

A Multimodal AI-Based System for Real-Time Harassment and Violence Detection Using Surveillance Cameras

Pradnya D. Bormane¹, Mitali Balkrishna Satpute², Onkar Chandrakant Rangate³,
Rajratna Kadu Salve^{4,*}

Abstract

Public safety in urban settings, as these areas are crowded and experience increasing incidents of harassment, verbal abuse, and physical violence. Traditional CCTV surveillance systems heavily rely on human operators who have to continuously monitor the video feeds, which is inefficient and prone to human error. AI-based harassment and violence detection systems automatically monitor public environments in real-time using video and audio analysis. The system is based on a multilayer guardrail architecture that includes the use of several artificial intelligence models to enhance the reliability of the detection. The proposed system identifies three major threats: physical violence, verbal abuse, and inappropriate physical contact in crowded public places. Video analysis is done with the SlowFast R50 deep learning model; pose-based motion analysis is implemented with MediaPipe with velocity tracking, and speech toxicity detection is done with a multilingual DistilBERT classifier. Additionally, object detection using YOLOv12 is used to analyze proximity between individuals to determine suspicious physical contact. The system combines these layers of detection through a fusion engine that computes a final risk score and triggers alerts in cases where the risk is above a certain threshold. When an incident is detected, the system captures a video clip, takes an incident report, and determines the nearest police station to respond to an emergency. Experimental evaluation is found to have a good performance, with 97% accuracy for violence detection and a 92% F1 score for multilingual toxicity detection. The system shows promise for AI-driven surveillance for public safety improvement and lessening of reliance on manual surveillance.

Keywords: Action recognition, artificial intelligence, computer vision, deep learning, harassment detection, object detection, violence detection, video surveillance

INTRODUCTION

Public places such as bus stations, railway stations, markets, and public transportation systems are

*Author for Correspondence

Rajratna Kadu Salve

E-mail: rajratnasalve2211@gmail.com

¹Assistant Professor, Department of Artificial Intelligence and Data Science, AISSMS Institute of Information Technology, Pune, Maharashtra, India

²⁻⁴Student, Department of Artificial Intelligence and Data Science, AISSMS Institute of Information Technology, Pune, Maharashtra, India

Received Date: April 03, 2026

Accepted Date: June 27, 2026

Published Date: June 29, 2026

Citation: Pradnya D. Bormane, Mitali Balkrishna Satpute, Onkar Chandrakant Rangate, Rajratna Kadu Salve. A Multimodal AI-Based System for Real-Time Harassment and Violence Detection Using Surveillance Cameras. Journal of Advancements in Robotics. 2026; 13(2):8–21p.

places where people get harassed, verbally abused, and physically beaten. Ensuring safety in these environments is a major challenge because of the large number of people and the constant activities that take place within crowded environments. Traditionally, surveillance systems using Closed-Circuit Television (CCTV) monitoring such environments have been used. However, conventional surveillance systems require constant observation by human operators to identify suspicious activities. This manual monitoring strategy is both time-consuming and inefficient and is also fraught with human error, as operators can easily overlook critical incidents because of fatigue or simply the large number of cameras that must be monitored at once.

With advances in artificial intelligence and computer vision technologies, modern surveillance systems can now automatically analyze video and audio streams in real time. AI-based surveillance systems help reduce the dependency on manual surveillance by allowing the system to automatically identify suspicious behaviors and abnormal activities. These systems can process large amounts of data quickly and accurately and can provide faster responses to potential threats, making them more effective public safety monitoring systems.

An AI-based surveillance system that can detect physical violence, toxic speech, and inappropriate physical contact as they occur. The proposed system is based on a multilayer guardrail architecture, where multiple detection models monitor different aspects of the environment independently to enhance the reliability of detection. The architecture has a number of modules such as video-based violence detection, pose-based motion analysis, speech toxicity detection, and proximity-based harassment detection. Each module analyzes a specific type of input and provides independent signals in terms of possible threats.

The outputs of these detection modules are combined using a fusion engine that computes a unified risk score for the monitored scene. When the calculated risk score is higher than a predefined threshold, the system automatically generates an alert, captures the relevant video evidence, and creates a detailed incident report. By incorporating multiple detection mechanisms based on artificial intelligence, the system enhances the detection accuracy and greatly lowers the number of false detections in comparison to the traditional surveillance approach of using a single model. This integrated approach allows for more reliable and efficient monitoring of public environments, leading to improved safety and quicker response to incidents.

LITERATURE REVIEW

Violence Detection Using Video Surveillance

Violence detection from video surveillance has become an important research area for the betterment of public safety and automated surveillance systems. Traditional surveillance systems depend on human operators to constantly monitor multiple camera feeds, which can be inefficient and susceptible to human error. To overcome these limitations, researchers have planned to use self-learning and deep learning-based methods in which surveillance footage is automatically analyzed for aggressive behaviors such as fighting, hitting, or other suspicious activities. A surveillance system that can detect violence and harassment in public spaces using deep learning has been proposed to demonstrate the efficacy of video analysis using automated techniques [1].

Several studies have been conducted on human activity recognition and motion analysis techniques for detecting suspicious behavior. A system that integrates human pose estimation and Long Short-Term Memory (LSTM) networks to analyze body movements and detect abnormal activities [2]. By monitoring the joint positions and movement patterns of human participants, these methods can be used to effectively detect violent human actions in surveillance settings. Recent research has also focused on the development of multimodal approaches incorporating multiple sources of information to increase detection accuracy.

An audio-visual surveillance system combines video analysis with audio signals to detect violent incidents more reliably [3]. By combining visual and speech information, multimodal systems can better understand complex situations in real-world environments and reduce the number of false detections.

Human Activity Recognition in Surveillance Systems

Human Activity Recognition (HAR) plays an important role in intelligent surveillance systems to identify suspicious or abnormal behaviors in public environments. By analyzing human movements and interaction patterns, HAR techniques can help automated systems recognize certain activities, such as fighting, pushing, or other aggressive actions. These systems generally rely on computer vision and deep learning models to extract spatial and temporal features from video sequences to classify various types of human activities.

A suspicious activity recognition system combines human pose estimation and Long Short-Term Memory (LSTM) networks [2]. The system analyzes the position of body joints and body motion patterns and identifies the abnormal action within the surveillance footage. By learning temporal dependencies in human movements, the model can differentiate between normal daily activities and potentially dangerous behaviors. Recent approaches have further improved the accuracy of activity recognition by using advanced deep learning architectures that can capture complex motion patterns in crowded environments. These techniques increase the dependability of automated surveillance systems and contribute to the real-time monitoring of public spaces for better safety and security.

Pose-Based Behavior Analysis

Pose-based behavior analysis is extensively adopted in surveillance systems to interpret human behaviors from the analysis of body motions and joint positions. Pose estimation methods identify important body landmarks, such as wrist, elbow, shoulder, and head positions, enabling the system to track human motion in real time. By analyzing the relationships and patterns of movement between these points of the body, surveillance systems can detect abnormal or aggressive acts, such as hitting, pushing, or other violent gestures.

Human pose estimation along with Long Short-Term Memory (LSTM) networks for detecting suspicious activities in surveillance videos [2]. Their approach tracked the movements of body joints for a series of frames and analyzed the temporal motion patterns to detect abnormal human behavior. This approach is used to differentiate between normal behavior and potentially violent activities based on the analysis of changes in body posture and movement. Pose-based analysis enhances the reliability of automated surveillance systems because it is based on human motion dynamics and not only visual appearance. This reduces false detections and promotes the ability of intelligent monitoring systems to identify suspicious behaviors in real-time surveillance environments.

Table 1. Comparison between the existing systems and the proposed system.

System	Techniques used	Features analyzed	Limitations	Proposed system improvement
CNN-Based Violence Detection System	Convolutional Networks for classification video Neural frame	Visual from frames features video	Cannot effectively capture temporal motion patterns and often produces false positives.	Uses SlowFast R50 architecture to analyze both spatial and temporal information in video sequences
Two-Stream Action Recognition System	Spatial and temporal streams for action recognition	Appearance and motion features	Requires high computational resources and does not consider audio or proximity information	Integrates video, pose, audio, and proximity analysis to improve detection accuracy.
Pose-Based Activity Recognition System	Human pose estimation and keypoint tracking	Body joint movements	Detects only body motion and cannot analyze speech or interaction context	Uses MediaPipe pose with wrist velocity tracking to detect slapping or punching actions
Audio Toxicity Detection System	Natural Language Processing (NLP) models for text classification	Speech and patterns content language	Focuses only on verbal communication and ignores physical violence	Combines speech toxicity detection with visual surveillance analysis
Object Detection Surveillance System	Object detection models for identifying people in CCTV footage	Presence and location of individuals	Cannot detect suspicious interactions or harassment scenarios	Uses YOLO-based detection with proximity and gender estimation
Proposed System	SlowFast R50, MediaPipe Pose, YOLO object detection, DistilBERT toxicity detection, Fusion Engine	Video behavior, human motion, speech toxicity, and proximity interactions	Minimizes through analysis limitations multimodal	Provides multimodal harassment and violence detection with real-time alerts and automated reporting

Speech Analysis and Toxic Language Detection

Speech analysis has become an important constituent of intelligent surveillance systems for detecting verbal harassment and abusive language. By converting spoken words into text using speech recognition techniques, automated systems can analyze the content of conversations and identify toxic or offensive language. Natural Language Processing (NLP) models are frequently employed with transcribed text to label the text as toxic or non-toxic, making it possible to detect verbal abuse in public settings. Recent research has focused on combining speech recognition with deep learning-based text classification models to improve the accuracy of detecting toxic language.

An audio-visual harassment detection system combines speech analysis and video processing to detect abusive interactions in surveillance scenarios [4]. By analyzing verbal and visual cues, such systems allow for a higher detection rate of harassment incidents. The combination of speech analysis and surveillance technologies makes it possible to monitor verbal communication in public spaces in real time. This approach is used to better identify threatening or abusive language and can help quickly respond to incidents of verbal harassment.

Limitations of Existing Systems

Although a large number of research studies have hypothesized intelligent surveillance systems for the detection of violence and harassment, there are still a great number of limitations in the current approach (Table 1). Many systems are based on a single detection method (video analysis, pose estimation, or speech recognition). This can affect the overall reliability of the system because relying on a single source of information can lead to inaccurate predictions or the possibility of missing an incident in complex, real-world environments.

Another drawback is that some systems only focus on visual analysis and do not include audio information for the detection of verbal abuse or toxic speech. As a result, verbal harassment incidents may go undetected. In addition, challenges such as crowded environments, lighting variations, and normal human activities can lead to false positives in automated surveillance systems. Therefore, there is a need for more advanced surveillance frameworks that use a combination of multiple detection techniques to enhance their accuracy and reliability. The integration of video analysis, motion analysis, and speech-based detection methods can help overcome the limitations of existing systems and provide more effective monitoring of public environments.

Object Detection for Crowd Monitoring

Object detection is an important part of surveillance systems for detecting and tracking individuals in crowded public places. Deep learning-based object detection models are widely used to detect the presence of people in video frames and analyze the spatial relationships between them. By creating bounding boxes around identified individuals, these systems can be used to monitor crowd density, movement patterns, and interactions among people in real time. Zhang et al. [5] presented an improved YOLO-based object detection model for intelligent monitoring applications. Their approach improves the accuracy and speed of detection, making it suitable for real-time surveillance systems. Object detection models, such as YOLO, can quickly detect multiple objects in a frame, which is necessary when monitoring a crowd. By analyzing the distance between detected individuals, object detection techniques can also be used to detect suspicious interactions and abnormal crowd behaviors. This capability supports automated surveillance systems in identifying potential safety risks and enhancing the effectiveness of crowd monitoring in public areas.

Multimodal Surveillance Systems

Suspicious activity recognition system based on human pose estimation and Long Short-Term Memory (LSTM) networks [2]. The system tracks the joints of the human body and analyzes the movement patterns of the joints over time to identify abnormal or aggressive behaviors. By using pose keypoints such as the wrists, elbows, and shoulders, the model can detect actions such as punching, pushing, or sudden physical movements. This is a way to improve the recognition of activities by not only analyzing raw video frames but also focusing on human motion.

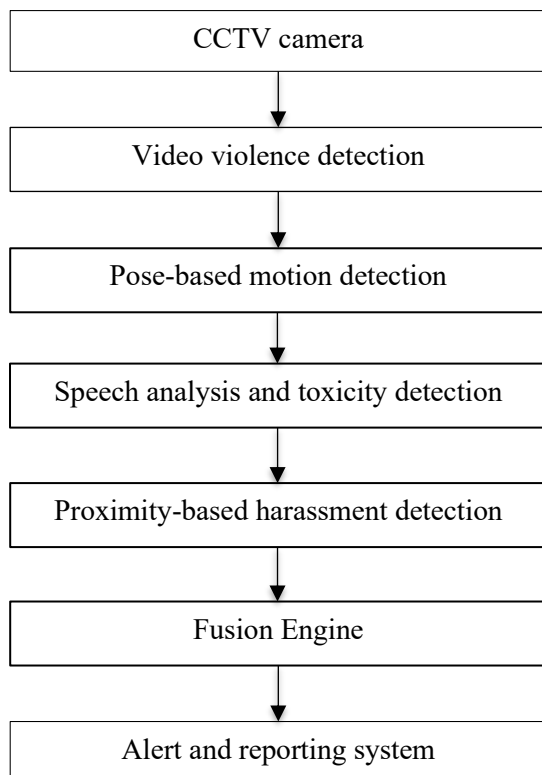


Figure 1. System architecture.

Multimodal surveillance based on audio and visual information for violence detection [3]. The system performs audio–video fusion by combining sound signals and video features to comprehend interactions between individuals in surveillance environments. This multimodal approach yields a higher accuracy in detection than single-modality systems because it considers not only the physical actions but also the audio cues surrounding them. However, the fusion of different data sources complicates the process and requires effective fusion strategies for real-time use [4, 6–10].

PROPOSED SYSTEM ARCHITECTURE AND Methodology

The system is designed to automatically detect harassment, violence, and toxic behavior in public spaces using artificial intelligence. The system takes video and audio feeds from surveillance devices and analyzes them using various detection modules. These modules collaborate via a fusion mechanism to create alerts and incident reports when suspicious activity has been detected (Figure 1).

CCTV Camera

The system starts with a video input from CCTVs mounted in public places, such as transport hubs, markets, and streets. These cameras constantly monitor the environment and provide real-time video streams. The captured footage is a primary source of data for analyzing human behavior and possible incidents.

Video Violence Detection

The Video Violence Detection module is used to analyze video frames captured from CCTVs to detect aggressive or violent activities in real time. The system uses the SlowFast R50 deep learning model to process video clips and classify them as violent or nonviolent. This model captures both spatial and temporal features from frame sequences [11–15]. The slow pathway is more engaged in spatial information such as body positions and context of the scene, whereas the fast pathway responds to fast motion patterns that may be indicative of sudden aggressive actions. The system extracts frames from the incoming video stream and analyzes the motion patterns and interactions of

people in the scene. Using models trained on datasets related to violence, the module can identify behaviors such as fighting, pushing, or sudden aggressive movements. By continually monitoring these patterns of motion, the system is able to pick out unnatural activities that may signal violent incidents. To limit false alerts, a confidence threshold of 70% was set to ensure that only high-confidence predictions were made of violence. When suspicious behavior is detected, the module produces a confidence score for the likelihood of violence. This output is then sent to the fusion engine, where it is then combined with the results from other detection modules to improve the overall accuracy of threat detection and reduce false alarms [16–20].

Pose-Based Motion Detection

The Pose-Based Motion Detection module analyzes human body movements in videos and identifies suspicious and/or aggressive behaviors. It is implemented with MediaPipe to detect and track 33 human body landmarks, such as wrists, shoulders, elbows, head, and facial points. By tracking these keypoints between video frames, the system can understand how people move and interact in the monitored environment. The module is especially limited to the movements of hands in relation to the face and upper body to detect possible slapping or punching gestures. To improve the accuracy of detection and avoid false positives from normal gestures, the system tracks wrist velocity. It calculates the speed of wrist movement between consecutive frames and analyzes whether the movement represents a violent action.

Predefined velocity values are used to categorize aggressive gestures: Slap detection at 300 pixels per second and Punch detection at 400 pixels per second. Only those movements that are larger than these thresholds are considered potentially violent. This filtering mechanism greatly helps minimize the misclassification of normal human actions, such as waving or casual hand movements. The identified patterns of movement and confidence scores are then sent to the fusion engine, where they are combined with data from other models, such as video violence detection and speech toxicity analysis. This multimodal integration helps improve the overall accuracy of the system and aids in minimizing false alarms.

Speech Analysis and Toxicity Detection

The Speech Analysis and Toxicity Detection module detects verbal abuse, offensive language, and toxic speech in the monitored environment. Audio signals from microphones are initially analyzed using a speech-to-text system that converts spoken language into text [21–25]. The system supports multiple languages such as English, Hindi, and Marathi, which helps it analyze the speech in diverse public settings. After transcription, the text was analyzed using a multilingual DistilBERT toxicity classifier, a Natural Language Processing (NLP) model trained to identify abusive or harmful language patterns.

The classifier was trained using a dataset of 30,000 samples equally distributed among the three supported languages. This training enables the system to successfully detect verbal harassment, threats, and offensive speech in various linguistic contexts. The model achieved a macro F1 score of 92%, indicating that it was capable of detecting toxic and abusive speech. Based on the analysis, the system determines a toxicity score that indicates the probability of harmful communication. This score is then passed to the fusion engine, where it is combined with the output from other modules, such as video violence detection and pose-based motion analysis. By combining both audio and visual information, the system helps to increase the accuracy of detecting incidents involving verbal harassment or aggressive communication.

Proximity-Based Harassment Detection

The Proximity-Based Harassment Detection module detects suspicious or uncomfortable interactions between people by analyzing the physical distance and contact patterns between two individuals. The system uses an object detection model to detect and track people in each video frame captured from the CCTV cameras. For each individual detected, the model produces bounding boxes representing the location of the individual in the frame. Using these bounding boxes, the system calculates the

proximity scores between pairs of individuals to determine how close they are to each other. It constantly tracks the distance and time of interactions. If two people are together for more than three seconds, the system identifies the situation as potentially suspicious [26–30].

To improve detection accuracy, the system's interactions in upper body zones are given more importance, as interactions in these areas may indicate possible harassment or inappropriate physical behavior. By examining both the distance and duration of interaction, the module can identify behaviors such as unwanted touching, stalking, or perimeter trespassing, especially in a crowded setting. The results of this proximity analysis are then sent on to the fusion engine, where it is combined with outputs from other modules, including video violence detection, pose-based motion analysis, and speech toxicity detection. This multimodal integration makes the system more adept at accurately detecting harassment incidents and minimizing false detections.

Fusion Engine

The Fusion Engine is the central part of the system that integrates the outputs from all the detection modules comprising Video Violence Detection, Pose-Based Motion Detection, Speech Analysis and Toxicity Detection, and Proximity-Based Harassment Detection. Each module examines a different aspect of the environment and provides a confidence score that indicates the likelihood of suspicious or harmful activity. The fusion engine uses a weighted decision-making mechanism to combine these results to compute the final risk score for the monitoring scene. The weighting factors assigned to each of the detection signals are: Video violence analysis: 50%, Pose-based motion analysis: 35%, and Audio toxicity detection: 15%. Integration of information from video behavior, body movements, speech content, and human proximity allows the system to make more reliable and accurate predictions than relying on a single detection method. To make the system more stable, temporal smoothing was applied, which means that the system does not experience sudden fluctuations in the risk score owing to momentary changes in detection. If the final risk score is higher than a predefined threshold, the fusion engine sends the outcome to the Alert and Reporting System, which in turn triggers notifications and logs the incident for further investigation.

Alert and Reporting System

The Alert and Reporting System is responsible for notifying authorities and documenting detected incidents. When the fusion engine concludes that the overall risk score exceeds the predefined alert threshold, an alert process is immediately triggered by the system. These alerts can be sent to security personnel, monitoring centers, or law enforcement agencies via notifications, alarms, or monitoring dashboards to allow for a quick response. Once the alert is triggered, the system automatically performs several tasks. It captures a video clip of the detected incident, related timestamps, and the results of the detection achieved by each of the analysis modules. Additionally, the system can also be used to determine the closest police station to the geographic location so that authorities can respond more quickly to the situation. These reports were also saved for further reference and analysis. The reporting feature helps authorities review incidents, analyze patterns of unsafe behavior, and support investigations as needed. By triggering both real-time alerts and structured reports, the system facilitates better response times and helps boost overall public safety monitoring.

IMPLEMENTATION

Datasets

The system utilizes multiple datasets for training and evaluation are as follows.

Real Life Violence Situation (RLVS) Dataset

The Real Life Violence Situations (RLVS) dataset is the main dataset for training the video violence detection model in the system. This dataset is widely used in research for building automated violence detection systems in surveillance environments. It contains 2,000 short video clips, 1,000 short violent videos, and 1,000 short nonviolent videos, which are collected from real-world surveillance footage. The dataset consists of realistic images recorded from CCTV cameras in public environments. Violent

clips contain activities such as physical fights, punching, slapping, kicking, and crowd fights, whereas non-violent clips contain normal human activities such as walking, talking, sitting, and crowd movement/public interactions.

Before training the model, several preprocessing steps were performed to prepare the data. Each video was first converted into frames, and 32 frames were sampled from each video clip to represent the motion sequence. The extracted frames were then resized to 224×224 pixels to suit the input requirements of the deep learning model. In addition, the pixel values were normalized between 0 and 1 to enhance the stable training of the model. The dataset was then split into training, validation, and testing sets to ensure that the model's performance was properly evaluated. This processed dataset was then used to train the SlowFast R50 video classification model, which combines information from spatial features (visual appearance) and temporal features (motion patterns) to accurately differentiate between violent and non-violent activities in surveillance videos.

Multilingual Toxic Speech Dataset

To identify abusive or threatening speech, a custom multilingual dataset was created comprising text samples in English, Hindi, and Marathi. The dataset has a total of 30,000 samples, of which 10,000 are from each language. Each language dataset was carefully balanced and contained 5,000 toxic and 5,000 non-toxic sentences, so that there were equal numbers of the two classes. This 50:50 balance helps prevent any model bias, and the reliability of toxicity detection improves. The English part of the dataset was gathered from sources such as the Google Civil Comments (Jigsaw) dataset, other profanity datasets, and manually curated toxic phrases. These samples can contain toxic sentences such as "I will beat you" or "Shut up idiot" as well as non-toxic conversational sentences such as "Where are you going?" or "Let's meet tomorrow."

The Hindi dataset was obtained from publicly available resources, such as the Paul/hatecheck-hindi and textdetox/multilingual toxicity datasets, which consist of samples of abusive and non-abusive language used in Hindi communication. For Marathi, as there are limited publicly available toxicity datasets, a custom dataset generation script was developed to generate both safe and toxic phrases. The safe category consists of topics that we use in everyday conversation, that are greetings, conversations in daily routine, journeys, food, bus words, and some casual talking, while the toxic category consists of insulting words, threatening words, harassment words, bad words, abusive expressions, etc. This multilingual dataset was used to train a multilingual toxicity classifier based on the DistilBERT model using the DistilBERT-base-multilingual-cased model from the Hugging Face Transformers framework. The model is used to perform binary classification of text, where the text is classified as either toxic or non-toxic, which means that the system can detect verbal harassment in various languages.

Harassment Detection Dataset

Unlike the violence detection module, the harassment detection component is not based on a specific dataset. Instead, it uses a pretrained YOLOv12 model for object detection and tracks people in video frames. The YOLOv12 model was trained using the COCO (Common Objects in Context) dataset, a large-scale dataset commonly used for object detection tasks. The COCO dataset has around 330,000 images with about 2.5 million annotated instances of objects belonging to 80 different object classes. In the system, the model uses the person class mainly to identify and locate people in the surveillance footage. Once people are detected, the system undergoes further analysis to spot potential harassment scenarios by analyzing the spatial and temporal patterns of interactions between people. This includes analyzing the bounding box overlap, calculating the distance between detected individuals, and monitoring the length of time that detected individuals are in proximity to each other. If two people are in close contact with each other for longer than three seconds, the interaction may be suspicious, depending on the context. These extra features enabled the system to spot potential inappropriate proximity or harassment scenarios in crowded settings without a specific harassment training dataset.

Pose Detection Dataset

The pose detection module in the system is based on MediaPipe Pose, a pretrained framework developed by Google for real-time human pose estimation. Instead of training a new pose detection model from scratch, the system is based on MediaPipe's model, which has already been trained on large-scale human pose data that contains a variety of images and videos of people performing different kinds of movements and activities. This pretrained model can accurately identify and track key points of the human body in real time. MediaPipe Pose is a model that detects 33 body landmarks that represent important joints and body parts, such as the nose, shoulders, elbows, wrists, hips, knees, and ankles. These landmarks capture detailed information on the posture and movement of the individuals in the video frame. In the system, these landmarks are detected and used to analyze the movement of the hands with respect to the face and upper body. By recording the wrist's position in one frame and the next, the system can calculate wrist velocity and movement patterns, and these are useful for determining rapid movements such as a slap or punch action. This pose-based analysis augments the ability of the system to detect violent behavior because, in addition to visual scene information, human body dynamics are understood.

Programming Language: Python

Python was the primary programming language used to develop the system. It is widely used in artificial intelligence, machine learning, and computer vision applications owing to its simplicity and powerful libraries. Python enables the easy integration of various modules, such as video processing, deep learning models, speech recognition, and alert generation. Python manages the entire system process. It captures video and audio input and processes the data using AI models to generate alerts when suspicious activity is detected. Python is also used for GPU acceleration (such as PyTorch), which enables deep learning models to run at a faster speed. Due to the huge support and documentation, Python makes it easier to develop complicated AI-based applications like real-time surveillance systems.

Video Processing: OpenCV

OpenCV (Open Source Computer Vision Library): Used for video input and real-time image processing. In OpenCV, the video frames are captured by the library and processed continuously. Each frame is extracted and sent to different AI models to perform various types of analyses, such as violence detection, pose detection, and object detection. OpenCV is also responsible for maintaining a buffer of the last few seconds of recorded video footage. When a threat is detected by the system, the buffered video is stored as an incident clip. This clip can be used as evidence or for further analysis. OpenCV is commonly used in computer vision applications, offering quick processing speed, real-time video streaming capability, and compatibility with Python-based AI frameworks.

Video Violence Detection: SlowFast R50

The Video Violence Detection module is the first layer of the surveillance system and is charged with the task of detecting physical violence from real-time video streams. This module is based on a video classification model called SlowFast R50, which was developed by Facebook AI Research. The model examines brief sequences of video frames taken from the camera and decides whether the activity under observation is violent or non-violent. The SlowFast architecture consists of two parallel processing pathways, namely the Slow pathway, which processes frames at a lower frame rate to capture detailed spatial features such as body posture and scene context, and the Fast pathway, which processes frames at a higher frame rate to capture rapid motion patterns such as punches, kicks, or aggressive movements. By combining these two pathways, the model can understand successfully the appearance of the scene and the dynamics of human motion successfully, making it suitable for the detection of violence in surveillance footage.

Deep Learning Models: PyTorch

PyTorch is a deep-learning library used for developing and deploying neural network models. It is known for its flexibility and has good GPU support. In this project, PyTorch is used to implement deep

learning models for tasks like violence detection and toxicity classification of speech. PyTorch enables models to be used for real-time inference with GPU acceleration, which is required for surveillance systems that handle constant video feeds. The violence detection model in the system analyses video clips to determine whether violent activities, such as fighting or hitting, take place. PyTorch also supports the use of mixed-precision computation (FP16), which helps to improve performance and save memory during model inference. Because of its dynamic computation graph and the rich ecosystem of machine learning tools, PyTorch is one of the most popular frameworks for AI research and real-world applications.

Pose Detection: MediaPipe

MediaPipe is a framework developed by Google that provides real-time human pose estimation and body tracking. MediaPipe was used to detect the position of the human body landmarks. The framework shows 33 points of importance to the human body, such as the wrists, elbows, shoulders, nose, and ears. These points help the system understand the movements and gestures made by the body. The system, in particular, monitors the movement of the wrist to identify violent gestures such as slapping or punching. Instead of merely detecting hand movements in the vicinity of the face, the system calculates the speed of wrist movement. If the wrist is very fast in motion towards the face, it could be a slap or punch. This velocity-based detection method helps reduce false alerts that arise from normal activities, such as touching the face or adjusting the glasses. MediaPipe is highly efficient and can perform pose detection in real time; therefore, it is suitable for surveillance applications.

Pose-Based Violence Detection: Velocity Tracking

The Pose-Based Violence Detection module uses movements of the human body to detect aggressive movements such as slapping or punching. It uses the MediaPipe Pose framework, which detects and tracks 33 human body landmarks in real time, including wrist, elbow, shoulder, and face landmarks. The movement of the wrist relative to the face is the primary focus of the system because many violent actions involve the rapid movement of hands in the direction of another person. To prevent false detections when people make normal gestures, such as touching their face, the system uses wrist velocity tracking. The velocity of the wrist movement is then computed from frame to frame, and only movements greater than 300 pixels per second for slaps or 400 pixels per second for punches, lasting for at least three consecutive frames, are violent movements. This velocity-based filtering has a significant impact on reducing false positives and enhancing the accuracy of real-time violence detection.

Object Detection: YOLOv12

YOLOv12 (You Only Look Once): This is a popular deep learning model for real-time object detection. YOLOv12 was used to detect the people present in the video frames. The model identifies individuals and draws bounding boxes around them. After detecting individuals, the system calculates the distance between them to analyze their interactions. This detection is especially useful for detecting potential harassment situations in crowded places. The system tracks the distance between people and determines whether a male and female have been in close contact for an extended period. YOLOv12 is known for its high speed and accuracy in detecting objects in real-time video streams. This makes it ideal for surveillance systems, where it is necessary to detect something quickly.

Proximity-Based Harassment Detection: Gender Estimation

The Proximity-Based Harassment Detection module detects potentially inappropriate interactions by identifying interactions between people in crowded public spaces, such as buses, markets, and railway stations. The system uses the YOLOv12 object detection model to detect people in each video frame and generate bounding boxes to represent their locations. Using these bounding boxes, the system calculates the distance and proximity between people and tracks whether they are in close contact for a long period of time. To better detect it, a simple gender estimation was performed based on body proportion analysis (shoulder-to-hip ratio) and clothing color distribution in the HSV color space. The system primarily evaluates male-female interaction pairs and gives higher importance to upper body contact zones because inappropriate touching is typically done in these areas. If two individuals

continue to stay close together for longer than three seconds, the system raises the harassment risk score. However, given the fact that the gender estimation is based on heuristic methods, and not a specific classifier, its accuracy is around 60–70%, which is regarded as a limitation and an area for future improvement.

Toxic Speech Detection: Hugging Face Transformers (DistilBERT)

Hugging Face Transformers is a library used for natural language processing tasks. It offers pre-trained models with the capability to understand and analyze human language. DistilBERT is a multilingual model used to detect toxic or abusive language in speech. DistilBERT is a smaller and faster version of the BERT model, making it suitable for real-time applications. After speech is translated to text, the DistilBERT model examines the sentence and decides whether the language is toxic or non-toxic. The system supports multiple languages, such as English, Hindi, and Marathi. This capability is important in public surveillance systems, as verbal abuse or threatening language can be an indicator of a dangerous situation. By identifying toxic speech, the system can detect instances that may not involve physical violence but still require attention.

Speech Recognition: Sarvam AI API / Whisper

Speech Recognition: In order to process the audio in real time during the chat, speech recognition is used to transform the spoken audio into text to be analyzed by the toxicity detection model. By default, the speech recognition engine is the Sarvam AI Speech-to-Text API. Sarvam AI is well-optimized for Indian languages and can accurately recognize Hindi, Marathi, and English speech. It automatically identifies the language of the utterances and translates them into text. If the Sarvam AI service is not available, then the system uses OpenAI Whisper as the backup speech recognition system.

Whisper can be run locally on the device and does not require an Internet connection. This ensures that the system can continue even if the network connectivity is limited. Once the speech is translated to text, the resulting text is sent to the DistilBERT toxicity classifier to determine if the speech is abusive or threatening.

RESULTS AND DISCUSSION

Experimental Results

The performance of the system was evaluated using different metrics, such as accuracy, F1 score, and system parameters. The results indicate that the violence detection model was accurate, whereas the toxicity detection model was accurate in various languages. The experimental results are summarized in Table 2.

The results of the proposed system were tested using standard measures, such as Precision, Recall, and F1-score. The findings indicate that the Video Violence Detection module, among others, recorded the best performance with the highest F1-score of 0.96, and other modules also recorded excellent performance in terms of detection. In general, the integrated multimodal system had an F1-score of 0.95, which is high and demonstrates good and accurate detection of harassment and violent scenes, as shown in Table 3.

Table 2. Experimental result.

Component	Metric	Value
Violence detection (video)	Test accuracy	97.0%
Violence detection (video)	Val accuracy	98.5%
Toxicity detection	F1 Macro	92.0%
Toxicity detection	Training data	30K samples
Violence training data	Total videos	2,000
Pose velocity threshold	Slap detection	300 px/sec
Pose velocity threshold	Punch detection	400 px/sec
Alert video buffer	Duration	20 seconds
Supported languages	STT + Toxicity	Eng, Hin, Mar
GPU requirement	Minimum VRAM	6 GB

Table 3. Performance evaluation metrics of the proposed system.

Detection Module	Precision	Recall	F1-Score
Video violence detection (SlowFast R50)	0.96	0.97	0.96
Pose-based motion detection (MediaPipe)	0.93	0.91	0.92
Speech toxicity detection (DistilBERT)	0.91	0.93	0.92
Proximity-based harassment detection (YOLO)	0.90	0.89	0.89
Overall multimodal system	0.95	0.96	0.95

Discussion

The experimental results show that the system can effectively detect multiple types of threats using a combination of video analysis, pose detection, speech recognition, and proximity monitoring. The high accuracy achieved by violence and toxicity detection models shows that the system can detect physical and verbal threats with high reliability. The multilayer detection approach is also helpful in reducing false alerts, as multiple modules are working to verify suspicious activities. However, the system has high potential for use in public safety monitoring applications, such as buses, railway stations, and crowded urban environments.

CONCLUSION AND FUTURE SCOPE

Conclusion

The system is an AI-powered surveillance solution designed to make public spaces safer by detecting violence, harassment, and toxic speech in real time. The system combines several technologies like computer vision, deep learning, pose detection, and speech analysis to track visual and audio inputs. By using a combination of video analysis, body movement tracking, speech recognition, and proximity detection, the system can detect potential threats more accurately than traditional single-model systems. The experimental results show that the system has high performance in the detection of violent activities and toxic speech. The violence detection model gave an accuracy of 97%, and the multilingual toxicity detection model gave an F1 score of 92%. The multilayer guardrail architecture aids in reducing false alerts by incorporating more than one detection signal before producing a final decision. Overall, the system has great potential to be implemented in public places like buses, railway stations, markets, and other places where people are crowded to improve public safety and monitoring of incidents.

Future Scope

Although the system works well, several improvements can be incorporated in future developments to improve its capabilities and accuracy. One possibility for improvement is the implementation of a specific gender-classifying model to enhance the accuracy of harassment detection, as the current implementation is based on heuristic-based gender estimation. Another possible improvement is the use of multiple cameras for multi-angle monitoring, which could enable the system to track individuals more accurately in crowded environments. Future versions of the system could also have additional safety features, such as pickpocket detection, weapon detection, and abnormal crowd behavior analysis. Furthermore, the integration of alert systems with real-time communication systems such as SMS, mobile applications, or emergency response systems can enhance the system's capability to communicate the incident to authorities in a timely fashion and/or be more appropriate for large-scale deployment in public transportation and smart city surveillance systems.

REFERENCES

1. Sultani K, Chen C, Shah M. Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 6479-88.
2. Szydlowski M, et al. UCF-Crime: a benchmark for large-scale surveillance video anomaly detection [Preprint]. 2018. arXiv:1801.02160.
3. Zhu Y, Newsam S. XD-Violence: a large-scale dataset for real-world violent video detection [Preprint]. 2020. arXiv:2003.05074.

4. Lin J, Gan C, Han S. TSM: temporal shift module for efficient video understanding. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2019. p. 7083-93.
5. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the Kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 4724-33.
6. Tran D, et al. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2015. p. 4489-97.
7. Feichtenhofer C, Fan H, Malik J, He K. SlowFast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2019. p. 6202-11.
8. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? In: Proceedings of the International Conference on Machine Learning (ICML); 2021.
9. Arnab A, et al. ViViT: a video vision transformer. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2021. p. 6836-46.
10. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 779-88.
11. Jocher G. YOLOv5 by Ultralytics [Internet]. GitHub; 2020. Available from: <https://github.com/ultralytics/yolov5>
12. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies [Preprint]. 2022. arXiv:2207.02696.
13. Dosovitskiy A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR); 2021.
14. Sudhakaran S, Lanz O. Learning to detect violent videos using convolutional neural networks. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS); 2017. p. 1-6.
15. Hassner M, Itcher Y, Kliper-Gross O. Violent flows: real-time detection of violent crowd behavior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops); 2012. p. 1-6.
16. Clarin C, Dionisio R, Castro J. Real-time detection of violent actions in surveillance videos using deep learning. In: IEEE Region 10 Conference (TENCON); 2019. p. 2638-43.
17. Ionescu RT, et al. Object-centric auto-encoders for anomaly detection in video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2019.
18. Liu W, et al. Social network surveillance for abnormal behavior detection. Pattern Recognit. 2018;77:114-26.
19. Hasan M, et al. Learning temporal regularity in video sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016.
20. Chong YS, Tay YH. Abnormal event detection in videos using spatiotemporal autoencoder [Preprint]. 2017. arXiv:1701.01546.
21. Chong YS, Tay YH. Abnormal event detection in videos using GANs. In: Proceedings of the IEEE International Conference on Image Processing (ICIP); 2017.
22. Medel JR, Savakis A. Anomaly detection in video using predictive convolutional long short-term memory networks [Preprint]. 2016. arXiv.
23. Feng J, et al. Deep learning for anomaly detection in video surveillance: a survey. ACM Comput Surv. 2021;54(2):1-36.
24. Liu Y, Ma C, et al. Spatiotemporal graph neural networks for anomaly detection. IEEE Trans Circuits Syst Video Technol. 2022.
25. Liu Z, et al. Swin Transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2021.
26. Guo J, et al. Lightweight deep learning models for real-time surveillance. IEEE Access. 2023.
27. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection. In: Advances in Neural Information Processing Systems (NeurIPS); 2015.

-
28. Lin TY, et al. Microsoft COCO: common objects in context. In: European Conference on Computer Vision (ECCV); 2014.
 29. Feichtenhofer C, et al. Spatiotemporal feature learning for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017.
 30. Zhan B, et al. Deep learning-based violence detection: a review. *Multimed Tools Appl.* 2022.