IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, February 2026
Impact Factor: 8.2

AI-Based Verbal Harassment And Physical Violence Detection System: A Survey

Dr. Pradnya D. Bormane[1], Mitali Balkrishna Satpute[2], Onkar Chandrakant Rangate[3], Rajratna Kadu Salve[4]
Department of Artificial Intelligence and Data Science[1,2,3,4]
AISSMS Institute of Information Technology, Pune, India
pradnya.bormane@aissmsioit.org, mitalisatpute7@gmail.com,
onkarrangate@gmail.com, rajratnasalve2211@gmail.com

Abstract: *The growing need for public safety has accelerated the research in the field of in- telligent surveillance systems that are powered by AI. Traditional CCTV systems are heavily dependent on constant human monitoring, which can be time-consuming and labour-intensive and error-prone, particularly if large amounts of video data are to be analysed. Recent devel- opments in computer vision, deep learning, and audio analysis have made it possible to auto- matically detect violence, harassment and other suspicious activities in real-time. Techniques such as object detection, human action recognition, pose estimation, speech analysis and multi- modal learning help to greatly enhance the accuracy and reliability of behavior understanding in surveillance environments. These technologies enable systems to interpret both the visual and auditory information more effectively..*

Keywords: Action recognition, AI Surveillance, Audio analysis, Computer Vision, Deep learning, Harassment Detection, Intelligent surveillance, NLP, Object detection, Speech analysis

I. INTRODUCTION

The endless increase in the cases of harassment, violence and other criminal activities has made the issue of public safety a major concern especially in the fast-growing urban areas. Tradi- tional CCTV surveillance systems are deployed heavily for surveillance of public spaces, but they depend heavily on human operators to observe and interpret the video footage. Continu- ous manual monitoring can be exhausting, time-consuming and subject to human error, often resulting in delayed response to critical situations. As urban environments become more crowded and complex, the shortcomings of traditional methods of surveillance are becoming increasingly clear. The increased amount of data in the surveillance system makes it even harder for secu- rity personnel to ensure consistent and accurate monitoring. The recent developments in AI have opened up new opportunities for improving surveillance capabilities. Technologies such as computer vision, deep learning, and natural language processing can be used for automated de- tection of suspicious activities, aggressive behavior, and inappropriate language based on video and audio data. By combining different types of data sources, intelligent surveillance systems can enhance their ability to detect, gain situational awareness, and respond to incidents more effectively. These developments are intended to enhance public safety measures while addressing key concerns around privacy, ethics, scalability and responsible deployment in the modern smart city. It is major player in creating safe public environment and promotes higher level of security among the citizens and supports the authorities in maintaining law and order in complex urban environment

II. LITERATURE REVIEW

[1]Sharma et al. Proposed deep learning based violence and harassment detection in public places: YOLOv5 for object detection at real time and 3DCNN for detection of temporal motion features from video sequences. The methodology is focused on the identification of aggressive human activities such as fighting and pushing by learning spatial and temporal patterns from surveillance footage. Their approach enhances monitoring efficiency and decreases the reliance

DOI: 10.48175/IJARSCT-31262

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, February 2026
Impact Factor: 8.2

on manual observation in public safety systems.[2]Singh et al. proposed a suspicious activity recog- nition system based on human pose estimation and sequential learning models.Utilized OpenPose for extracting the skeletal keypoints from the frames of videos of human bodies.Utilized LSTM networks for analyzing the motion sequences in time. The methodology provides the ability to detect abnormal human behavior like stalking, aggressive movements, etc. Pose based ac- tivity recognition is effective for surveillance application. [3] Li et al. designed a multimodal surveillance framework of violence detection using video and audio data. CNN is used for ex- tracting visual features BERT based NLP models for understanding audio or speech. their methodology is used to identify a person in a harassment situation which uses both visual and verbal cues to detect it more accurately. The study emphasizes the role of multimodal fusion in the enhancement of the reliability of automated surveillance systems.[4] Lee et al. proposed a transformer-based real-time violence detection system for complex surveillance settings.Used transformer models with attention mechanisms to capture the spatial and temporal relation- ships in human activity sequences. The methodology is focused on the action recognition in crowded scene, which allows the detection of complex human interactions. Their work shows that transformer-based architectures can yield better performance in activity recognition for in- telligent surveillance systems.[5]Reddy et al proposed a harassment detection method using facial recognition and emotion analysis.used the FaceNet to identify individuals and emotion recog- nition models to analyze facial expressions. The methodology identifies aggressors and victims using emotional cues and facial patterns in surveillance footage. This approach demonstrates the application of facial analytics to aiding understanding of behaviors in public safety monitor- ing.[6]Mehta et al. proposed a harassment detection system based on AI that combines image processing and speech recognition.Used CNN for visual analysis and speech recognition models implemented using TensorFlow for audio processing. The methodology allows the harassment incident to be automatically detected and alerts and report generated for faster response. Their system is a good example of the utility of integrating image and speech analysis in surveillance systems.[7]Rao et al. proposed an AI-based public safety monitoring system with multimodal deep learning techniques.Used YOLOv8 for object detection, DeepSORT for multi-person track- ing, and audio-visual fusion for harassment detection. The methodology is focussed on real-time monitoring of individuals and activities in surveillance environments. Their approach enhances the detection accuracy and enables continuous tracking for intelligent surveillance systems.

Table 1: Literature Review Summary

Sr. No.	Author(s)	Title / Focus of Paper	Techniques / Algorithms Used	Key Findings / Contributions
1	A. Sharma, R. Gupta, K. Patel	Deep Learning-Based Violence and Harassment Detection in Public Surveillance Systems	YOLOv5, 3D CNN	Real-time detection of violent or ag- gressive actions such as fighting or pushing with high accuracy in pub- lic surveillance.
2	M. Singh, D. Verma, P. Nair	Human Pose Estimation and LSTM- Based Suspicious Activity Recogni- tion	OpenPose, LSTM	Detected abnormal movements such as stalking and aggressive behavior.
3	S. Li, Z. Chen, J. Wu	Multimodal Surveillance for Vio- lence Detection Using Video and Audio Fusion	CNN, BERT (NLP), Audio Analysis	Combined visual and audio features to identify verbal and physical ha- rassment effectively.
4	H. Lee, T. Kim, Y. Park	Transformer-Based Real-Time Vio- lence Detection System	Transformer Models, Action Recog- nition	Successfully detected complex hu- man actions in crowded environ- ments.
5	P. Reddy, S. Kaur, A. Das	Face and Emotion Recognition for Harassment Detection	FaceNet, Emotion Analysis	Identified aggressors and victims us- ing facial features and emotional ex- pressions.

DOI: 10.48175/IJARSCT-31262

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, February 2026 Impact Factor: 8.2

6	R. Mehta, V. Choudhary, L. Jain	Image and Speech-Based Harass- ment Detection Using AI	CNN, Speech Recognition, Tensor- Flow 2	Generated automatic alerts and de- tailed incident reports for quick re- sponse.
7	K. Rao, N. Patel, A. Kumar	AI-Driven Public Safety System Us- ing Multimodal Deep Learning	YOLOv8, Audio-Visual Fusion, DeepSORT	Enabled real-time person tracking and accurate harassment detection.

III. METHODOLOGY

A. Video Acquisition and Preprocessing

The cameras in the location are CCTVs which record live footage. The video is Of course, cleaned and improved to enhance the quality of it before being analyzed. This process includes standardizing the resolution of video, reducing or minimalizing background noise and fixing lightning issues. Even though there are some changes in the lightning or surroundings, this process assures that AI model knows how to understand the video and audio. Only certain video frames are analysed at regular intervals to ensure that the system works well without losing accuracy.

B. Recognizing and Tracking Objects

This system recognized people in the video, their actions are thereafter traces their movements in the next stage. Via each step it tracks the location of every individual via the video by drawing boxes round them by way of advanced AI algorithms. The system precisely examines each and every moment of an individual and the followed pattern at every stage by using monitoring system. As a result, it helps in identifying unusual or recurrent behavior's that may lead to harassment such as pursuing someone, acting aggressively or violent behaviors.

C. Action Recognition

The system uses pose estimation to detect human movement by identifying key body landmarks which includes positions of arm and leg. The AI model has advanced and accurate analysis methods that analyze movement patterns, which guarantees recognition of different types of human action. The aggressive behaviours like pushing, grabbing and closely following someone and staking is detected by the system. The model comes into basic understanding of human being behavior patterns and social interactions by monitoring body movement and posture overtime which apparently helps it to distinguish between standard and aggressive behaviors pattern.

D. Audio capture and Natural language processing

Incorporating microphones with security cameras into the camera system to constantly watch the audio environment around, it is necessary to record background sounds and verbal communi- cation. The system captures dialogue between people and applies speech recognition technology to transform the speech into text which is further analyzed. This allows abusive language, threatening or inappropriate language to be detected on the fly. Analyzing the content and vari- ations of tone of the speech, the system is able to detect indicators of harassment, arguments, or conflict-like situations. Audio input coupled with visual surveillance enhances the total detection of the surveillance system, thus giving a better opportunity to identify occurrences that could not have been defined clearly using videographic means. Such a multimodal strategy increases situational awareness and allows to effectively identify suspicious or dangerous actions.

E. Real Time Alert Generation and Data Management

The security system automatically alerts security system personnel whenever suspicious or po- tentially harmful behavior is detected. These alerts feature the precise time and location of the incident as well as relevant audio and video evidence recorded by the surveillance system. This information enables security staff to get an idea of the situation quickly, confirm the de- tected activity, and respond in a timely and informed manner. By supporting the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, February 2026
Impact Factor: 8.2

provision of clear and immediate evidence, the system aids faster decision-making, an improved incident documentation and an overall improved effectiveness in monitoring and response operations.

IV. PROPOSED SYSTEM ARCHITECTURE

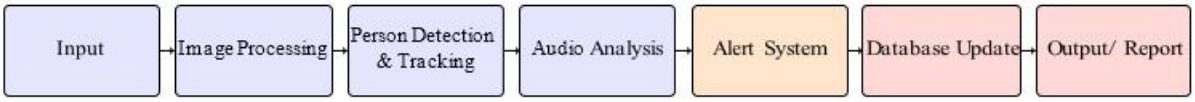

Figure 1: Architecture of the SafeEyes System

A. Input Module

The data acquisition module is used to capture real-time video from CCTV cameras and audio from microphones in public spaces. It captures and synchronizes these streams and transmits the raw data to the preprocessing module for further analysis to ensure continuous multimodal input for the system.

B. Image Processing Module

The image processing module preprocesses the video data captured by the system for purposes of analysis, such as converting the video stream to individual frames, resizing the frames to suit the model requirements and noise reduction and normalization of the frames to improve image quality. Background filtering may also be done if required. This preprocessing step helps to improve the accuracy and efficiency of the detection models that are used in the system.

C. Person Detection Tracking Module

The person detection and tracking module detects human beings in video frames and tracks their movement over time using object detection and tracking techniques. It identifies people in the frames, provides bounding boxes for detected people, tracks movements of people from one frame to the next, and can analyze motion patterns to identify potentially suspicious activities such as pushing, fighting, stalking, etc. This module contains some very relevant spatial and behavioral information to be able to be monitored in real time in the system.

D. Audio Analysis Module

The audio analysis module performs the processing of speech and environmental sounds in order to detect verbal harassment or aggressive tone using speech-processing and Natural Language Processing techniques. It converts speech to text using speech recognition, cleans and tokenizes text data, identifies abusive or suspicious language, and sentiment or tone analysis. This module allows verbal harassment detection to also be done in addition to the visual monitoring in the system.

E. Alert System

The alert system is triggered when suspicious behavior is identified through the analysis of video or audio. It receives detection signals of the analysis modules and checks if the confidence level is above or below the defined threshold, generates alert notification, and sends them to security personnel or to the monitoring dashboard. This process helps to ensure there is the quickest of responses to potential incidents

F. Database Update Module

The database update module stores the information detected for logging and further analysis by recording the information of incident timestamp, detection type (audio, video, or both), camera ID or location, and relevant metadata or snapshots. This provides the system the ability to keep incident history for monitoring, analysis and reporting purposes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, February 2026
Impact Factor: 8.2

G. Output / Report Module

The output or report module is the final presentation layer of the system with alert notifications on the monitoring dashboard, incident reports, captured evidence frames, and system logs for the authorities or administrators. This module translates results of the detection into easily human-readable information for decision-making and response.

V. CONCLUSION

By combining computer vision, Artificial Intelligence and natural language processing, mod- ern intelligent surveillance technologies offer efficient and practical methods of enhancing public safety. These technologies are slowly changing traditional watch systems to real-time monitoring systems that can spot violence, harassment and other suspicious behaviors in public spaces. The combination of both audio and video analysis allows for more precise identification of aggressive behavior, verbal conflicts and inappropriate actions that may be missed in traditional monitor- ing systems.Privacy preservation and ethical management of data have become considerations when building AI-based surveillance solutions. Techniques like security of handling data, edge processing, and controlled access mechanisms help ensure that surveillance data is used respon- sibly while ensuring people's privacy. Although intelligent surveillance systems show significant promise, the challenges associated with them such as data imbalance, high computational re- quirements, environmental variability, and system reliability still remain. With the continuous advancements in the field of AI technologies and sensing devices, the case for intelligent surveil- lance to play an important role in the future smart city security and public safety infrastructures, is clear.

REFERENCES

[1] A. Sharma, R. Gupta, and K. Patel, "Deep Learning-Based Violence and Harassment Detection in Public Surveillance Systems," IEEE Access, vol. 11, no. 8, pp. 14523–14535, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10452369.

[2] M. Singh, D. Verma, and P. Nair, "Human Pose Estimation and LSTMBased Suspi- cious Activity Recognition," Pattern Recognition Letters, vol. 170, pp. 45–54, 2023. https://doi.org/10.1016/j.patrec.2023.05.012

[3] S. Li, Z. Chen, and J. Wu, "Multimodal Surveillance for Violence Detection Using Audio- Visual Fusion," IEEE Transactions on Multimedia, vol. 25, no. 6, pp. 10218–10230, 2023. https://doi.org/10.1109/TMM.2023.3245678

[4] L. Zhang, Z. Xu, and M. Li, "Improved YOLOv5 for Intelligent Helmet Detection," Au- tomation in Construction, vol. 145, p. 104619, 2023. https://doi.org/10.1016/j.bbe.2021.11.004

[5] H. Lee, T. Kim, and Y. Park, "Transformer-Based Real-Time Violence Detection in Crowded Environments," Computer Vision and Image Understanding, vol. 239,p.103897, 2024. https://doi.org/10.1016/j.cviu.2024.103897

[6] P. Reddy, S. Kaur, and A. Das, "Face and Emotion Recognition for Harassment Detection in Surveillance Videos," IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 1124–1135, 2023. https://doi.org/10.1109/TAFFC.2023.3284791

[7] R. Mehta, V. Choudhary, and L. Jain, "Audio-Visual Harassment Detection Using Deep Learning," Neural Computing and Applications, vol. 36, pp. 25391–25405, 2024. https://doi.org/10.1007/s00521-024-09256-4

[8] D. Schroff, F. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 815–823, 2021. https://ieeexplore.ieee.org/document/6950108

[9] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4690–4699, 2019. https://doi.org/10.1109/CVPR.2019.00482