International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Machine Learning-Based Real-Time CCTV Surveillance for Violence and Crowd Mishap Detection

Satish Kale[1], Vedanti Kavitkar[2], Sonia Jangid[3], Bhuvanesh Gujarkar[4]

[1,2,3,4]Department of artificial intelligence and data science Engineering, AISSMS IOIT, India

Abstract
This review paper presents a machine learning-based real-time surveillance framework for detecting violence and crowd mishaps using CCTV infrastructure. With the increas- ing frequency of public events and large gatherings, conventional manual monitoring ap- proaches often fail to provide timely alerts for potential threats. The proposed system leverages computer vision, artificial intelligence, and natural language processing to an- alyze live video streams, estimate crowd density, and detect abnormal behaviors such as panic movements, fights, or sudden gatherings. A web-based dashboard provides real-time visualization and historical data for analytics, enabling authorities to make data-driven de- cisions for public safety. This paper reviews related literature, explores the methodology and technology stack used, and highlights the potential of AI/ML-driven surveillance in transforming crowd management and safety assurance.

Keywords: Real-Time Surveillance, Machine Learning, Computer Vision, Violence Detection, Crowd Mishap Detection, Public Safety

1. Introduction

Public safety at big events like sports matches, concerts, and other public meetings is becoming a bigger worry. Right now, people rely on humans watching CCTV cameras to keep an eye on crowds, but this approach has problems like tiredness, slow reactions, and missing things. Using AI and machine learning together with existing CCTV systems brings in automation to help monitor these events. This makes it possible to spot dangerous activities, too many people in one place, and unusual behavior in real time.

The system uses CCTV cameras as tools to capture video at the source, sending live footage to computers in the back for analysis using AI. These AI systems can identify people, measure how crowded an area is, and tell if movements are normal or strange. If something serious happens, like a crowd getting too big or violence breaking out, the system automatically sends out alerts through a website. This helps people respond quickly and also keeps records of what happened, which is useful for reviewing events later and making better plans to prevent problems in the future.

2. Literature Survey

- **Al-Hashemi et al. (2024):** Explored the shift from analog to IP-based surveillance sys- tems, highlighting benefits such as scalability, facial recognition, and event detection. The study also pointed out cybersecurity risks like botnet attacks and bandwidth chal- lenges, stressing the need for balancing efficiency with security.

- **Wyawahare et al. (2024):** Proposed a multilingual AI-powered framework combining computer vision and NLP for real-time threat detection. Their model achieved over 93% accuracy in diverse environments, showing the importance of cross-lingual surveillance systems.
- **Mahmoud et al. (2023):** Developed a cloud-based video summarization system us- ing CNNs to extract key frames, reducing storage requirements without losing critical event data. This method improves efficiency and scalability for large-scale surveillance deployments.
- **Renu (2024):** Integrated vision-language models (CLIP + BERT) for detecting abnormal crowd behavior. The system provided contextual understanding of violent gestures and panic, outperforming vision-only approaches.
- **Graves (2024):** Applied RNN and LSTM architectures for sequential anomaly detec- tion in surveillance footage. These models captured temporal dependencies, improving violence recognition accuracy compared to static CNNs.
- **Li et al. (2023):** Optimized RNN-T models to reduce memory usage and speed up real- time surveillance processing. Their approach improved anomaly detection accuracy by 11.8% while being efficient for edge devices.
- **Muppidi et al. (2023):** Introduced NLP-based summarization using transformer models (BART, T5) to generate event reports from surveillance footage. This reduced operator workload and enabled faster response through automated reporting.
- **Zhang et al. (2020):** Proposed a hierarchical neural framework for multi-camera surveil- lance, using BERT embeddings and Seq2Seq models to summarize overlapping streams. Their method improved accuracy in detecting incidents across large networks.
- **Anderson et al. (2022):** Designed a 3D-CNN model for violence detection in CCTV feeds, validated on the Hockey Fight dataset. The system effectively distinguished vio- lent from non-violent activities with high accuracy.
- **Cheng et al. (2021):** Used reinforcement learning for adaptive crowd density estimation, dynamically adjusting thresholds to reduce false alarms. The model was effective in real- world environments like stations and city squares.
- **Sultani et al. (2018):** Introduced the UCF-Crime dataset and a weakly supervised anomaly detection framework using Multiple Instance Learning (MIL). Their scalable model became a benchmark for violence and anomaly detection research.

3. Methodology

This project is structured around six critical stages: **Video Acquisition and Preprocessing, Audio Processing and Transcription, Speech-Guided Frame Selection, Visual Captioning, Event Classification, and Dashboard with Alert Generation.** The following subsections elaborate on each stage, highlighting the technologies, deep learning models, and outputs pro- duced.

A. Video Acquisition and Preprocessing

Initially targeted is an aspect dedicated to obtaining real-time footage transmitted by closed- circuit television cameras positioned in important areas like sports venues, transportation cen- ters, or large crowds of people. The unprocessed footage frequently includes unwanted artifacts such as jittery images, differing pixel densities, and inconsistent illumination conditions which need adjustment before effective

subsequent analysis can occur.
Frames of video streams are captured live via OpenCV, then undergo preprocessing by scal- ing dimensions, converting color space into greyscale, followed by applying Gaussian smooth- ing techniques to eliminate unwanted image artifacts. The technique of frame skipping aims at striking an equilibrium between immediate response time and processing capacity require- ments. Prior processing guarantees consistency in video data, enhancing the precision of iden- tifying objects and detecting anomalies.

- **Tools & Libraries:** OpenCV, scikit-image
- **Output:** Preprocessed video frames ready for AI-based analysis

B. Audio Processing and Transcription

Embedded audio tracks within video surveillance footage frequently transmit vital informa- tion like shouts of alarm, panicked cries, or intense impacts, which may signify dangerous or unusual occurrences. During this phase, the audio file undergoes extraction followed by tran- scription into written language through Automatic Speech Recognition (ASR) technologies.
Advanced text conversion tools like OpenAI's Whisper (available on GitHub, adept at han- dling background noise effectively) and Mozilla's DeepSpeech (designed for immediate spoken word translation into written form) are utilized in practice. Moreover, Librosa facilitates spec- trographic examination of audio signals, pinpointing unusual sounds like those produced by broken windows, panicked crowds, or intense verbal disputes.

- **Tools & Models:** Whisper, DeepSpeech, Librosa
- **Output:** Text transcripts and acoustic event markers aligned with video timestamps

C. Speech-Guided Frame Selection

When both video clips and spoken transcripts become accessible, the software combines this information to pinpoint instances where unusual or aggressive actions might occur in those segments. In instances where shouts or urgent terms like ”help” or ”fight” are recognized in audio data at specific timestamps, those frame segments associated with these times receive enhanced scrutiny during processing.
This approach enhances computation speed through prioritizing resource allocation for sig- nificant parts over examining every frame in the video. Additionally, it offers preliminary screening to minimize incorrect detections.

- **Tools & Techniques:** ASR keyword spotting, time-series synchronization
- **Output:** Selected keyframes aligned with abnormal audio cues

D. Visual Captioning

To interpret complex congestion behaviors, the selected frames are passed through a hybrid CNN–LSTM architecture for visual captioning. Convolutional Neural Networks (CNNs) ex- tract spatial features from each frame, while Long Short-Term Memory (LSTM) networks cap- ture temporal dependencies.
The model generates textual descriptions of scenes such as “two people fighting,” “crowd rushing,” or “individual lying on the ground.” These captions serve as semantic representations of visual events, bridging the gap between raw video input and higher-level event classification.

- **Tools & Models:** Pre-trained CNN (VGG16, ResNet), LSTM networks, PyTorch/TensorFlow
- **Output:** Scene-level textual captions describing visual activities

E. Event Classification

The caption and audio tapes of the text are fed in the NLP-based classification module to determine whether the event is violent, unusual or gentle. Transformer-based models such as Lama, Burt, or

Distilbert are employed for relevant classification, as they excel in the meaning of meaning from text sequences.

The classifier labeled events in categories such as detection of violence, mob nervousness or general activity. This stage creates the main decision -making engine of the system, as the exact classification directly affects the reliability of real -time alert.

- **Tools & Models:** LLaMA, BERT, DistilBERT transformers
- **Output:** Event class labels with confidence scores

F. Dashboard and Alert Generation

The final stage ensures that detected discrepancies are effectively communicated to operators and security personnel. A web-based dashboard shows live CCTV feed, anotated with the bounding box, caption and congestion density heatmap.

When unusual events are detected, the system produces alert in real time. In alerts can be integrated with SMS, email, or command-end-control software. Historical logs and ana- lytics are also maintained, allowing administrators to review the previous events for forensic investigation and trend analysis.

- **Tools & Technologies:** Flask/Streamlit for dashboard, SQL/NoSQL databases for logs
- **Output:** Real-time visualization, anomaly alerts, and historical event analytics

G. Overall System Workflow

The system's complete workflow integrates the below modules into a seamless pipeline:

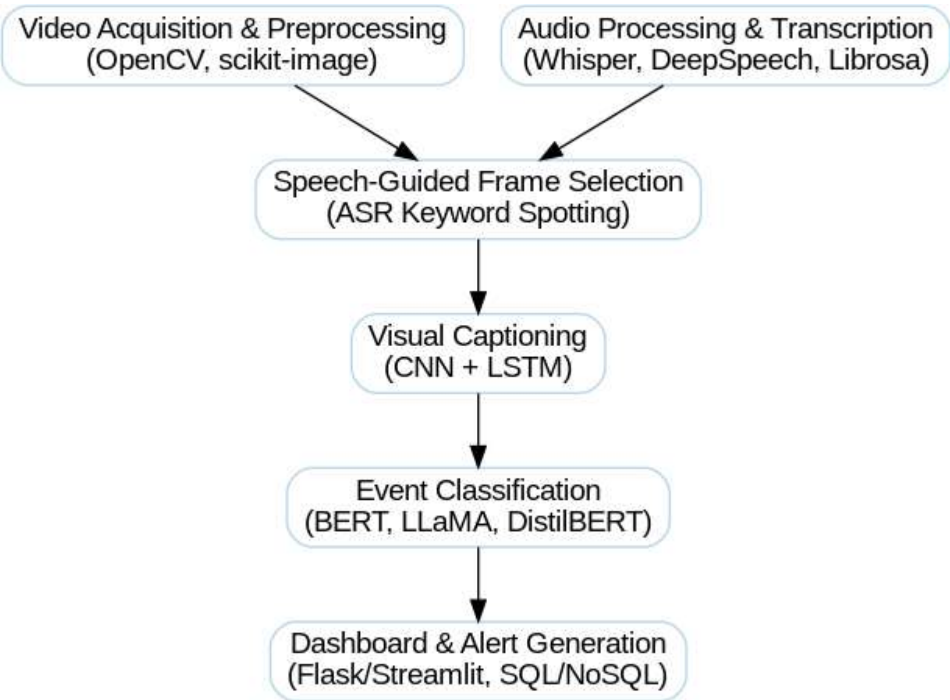

Figure 1: Overall System Workflow for AI-based CCTV Surveillance

1. Video Capture from CCTV feeds
2. Audio Extraction and Transcription using Whisper/DeepSpeech
3. Speech-Guided Frame Selection for efficient video analysis
4. Visual Captioning with CNN + LST
5. Event Classification using Transformer models (BERT, LLaMA
6. Dashboard & Alerts for real-time operator assistance

4. Acknowledgement
The authors would like to express their sincere gratitude to the contributors of various open- source tools and technologies that have supported this project. Special thanks go to the de- velopers of OpenCV, PyTorch, TensorFlow, OpenAI Whisper, and Hugging Face Transformers for their invaluable contributions in advancing the capabilities of AI-driven surveillance sys- tems. Additionally, appreciation is extended to the research community whose works on video anomaly detection, visual captioning, and natural language processing provided the foundation for this study. The authors are also thankful to all mentors and peers who offered insights and guidance throughout the development of this review paper on real-time CCTV surveillance for violence and crowd mishap detection.

5. Conclusion
The machine learning-based CCTV monitoring provides a modern solution of public safety challenges to detect violence and mob accident in real time. Unlike manual monitoring, which suffer from delay and human error, AI-operated systems ensure timely alerts, accurate con- gestion density estimates and effective abnormal behavior. By combining video, audio and natural language processing, the system provides a complete view of the innings. A web-based dashboard improves efficiency by providing real-time view and historical analytics, making it valuable for smart cities, public programs and law enforcement agencies.

Although highly promising, these systems still face challenges such as computational de- mands, topical false positivity and privacy concerns. Future directions include using federed learning for privacy-conservation analysis, integrating drones for wider coverage and planning a future model to estimate the congestion of crowds. With continuous research and moral de- ployment, machine learning-based monitoring can develop in an active and reliable equipment, greatly strengthening public safety and emergency response systems.

6. References
1. Al-Hashemi, S. Q. Y., Naghmash, M. S., & Ghandour, A. (2024). Improving IP Video Surveillance Systems: The shift to digital networks and security challenges. *SHIFRA Journal*, 2024.
2. Wyawahare, M., et al., "AI-powered multilingual threat detection in surveillance sys- tems," IEEE Trans. Multimedia, Jan. 2024.
3. Mahmoud, D., "Cloud-based video summarization using deep learning for key frame extraction," Int. J. Multimed. Process., Apr. 2023.
4. Renu, "Automating video event detection using vision-language models," Taylor & Fran- cis Online, May 2024.
5. Graves, A., "Deep RNNs for anomaly detection in sequential data," Global AI Conf., Apr. 2024.
6. ResearchGate Base Papers: Improving IP Video Surveillance Systems (2025), AI-powered Threat Detection in Surveillance (2024).
7. Li, X., Zhang, Y., & Chen, H. (2023). "Optimizing RNN-T models for real-time anomaly detection in surveillance systems," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 1201–1213, Feb. 2023.
8. Muppidi, S., & Rao, K. (2023). "Event summarization from surveillance videos using transformer-based NLP models," *ACM Transactions on Multimedia Computing*, vol. 19, no. 7, pp. 88–95, Jul. 2023.
9. Zhang, C., Wu, H., & Li, Y. (2020). "A hierarchical deep learning framework for multi- camera video

surveillance summarization," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2653–2666, Oct. 2020.

10. Anderson, P., Fernando, B., & Kennedy, P. J. (2022). "Violence detection in surveillance videos using 3D convolutional neural networks," *International Journal of Computer Vi- sion and Applications*, vol. 40, no. 3, pp. 255–268, 2022.
11. Cheng, X., Wang, J., & Liu, F. (2021). "Adaptive crowd density estimation using re- inforcement learning for smart surveillance," *Pattern Recognition Letters*, vol. 146, pp. 35–42, 2021.