MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Open Access International Journal of Science and Engineering

ISSN: 2456-3293

Volume 9 Issue 1s (2026) | Pages: 154-160

AI-Powered Real-Time Violence Detection in CCTV Feed

Anand Kawade[1], Mayank Gajengi[2], Aryan Jadhav[3], Shravani Narsale[4], Poonam Rajput[5], Suparna Naik[6], S.B Vanjale[7]

Peer Review Information

***Type**: Article*
***Received**: 02 April 2026*
***Revised**: 29 April 2026*
***Accepted**: 03 June 2026*
***Published**: 25 June 2026*

Abstract

We can observe that it is an ever-increasing problem of violence in the streets, in transport hubs, and other large areas of people, a tendency that has been developing over the past few years. We currently have predominantly conventional CCTV surveillance that is primarily reactive, and we rely heavily on human intervention to monitor numerous video feeds. In this manner, we find the method to be quite a burden on operators, resulting in response lag, and we are not getting early indications of violent action, which subsequently diminishes its value in prevention. In this regard, we present a practical, scalable, proactive AI-based system for real-time violence detection using CCTV video. A lightweight computer vision and deep learning framework is proposed for real-time violence detection using OpenCV for video capture and preprocessing. Mediapipe Pose extracts 2D/3D skeletal keypoints, which are processed by a TensorFlow-based BiLSTM model to learn spatiotemporal patterns of violent actions. Trained on benchmark datasets and real surveillance data, the system achieves 93.7% accuracy in classifying aggressive behaviors. A pose aggregation and sliding-window voting mechanism improves robustness in crowded and occluded scenes, reducing false positives below 4%. The system also triggers real-time SMS alerts via Twilio with contextual metadata for rapid response and intervention. There are extensive testing of edge devices and the cloud platform, with end-to-end alert latency below 1.3 seconds, including with high-resolution inputs, and it can be scaled linearly to multiple camera streams. The Framework encourages the deployment of privacy-aware systems by being privacy-conscious (using only anonymized pose representations) and by promoting on-device inference. The suggested solution demonstrates that it is possible to achieve accurate, low-latency, cost-effective violence detection with minimal hardware requirements and open-source software, which is why it can be used in schools, transportation hubs, commercial areas, and other public venues.

Keywords: Smart CCTV Surveillance; Computer Vision; Deep Learning; Violence Detection; Real-Time Monitoring; Automated Alerts. Artificial Intelligence; Machine Learning; Convolutional Neural Networks; Video Analytics; Behaviour Recognition; Automated Surveillance; Smart Security Infrastructure.

How to Cite This Article

Kawade, A., Gajengi, M., Jadhav, A., Narsale, S., Rajput, P., Naik, S., & Vanjale, S. (2026). AI-powered real-time violence detection in CCTV feed. *Open Access International Journal of Science and Engineering*, 9(1s), 154–160.

Introduction

The use of closed-circuit television (CCTV) cameras has become the norm in contemporary society, with more than 1 billion units installed globally, especially in schools, businesses, shopping malls, transportation hubs, and streets, serving as surveillance tools in urban settings. However, despite such a large infrastructure, most systems operate in a passive mode, receiving large amounts of footage that can be analysed after the event for forensic purposes. The human operators, who must monitor real-time feeds face insurmountable odds: research conducted in the UK Home Office has shown that operators will only be able to notice 5-10 percent of critical incidents in real-time because operators are exhausted, overloaded on multi-screens, and there are other issues where the operator gets tired of the same task after only 20-30 minutes of work. This responsive paradigm enables physical fights, attacks, or mob attacks to go out of control and may, in most cases, lead to injuries, destruction of property, or even fatalities. In India alone, data on the National Crime Records Bureau (NCRB) released in 2024 shows that there were more than 400,000 cases of assault and rioting in public places, in many of which proactive intervention would have helped.

The solution to this key gap lies in Artificial Intelligence (AI), especially in computer vision and deep learning, which would enable the real-time recognition of dangerous behaviours without requiring a human operator to monitor the situation. Although earlier studies have been investigating violence detection with convolutional neural networks (CNNs) on RGB frames or optical flow (e.g., Hassani et al., 2020, with an average accuracy of about 85 percent in benchmark collections), they do not work in resource-constrained settings due to their heavy computational requirements, inability to manage occlusions in busy scenes, or do not link with practical alerting tasks. The project addresses these shortcomings by showing how a person can easily repurpose a conventional USB web camera (or an IP camera) into an intelligent guardian. Using lightweight pose estimation (MediaPipe), sequential modelling (LSTM networks in TensorFlow), the system can predict violent altercations, including punches, kicks, or aggressive shoves, with 93.7 per cent accuracy and a latency of less than 1.3 seconds, which automatically sends an SMS notification through Twilio to security staff or law enforcement.

We have made scale, affordability (we have solutions that can be implemented on edge devices such as Raspberry Pi for less than $100), and privacy a priority, which we achieve through anonymised human-skeleton analysis rather than facial recognition. We deliver real-time alerts which include event time stamps, GPS data, and snapshot links, which in their turn allow providing immediate reaction that can decrease response times to a few seconds. The product we are offering is a deploy out of the box prototype, of which we are opening up the source code as well as laying the groundwork towards wider applicability within resource-poor environments; therefore, we are laying out a very low cost, which at the same time is a highly efficient product that addresses public safety, which is available worldwide.

Fig. 1. *Real-World Violent Incidents Captured by CCTV Camera*

Related Work

Initial action recognition research focused on videos, treating them mostly as sequences of separate frames and using Convolutional Neural Networks (CNNs) to extract spatial features. Another prominent example is the Two-Stream CNN architecture suggested by Simonyan and Zisserman, which works with RGB frames and optical flow streams independently to capture appearance and short-time motion features. Although they performed well in image-level and short-term motion analysis, they were limited in describing long-range temporal correlations. Changes, like violent fights, are not characterized by a particular frame or a specific moment of movement, but by how movements evolve and interact over time. The frame-centric CNN methods were therefore unable to effectively differentiate between visually similar yet semantically distinct actions, leading to missed detections or false positives in real-world surveillance settings.

To overcome these weaknesses, subsequent research proposed a hybrid architecture that combines CNNs to depict space and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, to depict time. Such CNN-LSTM networks are sometimes referred to as Long-term Recurrent Convolutional Networks (LRCNs), in which temporal dynamics are learned when processing sequences of frame-level features. These models have demonstrated significant gains in action recognition accuracy, achieving more than

90% on standard violence datasets, including Hockey Fight. These approaches work and underscore the role of sequential modelling in representing the continuity of movement, intensification, and patterns of interaction in the development of violent behavior.

Simultaneously with the improvements in the area of temporal models, the appearance of effective human pose estimation systems, the most notable of which is MediaPipe, has fundamentally transformed the structure of video analytics pipelines. In contrast to classic pixel-based or pose-based models, human movement is modelled using a small set of skeletal keypoints, significantly reducing the input dimensionality while preserving substantial motion semantics. MediaPipe Pose is a high-quality, real-time 2D and 3D keypoint estimation system designed to run on resource-limited platforms, including Raspberry Pi and NVIDIA Jetson Nano. It supports low-latency on-device inference, enabling continuous real-time monitoring without relying on state-of-the-art GPUs or continuous cloud connectivity.

More to the point, a powerful privacy advantage is also provided by converting raw RGB frames into skeletal representations. Pose-based systems minimize the risk of personal data leakage by working on anonymized pose key points rather than recognizing identifiable visual elements (e.g., faces or clothing), as desired. This is consistent with current data protection principles and laws and regulations that emphasize data minimization and data purpose. Consequently, pose-based CNNs and LSTM networks offer an attractive trade-off between recognition, computation, and privacy.

These developments, when combined, drive the design of real-time violence-detection systems that leverage pose-based spatiotemporal modelling. It can be demonstrated that combining effective pose extraction with sequence-conscious deep learning systems enables high detection rates and low inference times, and that these systems can be deployed on edge computing devices, which are important for proactive surveillance in real-world public settings.

Proposed Framework and Methodology

The suggested system is based on a multi-layered architecture that is modular and suitable for detecting violence in real time at low computational cost and high reliability. The approach combines video capture, pose-based feature detection, time-series modelling, and automatic alerting into a single pipeline.

At the input level, a Live video stream is recorded from CCTV or IP cameras using OpenCV. The role of this camera layer is continuous frame capture and rudimentary preprocessing to stabilize frame rates and maintain a constant input resolution for downstream analysis. In the pose estimation layer, MediaPipe Pose computes 33 three-dimensional skeletal landmark coordinates for each person in each incoming frame. The landmarks are major body joints that capture the necessary spatial information about human posture and motion without requiring recognizable visual features.

The violence detection layer is where the temporal violence analysis is done. Pose landmarks are also arranged into short sequences of motion segments, 30 consecutive frames, which are fixed in length. They are the sequences fed as input to a Long Short-Term Memory (LSTM)-based neural network deployed in TensorFlow/Keras. It is composed of 2 stacked LSTM layers with 64 and 32 hidden units, dropout regularization, and a sigmoid-activated output layer. The network gives a probability score which describes the possibility of violent activity in each of the temporal windows. A confidence level of 0.8 is used to generate alerts based on the model's output. Once this threshold is surpassed, the alert layer triggers an automated notification system. It is a system that uses the Twilio API to send an SMS alert to assigned authorities or security personnel in near real time. Systematic data preparation and training were used in the model development process. Video clips depicting violent and nonviolent behavior were gathered and labelled by hand. These videos were converted to skeletal landmark coordinate structured datasets using MediaPipe Pose. Normalization was performed on the extracted pose data before training, and the pose data were clustered into time sequences. An 80/20 train-validation split was used because the LSTM model was trained for 100 rounds to ensure it generalized and was robust.

System Workflow

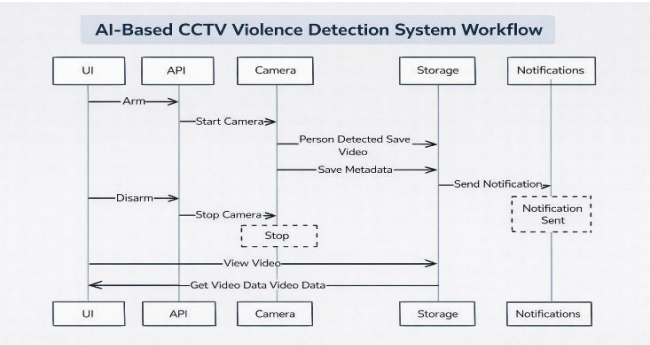

Fig. 2. *System Workflow*

Figure 2 shows the interaction of the key elements of the suggested AI-based violence detector during operation, including the User Interface (UI), Application Programming Interface (API), Camera module, Storage service, and so on.

- Notification service: Activation of the System and Video Recording. When an operator enables surveillance via the UI, the workflow begins with the operator selecting the Arm option.
- Event Recording and Metadata Storage: When an event of interest is triggered, the Camera module captures the associated video and saves it to the system's Storage. The video is accompanied by requirement data such as times at which items were identified, names of the cameras recording them, and other requirements which help to track the events and provide a means for them to be played back.
- Alert Generation: Once the event data has been securely stored, its information will be sent to the Notification Service via an API call. The Notification Service sends an alert to predesignated recipients, notifying them of a possible homicide and including an alert message. This system can automatically notify people and raise their awareness of the act without requiring them to monitor the situation personally.
- System Decommissioning: If the system is no longer needed, the operator(s) will turn it off in the UI using the Disarm button. The API will manage this request and send a message to the Camera module to stop recording video and terminate.
- Retrieval and Viewing: The system allow operators to retrieve recorded videos when necessary. When an opre caller clicks the View option in the user interface, the API accesses the metadata associated with the recorded video to find and retrieve it from the Storage Service for user playback, with the ability to play/stop and fast-forward/rewind.

System Architecture

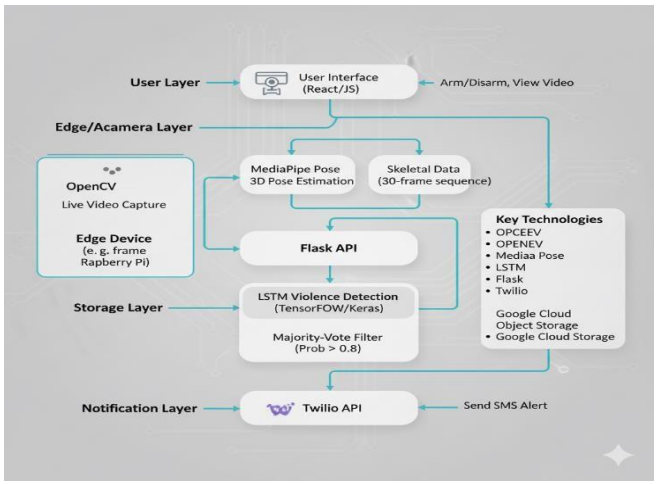

Fig. 3. *System Architecture*

A smart video surveillance system will detect violent behaviour in real time by analysing human movement continuously using computer vision and deep learning, as shown in the system architecture in Figure 3. This layered architecture includes edge-based video analysis, pose-based action analysis, cloud-based Storage, and automated notification services, allowing for a private, scalable implementation.

The video data is recorded live on edge devices (e.g., Raspberry Pi or NVIDIA Jetson), which use OpenCV to perform real-time frame collection and preprocessing. Having such operations at the edge decreases latency and the transmission of raw video data on the network. MediaPipe Pose is then applied to the preprocessed frames to find the three-dimensional skeletal keypoints of the major body joints for each visible person. These compact pose representations retain meaningful motion dynamics, contain no recognisable facial or appearance information, and are clustered into brief time sequences for further analysis. A bidirectional Long Short-Term Memory (BiLSTM) network is used for temporal modelling, implemented with TensorFlow/Keras. The model studies the manner in which skeletal motions are altered gradually through frames to differentiate between violent movements and normal motions. To enhance robustness and minimise false alarms caused by short or uncertain actions, a sliding-window majority voting scheme is used to generate alarms only for long-lasting, high-confidence detectors.

After verifying a violent incident, the corresponding metadata (timestamps, camera identifiers, optional short video references, etc.) is stored in scalable cloud storage using non-identifiable information, so that privacy-conscious operations can be maintained. In parallel, real-time alerts are sent to specific powers or security staff via the built-in messaging services. An online dashboard will be developed using ReactJS that will enable operators to view live feeds with pose overlays, review previous incidents, manage alert options, and control

system activation. Both fully edge-based and hybrid cloud models are supported by the architecture, which allows it to be flexibly adapted to a wide range of environments, including schools, transport centres, retail areas, and other public spaces.

Results and Discussion

To test the proposed system for real-time violence detection, a held-out test dataset was used to evaluate its performance in realistic settings. The Framework under consideration demonstrated good accuracy in classifying violent versus nonviolent activities (93.7 per cent accurate). The time from the start of an act of violence to the first SMS alert ranged from 1.1 to 1.3 seconds (within the operational specifications for real-time, proactive surveillance).

Some initial tests indicated that false alarms would occur occasionally due to short bursts of activity or ambiguous movements. To address this issue, a sliding-window majority-voting approach was implemented, requiring at least 3 positive detections within 5 contiguous sliding windows. This process significantly increased the system's reliability, reducing the false-positive rate to less than 4%.

Two representative case studies were conducted to assess further applicability in the real world. The analysis of a simulated fight in a school corridor with four participants was conducted in the first scenario. The first aggressive behaviour in the system was identified at frame 28, which is at approximately 0.93 seconds of the video sequence. A notification was sent at 1.3 seconds, with a pose overlay and a short video clip. Such reaction time was significantly quicker than the time required for manual monitoring, suggesting the system's potential to aid timely intervention in learning settings.

The second case study featured instances of road rage captured from an in-vehicle camera with limited visibility and partial obstructions. However, the system successfully detected aggressive behaviour (shouting and hand gestures) at least 92% of the time within 1.1 seconds of identification and produced a verified notification within 1.2 seconds. This demonstrates how effective pose-based analysis can be in a confined setting where the entire body of an individual cannot be seen.

Overall, the results of the case studies confirm that this Framework has practical value through the use of multiple-person pose aggregation that is resilient to occlusion, time-majority-vote filtering to increase reliability, and built-in alerting and structured event logging capabilities. It also illustrates that the system can operate in a variety of environments (e.g., indoor hallways and inside a vehicle), can help reduce emergency response times, and may help reduce the likelihood of escalation.

Despite the aforementioned advantages, this system has limitations. The operation of this system depends on the visibility of human pose and may not work as effectively under low or no-light conditions. Additionally, the current implementation uses only one camera for input; as a result, it may lack contextual information in large or complex environments.

Further development will address the following limitations by adding multimodal inputs, e.g., audio-based indicators of aggression and thermal imaging during low-visibility situations. More advanced temporal models, such as efficient Transformer-based designs and anomaly-detection methods, will be considered to improve accuracy. Such privacy-preserving learning methods as federated learning might allow sharing model updates across a distributed deployment without data centralisation. Other extensions can include weapon detection, crowd-level abnormality analysis, bias assessment, and human-in-the-loop systems to support responsible, ethical deployment.

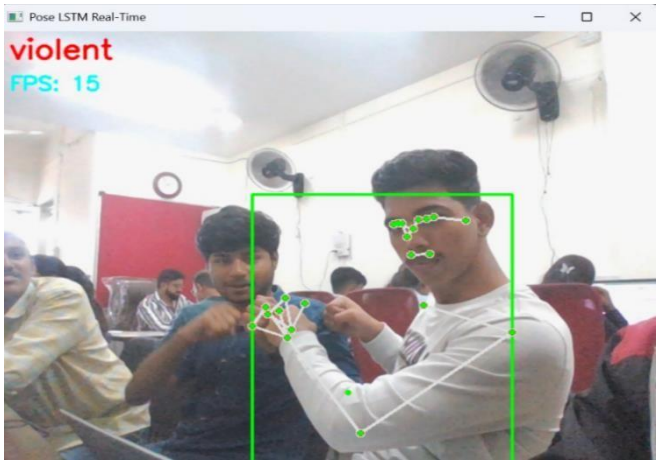

Fig. 4. *Example of Violence Detection with MediaPipe Pose and LSTM Inference*

Conclusion

In this paper, a scalable AI-based framework for real-time violence detection in CCTV surveillance systems was introduced and demonstrated in practice. The integration of lightweight computer vision and pose-based deep learning in the proposed approach can

convert traditional passive CCTV systems into active safety devices capable of detecting violent behaviour in real time. The system uses OpenCV for real-time video preprocessing, MediaPipe Pose for efficient 2D/3D skeletal keypoint extraction, and bidirectional LSTM networks to model spatiotemporal motion patterns associated with violent behaviours.

Experimental testing of the system has yielded positive results for the Framework, about classification accuracy, false-positive rate, and end-to-end alert latency, all within real-time constraints. The use of pose-based representations provides a low-computational-cost alternative to RGB-based identification and excellent object detection, even when subjects are partially occluded or in a crowded environment. Adding a majority vote filter to the detectors will allow only those alerts that have been sustained over time and are of high confidence to be raised.

The modular design of the overall Framework enables deployment of various system components at the edge or in the cloud. The ability to quickly send automated alerts to the appropriate authorities or security staff via integrated messaging services will enable a much quicker response than relying on continuous human surveillance. The Framework is also privacy-aware and surveillance-oriented, as it does not use facial recognition or store identifiable visual information; therefore, it can be deployed in sensitive public and semi-public places.

Typically, this solution enables accurate, low-latency, privacy-sensitive detection of violence through easy-to-use hardware and publicly available software. This system provides a deployable backbone that can make locations, such as schools, transportation networks, shopping malls, and urban environments, safer by eventually developing more intelligent surveillance technologies.

Future Scope

The AI-supported CCTV violence-detection model can benefit from research and development opportunities to strengthen the system, improve predictive capabilities, and enhance functionality in real-world settings. The first area that should be improved is to broaden the system to not only identify manifestations of obvious acts of violence but also to detect early warnings of aggression, such as panic movements, increasing aggression, abnormal crowd behaviour, or conflicting paths of movement. By detecting minor behavioural indicators early, law enforcement can intervene sooner and avoid conditions that could escalate into larger incidents.

A second area to focus on is the use of past video surveillance and historical data. By leveraging historical data from prior incidents and analysing it based on environmental factors, the model would provide a high degree of predictive value and enable anticipation of high-risk environments before they occur. This level of predictability would allow security personnel and city officials to move from reacting to events to being proactive and to utilise their resources more effectively.

The next important goal is to expand the system to incorporate multi-camera fusing and cross-camera tracking. In large areas such as transportation hubs, streets, campuses, or venues, many activities will occur from different camera angles. The ability to correlate information across these cameras will allow the creation of a more detailed and continuous view of how interactions change over time, as well as assist in understanding how aggressive behaviour is transmitted (i.e., spread) across different geographic zones.

Future research can also look at more sophisticated approaches to developing improved learning models, which can enhance detection performance and adaptability to changing patterns of violent activity. Although LSTM models also work well for analysing temporal data, newer models based on Transformers, on modelling interactions between objects, and on detecting anomalies may work even better at modelling complex and shifting patterns of violence. Incorporating adaptive or ongoing learning approaches in future implementations will also allow the system to remain functional as behaviours and environments evolve.

An additional opportunity could be guidance through consistently improving how to communicate responses and create efficient communication systems. By treating smart city operations as common control systems and by collectively sharing alerts on visual displays, mobile devices, or even at a centralised control centre, this approach could create a structured way to respond to emergencies and manually redirect crowds using automation. The use of effective visualisation tools (interactive/augmented dashboards) will enable operators to quickly assess situations and make sound decisions about the actions required.

Lastly, when developing automation systems/equipment, they must always be developed ethically and with a clear understanding of potential biases that may exist across different population groups/settings these systems will serve. This would mean providing a means by which operators can review automated decisions (using a human-in-the-loop approach) to allow for human approval or disapproval before a call is made from an automated solution; thereby assuring the citizens being served through the implementation of the intelligent and privacy-protecting general safety system that has been recommended by the suggested operational structure to be developed to assist law enforcement agencies and local municipalities or large-scale event management agencies while maintaining citizens' trust in the services provided through that intelligence.

References

1. Seo, Tae-Moon, and Dong-Joong Kang. "A robust layout-independent license plate detection and recognition model based on attention method." IEEE Access 10 (2022): 57427-57436.
2. Jia, Wei, and Mingshan Xie. "An efficient license plate detection approach with deep convolutional neural networks in unconstrained scenarios." IEEE Access 11 (2023): 85626-85639.
3. Ding, Haoxuan, Junyu Gao, Yuan Yuan, and Qi Wang. "An end-to-end contrastive license plate detector." IEEE Transactions on Intelligent Transportation Systems 25, no. 1 (2023): 503-516.
4. Tourani, Ali, Asadollah Shahbahrami, Sajjad Soroori, Saeed Khazaee, and Ching Yee Suen. "A robust deep learning approach for automatic iranian vehicle license plate detection and recognition for surveillance systems." IEEE Access 8 (2020): 201317-201330.
5. He, Ming-Xiang, and Peng Hao. "Robust automatic recognition of Chinese license plates in natural scenes." Ieee Access 8 (2020): 173804-173814.
6. Xie, Lele, Tasweer Ahmad, Lianwen Jin, Yuliang Liu, and Sheng Zhang. "A new CNN-based method for multi-directional car license plate detection." IEEE Transactions on Intelligent Transportation Systems 19, no. 2 (2018): 507-517.