kaggle, to evaluate our violence detection model. This dataset is specifically designed for training and testing models aimed at recognizing fight scenes in videos. It consists of a total of 160 videos, categorized into two classes: 85 videos depicting fights and 75 videos showcasing non-fight scenarios. Each video has a duration ranging from 1 to 5 seconds, making it suitable for analyzing brief but intense interactions characteristic of fight sequences. These clips have been carefully selected to represent different fighting techniques and scenarios, contributing to the robustness of the model during training. Conversely, the non-fight videos encompass a range of non-violent interactions, ensuring that the model learns to distinguish between violent and non-violent behaviors effectively.

4. Conclusion and Future Scope

In recent years, the field of violence detection in surveillance videos has witnessed significant advancements, driven by the rise of deep learning techniques, particularly Convolutional Neural Networks (CNNs), Three-Dimensional Convolutional Networks (3D CNNs), and Long Short-Term Memory (LSTM) networks. These advancements have revolutionized the way violent

events are detected in real-time, offering substantial improvements in both accuracy and speed. Techniques such as multi-stream deep learning approaches, action recognition methods, and hybrid systems incorporating CNNs and LSTMs are particularly promising in addressing the challenges of dynamic, real-time video surveillance.

The papers reviewed highlight various methods, each contributing valuable insights into how violence detection systems can be enhanced through advanced machine learning models, improved training datasets, and the use of high-quality video frames.

The literature suggests that deep learning models, particularly those leveraging 3D CNNs and advanced architectures like InceptionResNet, offer the potential for improved recognition of violent actions, even in cluttered and complex environments. [1] and [2] emphasize the importance of temporal data captured through 3D CNNs for more precise identification of violence, as it allows for the analysis of actions in context over time, rather than relying solely on isolated frames. Furthermore, multi-stream approaches, such as those explored by Dong and Qin [4], demonstrate the efficacy of utilizing multiple data streams to analyze different aspects of a scene, which not only improves detection but also reduces the likelihood of false positives in real-world applications.

Another critical advancement lies in the integration of CNNs with other neural network structures like LSTMs for better contextual understanding and temporal sequence modeling. The combination of CNNs with LSTM networks allows for the modeling of sequential dependencies between frames, which is crucial in recognizing violent actions that may not be apparent in a single image or isolated frame. For instance, the IoT-based violence detection node proposed by Albunni et al. [5] demonstrates the effective use of this hybrid approach to build real-time, low-latency detection systems that can function efficiently in large-scale, industrial surveillance systems.

The use of edge-based vision systems, as discussed in the works of Ullah and Muhammad [6], introduces an exciting direction for deploying violence detection models in IoT-based industrial networks, where low-latency processing and real-time decision-making are critical. By processing video data locally, these systems reduce the reliance on central servers, thereby enhancing both speed and privacy. Furthermore, integrating AI-assisted edge vision into surveillance systems opens up the potential for proactive, on-site responses to violent incidents, which is particularly beneficial for industrial and urban security applications.

Looking ahead, there are several avenues for further research and development in violence detection systems. The future of violence detection will likely involve the refinement of existing models and architectures, focusing on improving their accuracy, scalability, and adaptability to

various real-world scenarios. Researchers could focus on developing models that can generalize across different environments and scenarios, which would allow for the deployment of violence detection systems in a wide range of industries, from industrial surveillance to public safety. Additionally, the integration of multi-modal data, such as audio alongside video, could improve the context and reliability of violence detection systems. For instance, combining visual and audio cues can help in detecting violent actions that may not be captured effectively by the camera alone.

REFERENCES

1. Toward Fast and Accurate Violence Detection for Automated Video Surveillance Applications, Huszár et al.
2. Violence Detection Based on Three-Dimensional Convolutional Neural Network with InceptionResNet, Shen Jianjie
3. Action Recognition Method Based on a Novel Keyframe Extraction Method and Enhanced 3D Convolutional Neural Network, Qiuhong Tian

4. Multi-Stream Deep Learning for Person-to-Person Violence Detection in Videos, Dong, Qin
5. A. Albunni, Convolutional neural network–long short term memory based IoT node for violence detection.
6. F. U. M. Ullah, K. Muhammad, AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks.
7. Violence Detection from Industrial Surveillance Videos Using Deep Learning HAMZA KHAN 1 , XIAOHONG YUAN 2 , LETU QINGGE , 3 , KAUSHIK ROY
8. A real time crime scene intelligent video surveillance systems in violence detection framework using deep learning techniques Author links open overlay panel Kishan Bhushan Sahay a , Bhuvaneswari Balachander b , B. Jagadeesh c , G. Anand Kumar c , Ravi Kumar d , L. Rama Parvathy