Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 46 No. 5
May 2025

Comparison of Deep Learning Techniques for Violent Content Detection in Video Stream

Akash Bhilwande [1], Dr. A. P. Patil [2], Aditya Toglekar [3]

[1, 2, 3] Walchand College of Engineering Sangli, Department of Information Technology, India
Email Id: [1] akash.bhilwande@walchandsangli.ac.in, [2] anita.patil@walchandsangli.ac.in,
[3] aaditya.toglekar@walchandsangli.ac.in

Abstract

The widespread adoption of surveillance systems has underscored the need for automated violence detection methods to bolster public safety. This survey explores both conventional machine learning techniques and modern deep learning approaches employed for violence detection. It examines the progression of these methodologies, evaluates their effectiveness on standard datasets, and sheds light on current challenges, including the complexities of real-time analysis and ethical dilemmas. Additionally, the paper suggests prospective research avenues aimed at overcoming these obstacles and enhancing the accuracy and reliability of violence detection systems.

Keywords: Violence detection, video analysis, pose detection, audio spectrogram, CNN, LSTM, multi-modal analysis.

1. Introduction

The rapid expansion of multimedia content on the internet has significantly increased the exposure of younger audiences to violent material. Studies have shown that Exposure to violent media content may increase the chances of aggressive behaviour in both short-term and long-term situations. The implications of exposure to the harsh violent content are not limited to individuals but extend to society as a whole, contributing to a potential increase in real world violence and fostering environments that normalize aggressive behaviours. Given this growing concern, the development of automated systems for violence detection has become an area of significant research interest. With the rise of video-sharing platforms, streaming services, and surveillance systems, the volume of video data being generated has reached unprecedented levels. Traditional manual monitoring methods are no longer sufficient to manage this influx, making automated violence detection systems critical for ensuring public safety and protecting vulnerable groups from exposure to harmful content. Automated violence detection systems offer multifaceted applications, ranging from content moderation on digital platforms to enhancing public safety in crowded environments However, designing effective violence detection systems is fraught with challenges. The nature of violent events can vary widely, encompassing both overt physical aggression and subtler forms of violence that may only be discernible through audio cues or contextual analysis. Furthermore, the performance of these systems often depends on the quality and diversity of the training datasets, which must capture the wide range of scenarios in which violence can occur. The inclusion of multiple modalities, such as audio and visual streams, adds complexity to the system design but also enhances its robustness. This survey aims to provide a comprehensive overview of the advancements in violence detection techniques. We categorize these methodologies into traditional machine learning approaches and contemporary deep learning frameworks, emphasizing the evolution of the field over time. The survey also explores the integration of audio-visual features, discusses the challenges inherent in real-world applications, and identifies potential research directions to address these issues. By systematically analyzing existing literature, this paper seeks to highlight the strengths, limitations, and opportunities for future work in the field of automated violence detection in video streams.

2. Background

The detection of violent content in multimedia has become a critical area of research due to the rapid proliferation of video and audio content across various platforms. Violence in multimedia can be broadly defined as physical aggression, harm, or behaviours intended to inflict injury or distress. Detecting such

content involves analysing both visual elements, such as fights and physical altercations, and audio cues, including screams, gunshots, or distress signals. However, this task is fraught with challenges. The complexity of violent scenarios, ranging from subtle acts like verbal aggression to overt physical confrontations, requires sophisticated analytical methods. Environmental factors, such as variations in lighting, occlusions, background noise, and diverse camera angles, further complicate the detection process. Additionally, creating diverse and ethically sourced datasets with balanced representations of violent and non violent content remains a significant hurdle. Traditional machine learning approaches relied heavily on handcrafted features from audio and video, such as Mel-frequency cepstral coefficients (MFCCs), optical flow, and motion trajectory analysis, paired with classifiers like Support Vector Machines and Hidden Markov Models (HMM). While these methods achieved success in controlled scenarios, their scalability and generalization across diverse datasets were limited. The advent of deep learning has transformed this domain, enabling automated feature extraction and better modeling of complex patterns. Convolutional Neural Networks excel in capturing spatial features from video frames, while Recurrent Neural Networks and Long Short-Term Memory networks handle temporal dependencies in sequential data. Additionally, combining audio and visual features through fusion techniques enhances robustness, and the use of pre-trained models such as ResNet and Inception has significantly accelerated progress in the field. Applications of violence detection systems are vast and impactful. They are employed in surveillance and security for real-time monitoring, content moderation on digital platforms, law enforcement for evidence collection, and parental control systems to safeguard children from harmful content. This evolving field has immense potential, but significant challenges remain, necessitating continued exploration of innovative approaches and technologies.

3. Literature Review

Violence detection in videos has been an active research area, leveraging deep learning techniques to improve accuracy and real-time performance. Various approaches have been explored, integrating spatial and temporal features through different architectures.A two-stage system combining audio and video processing using a C3D model pre-trained on the Sports-1M dataset was proposed in [1]. This method enhances detection accuracy by effectively capturing both spatial and temporal features. Another study developed a hybrid CNN that integrates 3D and 2D CNNs without pre-training, leveraging spatial and motion information to classify violent activities [2]. Similarly, a YOLOv5-based approach incorporating CNN-LSTM was introduced, where YOLOv5m weights were utilized alongside an LSTM trained from scratch to capture spatial and sequential patterns in violent events [3].

Further advancements include the integration of multiple YOLO versions, such as YOLOv5 and YOLOv8, with CNN-LSTM architectures like ResNet and Bi-LSTM for weapon and violence detection [4]. These architectures focus on deep spatial representations and sequential dependencies to improve classification accuracy. A lightweight approach based on MobileNetV2 integrated cloud connectivity for real-time alerts, using CNNs for efficient feature extraction [5]. Moreover, pre-trained 3D ConvNets were applied to multiple datasets, employing volumetric feature extraction to capture motion cues across frames [6].

Custom X3D models trained from scratch for spatiotemporal violence detection have also been proposed, optimizing architectures for efficient video classification [7]. Another system integrated pre-trained YOLOv3 and VGG19 for object detection, coupled with a custom LSTM model to recognize violent actions such as punching and kicking in surveillance footage [8]. A lightweight approach using MobileNetV2 pre-trained on ImageNet was combined with Cloud Firestore and Telegram bot integration for real-time alerts in violence detection applications [9]. Additionally, VGG19 was employed for spatial feature extraction, combined with an LSTM for temporal learning, facilitating motion pattern analysis based on body movements [10]. Custom X3D-XS and X3D-S models were trained from scratch to enhance spatiotemporal violence detection, utilizing 3D convolutional networks to extract fine-grained motion details across frames [11]. Optical flow integration with 3D CNNs has been explored in another study, where a dual-stream network separately processes spatial and motion features before classification, aiming to enhance recognition accuracy in surveillance footage [12].Some researchers have transformed videos into planar spatio-temporal representations, reducing reliance on 3D CNNs by employing a Random Walk

model to capture motion patterns efficiently, allowing a 2D CNN to process extracted features while maintaining key temporal information [13]. Another approach merged 2D depthwise and 3D convolutions in a hybrid CNN model, balancing spatial and temporal feature extraction while optimizing computational efficiency for real-time applications [14]. A lightweight 3D CNN model designed for video-based violence detection was introduced, incorporating depthwise convolutions and optimization techniques to improve feature extraction while maintaining low computational costs across datasets [15].

4. Comparison of Technologies

Violence detection in videos has been explored through various deep learning techniques, each with distinct architectures, datasets, and challenges. A two-stage system that integrates audio and video processing [1] leveraged a pre-trained C3D model on Sports-1M, achieving high accuracy across datasets such as Violent Flows (96.179%), HockeyFights (96.8%), Movies (97.575%), and RLVS (95.5%). However, this approach struggled with audio misclassification, particularly in noisy settings where sudden background noises could lead to incorrect predictions. Hybrid CNN models combining 3D and 2D architectures for spatial and temporal feature extraction [2] improved video analysis by capturing more contextual details. Despite achieving an impressive 99.93% training accuracy, the validation accuracy was significantly lower (84.5%), indicating dataset-specific overfitting. This highlights a common limitation in deep learning models: high accuracy during training that does not always translate well to real world generalizability due to small and non-diverse datasets. Another approach combined YOLOv5 for object detection with a CNN-LSTM pipeline [3] to detect violent activities such as fights. The model demonstrated strong classification results, with an LSTM accuracy of 93.5% for fights and YOLOv5 achieving 84-88% confidence in object detection tasks. However, the approach lacked real-world deployment and integration, making it less practical for immediate use in live surveillance scenarios. A more sophisticated system extended the use of YOLO object detection by integrating various CNN and LSTM architectures, including MobileNet, ResNet, and Bi-LSTM [4]. This multi-model approach achieved strong performance across datasets, with YOLO reaching ~92% accuracy, CNN+LSTM at 84%, and the best results from ResNetv2+Bi LSTM at 95%. While effective, its primary limitation was scalability in handling multi-class violent events, where real world complexity often exceeds the scope of pre-trained models. A fusion-based approach [5] combined a custom 3D CNN for spatial feature extraction with a CNN-LSTM model for temporal sequence analysis. The 3D-CNN achieved 86.11% accuracy, while the CNN-LSTM reached 87.60%. However, dataset-specific overfitting remained a concern, particularly in high-density crowd scenarios where movements may be subtle and misclassified. Another hybrid CNN-LSTM model [6] was trained from scratch using a surveillance dataset containing over 11,000 images. It reached 98.63% accuracy, showcasing strong generalizability in controlled environments. However, due to its lack of large-scale real-world testing, it remains uncertain how well it would perform in practical applications with diverse lighting conditions and occlusions. A YOLOv3-based model [7] integrated VGG-19 for object detection and an LSTM for temporal learning, focusing on real-world surveillance videos. The classifier achieved high performance for specific violent activities, such as punching (94.0032%) and kicking (96.1199%). However, real-time deployment posed challenges due to high computational costs, and the model struggled with occlusions in crowded environments, a common issue in public surveillance footage. A lightweight approach [8] employed MobileNetV2 for video frame classification, integrating Cloud Firestore for data management and a Telegram bot for alerts. While efficient, achieving 96% training and 95% testing accuracy, its reliance on network connectivity limited its real-time applicability, as any connectivity issues could disrupt functionality. Another study applied CNN-LSTM architectures with transfer learning via VGG19 [9] to extract spatial and temporal features from a combined dataset (Hockey Fight, Violent Flows, Movies, and YouTube videos). This model achieved 98% accuracy on the Hockey Fight dataset, 100% on Movies and YouTube, and an overall combined accuracy of 94.765%. However, due to limited training data, it struggled with detecting violence in crowd scenarios, where small-scale features and occlusions made classification difficult. A 3D CNN-based model [10] used pre-trained MobileNetV2 alongside 3D ConvNet for violence classification on datasets such as Real-Life Violence Situations and UCF Crime Dataset. Achieving 92-97% accuracy, it was highly effective but sensitive to frame rate inconsistencies, limiting its performance across different camera settings. Finally, a fully custom-built

3D CNN model [11] used X3D XS and X3D-S architectures trained from scratch for surveillance footage. It achieved high accuracy across datasets like RWF-2000 (94%) and ViolentFlows (98%), but overfitting remained an issue due to low sample sizes. This raised concerns about scalability, requiring additional data for improved generalization. Another study proposed a violence detection system leveraging 3D DenseNet models and optical flow-based feature fusion [12]. The approach introduced the AICS-Violence dataset, containing 7,576 high-resolution surveillance videos, with test sets designed to evaluate generalization across different camera angles. The system employed YOLOv4 for human detection, followed by candidate box extraction to focus on relevant regions in each frame. Two fusion techniques were explored: 3D DenseNet Fusion OF RGB and 3D DenseNet Fusion OFnom RGB, which combined spatial and motion-based features to enhance classification. The best-performing model achieved 97.675% accuracy on the Cam1 test set and 93.55% on the more challenging Cam2 test set, outperforming standard C3D (76.25%), ConvLSTM (91.75%), and standalone 3D DenseNet (96.55%). While the fusion-based models demonstrated superior performance, the computational complexity remained a challenge, limiting real-time deployment on low-resource hardware. Another approach, VDstr, proposed a novel spatio-temporal representation method for violence detection by transforming videos into planar representations (Prep), which retain both spatial and temporal information while allowing classification via 2D CNNs instead of computationally expensive 3D CNNs [13]. The method employed a Random Walk (RW) model to extract meaningful 1D structures from video frames, converting them into stacked temporal representations. A SqueezeNet-based CNN was then trained on these representations to classify violence. The VDstr model achieved 93.8% accuracy on RWF-2000, 98.5% on Movies Fights, 94.4% on Hockey Fights, and 89.8% on Crowd Violence, consistently ranking among the top three methods across datasets. While VDstr reduced computational costs compared to traditional 3D CNNs, it struggled with crowd complexity in surveillance footage and unstable camera movements in sports videos, which led to occasional misclassifications. A hybrid CNN model integrating 3D and 2D depth-wise convolutional layers was proposed for violence detection in videos, balancing spatial and temporal feature extraction while reducing computational complexity [14]. The architecture combined 3D CNNs for temporal feature extraction with 2D CNNs for spatial attention, leveraging depth-wise convolutions to improve computational efficiency. The method was tested on Hockey Fight, Surveillance Fight, Violent Flows, and Action Movies datasets, achieving 99.3%, 98.46%, 99.92%, and 98.2% accuracy, respectively, surpassing several state-of-the-art techniques. Despite its high accuracy across multiple datasets, the model's computational efficiency advantages over standard 3D CNNs were offset by challenges in handling complex occlusions and varying video quality, which could affect generalization in real-world scenarios. A modified X3D network was applied for violence detection across multiple surveillance datasets, leveraging hyper-parameter optimization and a custom data augmentation strategy [15]. The model was evaluated on RWF-2000, SCFD, and ViolentFlows datasets, incorporating techniques such as Bayesian hyper-parameter tuning, adaptive temporal sampling, and Grad-CAM visualization to enhance detection performance. The best model achieved 94.0% accuracy on RWF-2000, outperforming previous methods, while on SCFD, it achieved 88.7% accuracy, ranking second among state-of-the-art techniques. Results showed that longer temporal segments improved accuracy, but challenges remained in detecting subtle violent actions and handling occlusions in surveillance footage.

5. Challenges

Building a system to detect violent content in video and audio data involves several challenges that impact accuracy, efficiency, and ethical considerations. Defining violence is subjective, influenced by cultural and contextual differences, making it difficult for models to distinguish between real violence and fictional depictions. Dataset imbalance, where violent content is underrepresented, can lead to biased models, requiring resampling techniques or modified loss functions. Handling multi-modal inputs from video and audio is complex, demanding advanced fusion techniques. Real-time processing is computationally intensive, posing challenges for live monitoring applications. Models trained on the specific content types struggle to adapt across different type of data, necessitating diverse training data. False predictions can have serious consequences, requiring a balance between sensitivity and accuracy. Ethical concerns, including privacy and bias, must be addressed to ensure fairness. Model explainability is essential for

transparency, especially in content moderation. Scalability remains a challenge as user-generated content grows, requiring efficient processing methods. Additionally, adversarial attacks, where slight modifications in input data deceive models, highlight the need for robust defences. Computational cost is another major hurdle, as training deep learning models on large-scale video and audio datasets requires high-performance GPUs, which may not be accessible, making optimization techniques like model quantization, pruning, and efficient architectures essential for practical implementation. Focusing on these challenges is important for developing an effective and reliable violent content detection system.

6. Future Scope

The future of violent content detection systems is set to improve through advancements in deep learning, cross-modal fusion, and real-time processing. Leveraging transformers and attention mechanisms can enhance accuracy by capturing long-range dependencies in video and audio. Improved fusion techniques will allow better synchronization of multimodal inputs, strengthening detection reliability. Real-time processing can be optimized with edge computing, reducing latency and dependence on cloud resources. Customizable detection systems that consider cultural and linguistic differences can help reduce false positives and negatives. Expanding multilingual adaptability will make these systems more inclusive and effective across different regions. Training on diverse datasets will enable better performance in complex real-world conditions, such as crowded spaces or partial occlusions. A balanced approach combining AI automation with human moderation can refine decision-making, ensuring more context-aware assessments. Ethical considerations, including fairness and privacy protection, will remain vital in developing trustworthy systems. Additionally, optimizing computational efficiency through model compression and lightweight architectures will enable broader deployment, especially in low-resource environments. These innovations will drive the evolution of violent content detection into a more precise, scalable, and accessible technology.

7. Acknowledgement

We express our sincere thanks to all the authors, whose papers in the area of violent content detection are published in various conference proceedings and journals.

References

[1] D. Kinhal, R. K. S. Rishab, P. M. R. Pranav, M. P. Mayuravarsha, and R. Ravish, "Detection of Violent Content in Videos using Audio Visual Features," in Proceedings of the 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS), Bengaluru, India, April 2023, DOI: 10.1109/ICAECIS58353.2023.10170034.

[2] A. Jayasimhan and P. Pabitha, "A hybrid model using 2D and 3D Convolutional Neural Networks for violence detection in a video dataset," in Proceedings of the 2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India, Dec. 2022, DOI: 10.1109/C2I456876.2022.10051324.

[3] A. Khayrat, P. Malak, M. Victor, S. Ahmed, H. Metawie, V. Saber, and M. Elshalakani, "An intelligent surveillance system for detecting abnormal behaviors on campus using YOLO and CNN-LSTM networks," in Proceedings of the 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, May 2022, pp. 104-109, DOI: 10.1109/ MIUCC55081.2022.9781786.

[4] F. A. Razzaq, W. Tariq, M. A. Chaudary, M. Waqas, S. Fareed, and S. Javaid, "Enhancing public safety: Detection of weapons and violence in CCTV videos with deep learning," in Proceedings of the 2023 25th International Multitopic Conference (INMIC), Lahore, Pakistan, Dec. 2023, DOI: 10.1109/ INMIC60434.2023.10465800.

[5] S. M. Shanmughapriya, G. S. Gunasundari, J. R. Fenitha, and S. R. Sanchana, "Fight Detection in surveillance video dataset versus real-time surveillance video using 3DCNN and CNN-LSTM," in Proc. 2022 Int. Conf. Computer, Power and Communications (ICCPC), 2022, pp. 313–317, DOI: 10.1109/ICCPC55978.2022.10072291.

[6] R. G. Tiwari, H. Maheshwari, A. K. Agarwal, and V. Jain, "Hybrid CNN-LSTM Model for Automated Violence Detection and Classification in Surveillance Systems," in Proceedings of the 12th International Conference on System Modeling & Advancement in Research Trends (SMART), 2023, pp. 169–175, DOI: 10.1109/ICCPC55978.2022. 10072291.

[7] D. S. S., S. Govindraj, and S. N. Omkar, "Real-time Violence Activity Detection Using Deep Neural Networks in a CCTV Camera," in Proceedings of the 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2021, DOI: 10.1109/CONECCT52877. 2021.9622739.

[8] M. Thomas and P. Balamurugan, "Real-Time Violence Detection and Alert System using MobileNetV2 and Cloud Firestore," in Proceedings of the 2nd International Conference on Networking and Communications (ICNWC), 2024, DOI: 10.1109/ICNWC60771.2024.10537319.

[9] A. R. Abdali and R. F. Al-Tuma, "Robust Real-Time Violence Detection in Video Using CNN and LSTM," in Proceedings of the 2nd Scientific Conference of Computer Sciences (SCCS), University of Technology, Baghdad, Iraq, Mar. 27–28, 2019, pp. 104-108, DOI: 10.1109/SCCS.2019.8852616.

[10] N. Suba, A. Verma, P. Baviskar, and S. Varma, "Violence detection for surveillance systems using lightweight CNN models," in Proc. 7th Int. Conf. Computing in Engineering & Technology (ICCET), 2022, DOI: 10.1049/icp.2022.0587.

[11] J. Su, P. Her, E. Clemens, E. Yaz, S. Schneider, and H. Medeiros, "Violence detection using 3D convolutional neural networks," in Proc. 18th IEEE Int. Conf. Advanced Video and Signal-Based Surveillance (AVSS), 2022, DOI: 10.1109/ AVSS56176.2022.9959393.\

[12] A. H. U. Rahman, A. A. Munir, F. Hafeez, and M. A. Khan, "Violence Detection using Feature Fusion of Optical Flow and 3D CNN on AICS-Violence Dataset," *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*, 2022, pp. 141-146, doi: 10.1109/ ICCE55644.2022.9852065.

[13] M. Chelali, C. Kurtz, and N. Vincent, "Violence Detection from Video Under 2D Spatio-Temporal Representations," *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 1234-1238, doi: 10.1109/ICIP42928.2021. 9506142.

[14] J. Mahmoodi, H. Nezamabadi-pour, and B. Mirzaei, "Improved Violence Detection in Video Analysis with a Hybrid CNN Approach Using 3D and 2D Convolutional Networks," *2024 19th Iranian Conference on Intelligent Systems (ICIS)*, 2024, pp. 30-35, doi: 10.1109/ICIS64839.2024.10887525.

[15] J. Su, P. Her, E. Clemens, E. Yaz, S. Schneider, and H. Medeiros, "Violence Detection using 3D Convolutional Neural Networks," *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2022, pp. 1-6, doi: 10.1109/AVSS56176.2022.9959393.