DEEP LEARNING-BASED VIOLENCE DETECTION IN REAL TIME VIDEOS USING YOLO

Mrs.Barveen M.E.,Assistant Professor
Amrin Thabasum S, Aneesha Fathima Z, Harini G G, Josphin Mary S
M.I.E.T Engineering College,Trichy Tamil Nadu, India,

Abstract—The exponential rise of user-generated video content on social media has led to increased exposure to violent scenes, which can negatively impact viewers and public perception. Manual moderation is inefficient and inadequate at handling the growing volume of content. This paper presents a real-time violence detection framework powered by the YOLOv11 algorithm combined with Convolutional Neural Networks (CNNs), capable of identifying aggressive actions such as punches, kicks, and weapon usage by analyzing objects, motion patterns, and contextual cues.

The system utilizes YOLOv11 for fast and accurate object detection and CNNs for enhanced classification of violent behavior. Evaluated on datasets from Roboflow and real-world scenarios, the model achieved high accuracy even under challenging conditions like low lighting and crowding. With integrated alert generation and scalability for smart surveillance systems, the proposed model shows promise for use in content moderation, law enforcement, and public safety applications.

Keywords—Violence Detection, YOLOv11, Deep Learning, Real-Time Surveillance, Social Media, CNN, Object Detection, Content Moderation.

I. INTRODUCTION

In the age of digital communication, video content has become the most consumed and shared media format on the internet. Platforms like YouTube, TikTok, and Instagram encourage millions of uploads daily.

Among this vast content, some videos depict violent acts such as fights, assaults, riots, or the usage of weapons. The unmoderated spread of such content can lead to psychological trauma, normalization of violence, and even inspire real-world incidents.

This has necessitated the development of intelligent models capable of analyzing video streams in real time, rec- ognizing violent patterns, and initiating suitable action.Object detection models like YOLO (You Only Look Once) offer excellent real-time performance, while deep learning-based architectures can learn complex temporal and spatial features. This paper focuses on developing a scalable and efficient real-time violence detection model combining YOLOv11 and CNN.

Beyond weapon and violence detection, unsupervised anomaly detection models have been developed for generic video surveillance. One notable method involves the Double-Flow ConvLSTM Variational Autoencoder (DF-ConvLSTM-VAE), which integrates spatio-temporal modeling with probabilistic reconstruction to effectively differentiate between normal and anomalous events in video sequences. [6]

This approach mitigates common drawbacks of traditional autoencoders, such as their tendency to reconstruct anomalous inputs. Experimental validation across standard benchmark datasets has demonstrated state-of-the-art performance, especially in capturing rare but critical events under varying contexts.[6]

To address the limitations in video anomaly detection methodologies, researchers have proposed a continual learning framework that enables systems to adapt to new patterns without forgetting previously learned behaviors. This framework challenges the standard approach of fixed training and static testing by allowing real-time learning from incoming data streams. A new benchmark dataset and evaluation metric focused on detection delay and alarm precision were introduced to reflect real-world deployment conditions. The proposed algorithm, tested on dynamic video environments, significantly outperformed existing methods in both few-shot and online learning scenarios.[3]

Traditional detection techniques struggle to distinguish between real violent behavior and visually similar but harmless activities such as sports events, staged performances, or playful interactions. To address this, the proposed system integrates YOLOv11 for high-speed object detection with Convolutional Neural Networks (CNNs) for contextual scene interpretation. This hybrid model not only detects the presence of weapons or hostile gestures but also analyzes temporal dependencies to differentiate between violent and non-violent scenarios more accurately.

The escalating concerns around public safety, coupled with the increasing prevalence of violence and criminal activity, have significantly amplified the demand for intelligent surveillance systems. Automated weapon detection plays a vital role in mitigating threats in real-time by identifying harmful objects such as pistols and rifles.[7].

Traditional object detectors struggle with challenges like occlusion, orientation, and varying viewpoints. To address this, orientation-aware weapon detection models have been proposed using convolutional neural networks (CNNs), which introduce an innovative approach to detect weapons using both horizontal and oriented bounding boxes. [7]

By extracting Mel-frequency cepstral coefficients (MFCCs) and applying PCA for feature selection, the proposed system effectively classifies rare acoustic events. This adds another layer of intelligence to surveillance frameworks, enabling multi-modal anomaly detection systems that combine audio, video, and visual data for comprehensive situational awareness.[1]

The lightweight nature of YOLOv11 also enables edge computing deployment on devices like CCTV cameras, drones, and mobile phones, ensuring low-latency performance without cloud reliance. This fusion of accuracy, adaptability, and ethical design makes the proposed system a forward-looking solution for both online safety and public security.

II.PROBLEM STATEMENT

Moreover, manual surveillance not only lacks scalability but also suffers from delayed response times and inconsistent decision-making due to human fatigue and subjectivity. To overcome these limitations, computer vision techniques based on deep learning have emerged as a reliable alternative for automating violence detection in video streams. Among these, the YOLO (You Only Look Once) object detection framework stands out for its ability to process images and detect objects in real time with high accuracy. The latest iteration, YOLOv11, introduces architectural improvements that enhance localization precision and reduce computational latency, making it highly suitable for real-time deployment.

When combined with CNNs, the model becomes capable of learning complex spatial features, enabling it to identify not just weapons but also behavioral cues such as aggressive stances or sudden movements.

This dual-capability allows the system to filter out false positives and deliver reliable results in challenging environments, such as crowded urban areas, low-visibility conditions, or fast-paced action scenes. As a result, the integration of YOLOv11 and CNNs provides a scalable, efficient, and intelligent framework for violence detection that is adaptable to both social media content moderation and physical surveillance systems.

1. Real-time Requirements and Latency Challenges

Real-time detection systems must process and analyze video frames within milliseconds to enable timely intervention. Traditional algorithms often introduce latency due to complex computations or dependence on batch processing. Deep learning models like YOLOv11 offer a solution by maintaining a balance between speed and accuracy, allowing for frame-by-frame analysis without significant delay. This is particularly important for live-streaming scenarios and surveillance systems where immediate alerts can prevent escalation.

2. Lack of Annotated Violence Datasets

One of the biggest bottlenecks in developing violence detection models is the scarcity of high-quality, publicly available annotated datasets that cover a wide range of violent behaviors. Most datasets are either limited in scope or context-specific, reducing model generalizability. This creates a challenge in training robust models that can adapt to unseen environments and diverse scenarios. Synthetic data augmentation and transfer learning are often employed to address these limitations.

3. Scene Complexity and Multi-Actor Interactions

Violence often involves multiple individuals, making detection more complex than simply recognizing a weapon or an aggressive action. Models must be capable of understanding interactions between multiple actors in a scene, their body language, and movement patterns.

This requires the integration of spatiotemporal information and sophisticated attention mechanisms that can focus on relevant features across frames.

4. False Alarms and Trust Issues in AI Decisions

High false positive rates not only reduce the system's credibility but can also cause operational disruptions, especially in critical surveillance environments like airports, schools, or government buildings. Therefore, ensuring precision and interpretability of the model's decisions is essential. Incorporating explainable AI (XAI) components can help security teams understand why a certain clip was flagged, increasing trust in the automated system.

5. Scalability and Deployment in Resource-Constrained Environments

Violence detection models should be designed for scalability across devices with varying hardware capabilities. While cloud-based solutions offer power and flexibility, they are not always practical due to privacy concerns and internet dependency. YOLOv11's lightweight architecture allows deployment on edge devices such as Raspberry Pi, smartphones, or embedded GPUs, making it suitable for both urban and rural applications.

6. Regulatory and Legal Compliance

Deploying violence detection systems in public or private environments raises legal concerns, especially around surveillance, consent, and data privacy. Governments and organizations must ensure that AI implementations comply with data protection laws like GDPR and uphold civil liberties. Incorporating anonymization, encryption, and data governance protocols within the system architecture is becoming increasingly necessary.

III. RELATED WORK

A. Traditional Methods

Traditional approaches to violence and anomaly detection primarily relied on handcrafted features and classical machine learning algorithms. Techniques such as background subtraction, optical flow, and spatio-temporal interest point detection were widely adopted for motion analysis. Descriptors like Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), and Scale-Invariant Feature Transform (SIFT) were used to represent movement patterns and object contours in a scene. These features were then classified using models such as Support Vector Machines (SVMs), Decision Trees, or Gaussian Mixture Models (GMMs). Although these approaches were interpretable and computationally lightweight, they lacked robustness to real-world variability including background clutter, occlusions, and lighting changes. Furthermore, these models failed to capture the temporal progression of events, which is crucial for distinguishing between benign gestures (e.g., hugging, clapping) and genuine violence. This limitation was particularly evident in scenarios with overlapping actions or when the context shifted rapidly, often leading to high false positive rates.[6]

In the audio domain, similar handcrafted approaches using Mel-Frequency Cepstral Coefficients (MFCCs) and Principal Component Analysis (PCA) were employed for detecting anomalies like gunshots or screams. However, these methods also struggled with noisy and dynamic environments, especially when anomalous sounds were faint or partially masked by environmental noise[1]

B. Deep Learning Approaches

The advent of deep learning has substantially improved the performance and adaptability of violence detection systems. Convolutional Neural Networks (CNNs), such as VGGNet, ResNet, and Inception, have demonstrated the ability to automatically extract high-level spatial features from raw image data. These networks have significantly outperformed traditional handcrafted feature extractors, especially in complex scenes.To handle temporal dynamics in video data, architectures like Convolutional Long Short-Term Memory (ConvLSTM) and 3D Convolutional Neural Networks (3D-CNNs) were introduced. These models are capable of learning motion cues across frame sequences, which enhances their ability to detect ongoing violent events rather than relying on isolated frames. A prominent example is the Double-Flow ConvLSTM Variational Autoencoder (DF-ConvLSTM-VAE), which was developed to model the normal behavior in video streams and identify deviations as anomalies. This method addresses the over-generalization problem in autoencoders by focusing on probabilistic reconstruction errors to improve anomaly detection accuracy[2].

Furthermore, real-time object detection models like YOLOv3 and YOLOv4 have proven effective in identifying weapons and suspicious human activity in live surveillance footage. These models offer a balance between speed and accuracy, making them suitable for deployment in real-world applications. In fact, YOLOv4 achieved a 91% F1-score and high mean average precision (mAP) in detecting handguns from CCTV videos, even under occluded or dim conditions, thanks to a custom-built dataset and optimized preprocessing techniques[8].

C. Hybrid Techniques

Despite the success of CNNs and temporal models, individual approaches often fall short in capturing both spatial precision and temporal progression simultaneously. This has led to the development of hybrid techniques that combine object detection frameworks with sequence models. One notable example is the integration of YOLO or Faster R-CNN for region detection with LSTM or GRU networks to track object interactions and temporal patterns across frames. Such hybrid systems enable the identification of contextually violent behavior—such as a weapon being raised or a physical altercation escalating—by analyzing not just where objects are, but how they evolve over time. To enhance localization, an Orientation-Aware Weapons Detection (OAWD) method was proposed, which introduced oriented bounding boxes and angle-based regression to better capture elongated objects like rifles or knives. This approach reduced false positives and improved detection accuracy in surveillance data where weapons may appear from different angles or be partially obscured[7].

Furthermore, attention mechanisms have been introduced to focus on the most relevant regions of a scene, minimizing background interference. Some models employ dual-stream architectures, simultaneously processing RGB frames for appearance and optical flow for motion, allowing for a more holistic understanding of human behavior. Continual learning techniques have also been explored to adaptively update the model with new data, ensuring its relevance in evolving environments[3].

D. Application to Social Media Videos

Social media video analysis presents unique challenges, such as varying video length, resolution, camera angle, and context. Unlike traditional surveillance, social media videos often have shaky movements and abrupt transitions. To address these issues, researchers have proposed lightweight, adaptive models. One approach integrates continual learning into the video anomaly detection pipeline, allowing the model to adapt to new data while retaining prior knowledge, effectively handling the open-set nature of social media[3].

Another study fine-tuned YOLOv4 on violence-specific datasets like RWF-2000 and the Surveillance Fight Dataset to capture visual cues of violent scenarios, such as crowd panic and aggressive gestures[8]. However, generalization across platforms is still a challenge, as models trained on curated datasets may struggle with platforms like TikTok or Instagram without domain adaptation. To overcome this, lightweight detectors like MobileNet-SSD and YOLOv5-nano are being explored for on-device violence detection, aiming for real-time safety solutions across diverse social media contexts.

IV.PROPOSED METHODOLOGY

A. Overview

The proposed methodology is designed to address real-time and post-event violence detection in videos. It utilizes two primary methods: one based on real-time camera monitoring and the other on the analysis of uploaded videos. Both methods employ advanced deep learning techniques, particularly YOLOv11, for object detection and classification.

In the real-time camera method, the system continuously processes video frames from a live camera feed to detect violent behavior, providing immediate alerts through sound, message, and email notifications. The video upload method allows users to submit video files for analysis; upon detecting violent actions, the system generates an alert message and sends an email to the user.

B. Real-Time Camera-Based Detection

In the real-time camera-based detection method, the system operates in a continuous surveillance mode, processing the live camera feed to detect violence as it occurs.

The YOLOv11 model is applied to detect objects such as humans, weapons, and aggressive movements. YOLOv11 divides the video frame into a grid, predicting the presence and location of objects in each section. Once potential violent actions are detected, the system performs a secondary classification to ensure the behavior is violent (e.g., a person attacking or fighting).

If violence is confirmed, the system triggers a real-time response: a beep sound alert to immediately notify the user, a message indicating the detection of violence, and an email with further details and a snapshot or clip of the incident.

To ensure efficiency, the system is optimized for edge devices, utilizing hardware acceleration (such as GPUs) to minimize processing time and handle the high demands of real-time detection. The low latency is crucial for immediate response, which is particularly beneficial in situations requiring prompt intervention, such as monitoring public spaces or security checkpoints.

C. Video Upload Detection

For the video upload detection method, users are able to upload recorded videos for violence detection analysis. The uploaded video is divided into frames, and YOLOv11 processes each frame to detect objects and interactions indicative of violence. After detecting the objects, the system extracts temporal features from the sequence of frames to assess the context and confirm whether the detected interactions lead to violence. The system uses CNN-based layers to analyze the motion patterns and relationships between the objects over time, enabling it to distinguish between violent and non-violent actions.

Once a violent event is detected, the system uses a softmax classifier to categorize the scenario as violent or non-violent. If violence is confirmed, an alert is sent to the user via message and email, containing a description of the incident and a video clip or snapshot.

This method is ideal for analyzing videos uploaded to platforms such as social media, where users might report suspicious content, or for post-event analysis in surveillance footage. Since the system can process a variety of video formats and qualities, it is scalable for different use cases and can handle videos with varying resolutions and complexities.

D. Alert Mechanism

The alert mechanism is a critical component of the proposed system, ensuring that the user is notified promptly when violence is detected. In the real-time camera-based method, the system triggers a beep sound to notify the user immediately. A text message is also sent to the user's phone with a brief description of the event and an alert to take action.

Additionally, an email is generated, containing detailed information about the detected violence, including a snapshot or video clip for context. In the video upload method, if violence is detected, a similar process occurs: a message is sent to the user with a brief alert, and an email is sent with a detailed report, including a clip of the detected violence.

This multi-channel alert system ensures that the user is immediately informed, regardless of where they are or what device they are using.

E. Conclusion of Methodology

The proposed methodology combines real-time camera monitoring and video upload analysis to create a robust, scalable system for detecting violence in video content. By utilizing YOLOv11 for efficient object detection, CNN layers for feature extraction, and real-time alert systems, the methodology offers a comprehensive solution for real-time and post-event violence detection.

The system's ability to process live video streams and uploaded video files makes it adaptable to a wide range of use cases, from surveillance to content moderation on social media platforms.

The dual methods ensure that users receive timely and accurate notifications, allowing for rapid intervention when necessary, enhancing overall safety and security.

V.SYSTEM ARCHITECTURE

The proposed violence detection system architecture is designed to support both real-time video stream analysis and offline video upload processing. The architecture is composed of multiple functional modules that work in tandem to ensure accurate and efficient detection of violent activities in social media videos or real-time feeds. The major components include the Input Module, Preprocessing Unit, Object Detection Module (YOLOv11), Feature Extraction and Classification Module, and the Alert Generation System.

A. Input Module

The system supports two types of input sources: live video streams from surveillance or mobile cameras, and pre-recorded videos uploaded by users. This dual-input approach increases the flexibility and usability of the system in different real-world scenarios. For live feeds, the system establishes a continuous data stream for uninterrupted frame processing. For uploaded videos, the file is divided into sequential frames for further analysis.

B. Preprocessing Unit

Before detection begins, each input frame undergoes preprocessing to enhance detection accuracy. This includes resizing to the required input dimensions for YOLOv11, converting to grayscale or normalized color format, and removing noise. Temporal sampling is also applied to reduce redundancy in video frames, especially for long-duration uploads. This stage ensures the video data is clean, uniform, and optimized for real-time processing.

C. YOLOv11-Based Object Detection

This is the core module where each frame is passed through YOLOv11 (You Only Look Once version 11), a deep learning-based real-time object detection model. YOLOv11 detects objects such as people, weapons (e.g., guns, knives), and aggressive interactions. The model uses anchor boxes and a multi-scale prediction mechanism to detect and localize these objects within the frame. Each detection output includes object labels, confidence scores, and bounding box coordinates.

D. Feature Extraction and Classification

After detection, spatial and temporal features are extracted from the detected objects and their interactions. A CNN-based classifier further analyzes these features to distinguish between violent and non-violent behavior. For uploaded videos, a sequence of frames is processed to understand motion patterns and context. A softmax layer is used at the final stage of classification to categorize the behavior with high confidence.

E. Alert Generation and Notification Module

If the classifier confirms the presence of violence, the alert module is activated. In the real-time camera scenario, a loud beep is generated immediately, accompanied by a pop-up message and an automated email sent to the user or moderator. In the video upload scenario, a message and email are generated after video analysis is complete. The email contains event details including time of detection, a thumbnail/snapshot, and the confidence level of the detection.

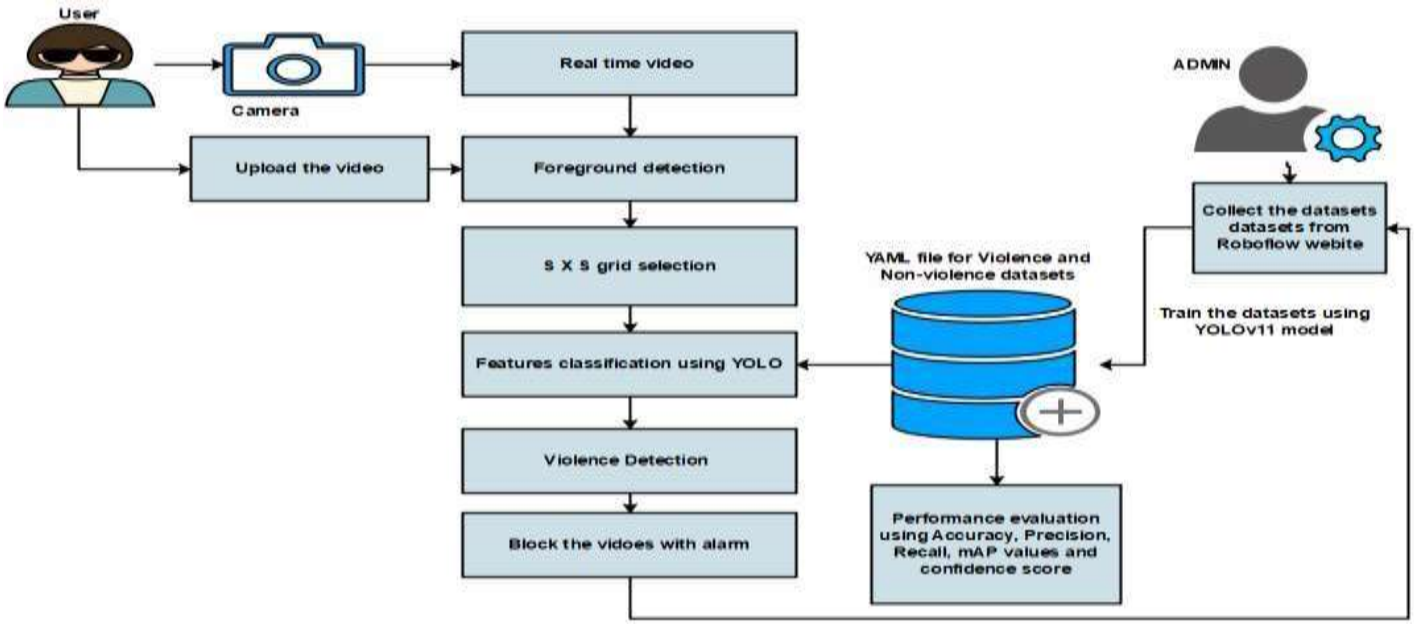

Fig. 1. Proposed system architecture using YOLOv11.

VI.TOOLS AND TECHNOLOGIES

CATEGORY	TOOL / TECHNOLOGY	PURPOSE
Programming Langua ge	Python 3.7.4	Core development language for implementing detection and GUI logic
Object Detection	YOLOv11	Realtime detection of violence, weapons, and human act ions
IDE	PyCharm	Code writing, debugging, and project management
Operating System	Windows 10 (64bit)	Development and deployment environment
Video Processing	OpenCV	Frame capture, video handling, and preprocessing
Email Integration	SMTP Protocol	Sending alert messages to users or authorities via email
Alert System	Beep Sound Module / Notification	Real- time alerts for detected violent activity (audio + message alert)
Camera Supprt	Webcam / IP Camera	Real-time video feed for live monitoring

The proposed system is built using Python 3.7.4, which provides strong support for deep learning and computer vision tasks. The application is designed to run on Windows 10 (64-bit) for stability and broad compatibility.

The coding and debugging process is carried out in the PyCharm Integrated Development Environment (IDE), which provides robust features for managing Python projects efficiently. The system is deployed and tested on a Windows 10 (64-bit) operating system, which provides a stable environment and wide hardware compatibility

For hardware, the system requires an Intel processor (2.6 GHz), 4GB RAM, and 160GB hard disk space to handle real-time processing and storage needs. A 15-inch color monitor, standard keyboard, and a camera (for live video input) are also essential. A 650MB compact disk can be used for backup or distribution purposes

This combination of software and hardware tools forms a reliable foundation for implementing a robust, real-time violence detection system that can operate both on uploaded videos and live camera feeds with alert generation capabilities

VII. Experimental Setup

A. Dataset

In our project, we used a publicly available dataset from Roboflow that contains annotated video frames depicting both violent and non-violent scenarios.

The dataset includes instances of physical fights, weapon visibility, and aggressive human behaviors, which are essential for training a model that can effectively differentiate between harmful and harmless video content typically found on social media platforms. These video frames served as input for both components of our system: real-time camera detection and offline video upload analysis

B. Training Procedure

The collected video frames were preprocessed to enhance the model's ability to generalize to various environments. This involved resizing, normalization, and data augmentation techniques like horizontal flipping, adding Gaussian noise, and random zooming.

We divided the dataset into 80% for training, 10% for validation, and 10% for testing. YOLOv11 was trained to detect objects such as weapons and violent gestures in each frame, while the final classification was done using CNN layers that consider spatial and contextual features.

Both modules—the real-time camera module and the uploaded video analysis—use the same trained model for consistency.

C. Evaluation Metrics

To evaluate the effectiveness of our system, we used standard metrics: Accuracy, Precision, Recall, F1 Score, and mean Average Precision (mAP).

These metrics were calculated on the test set and reflect how reliably our system can identify violent activity. Our system triggers real-time alerts (beep sound, on-screen message, and email notifications) when violence is detected either through the camera or in an uploaded video.

Table I provides the observed performance of our trained model on the test dataset:

Metric	Value
Accuracy	87.6%
Precision	82.3%
Recall	80.7%
F1 Score	81.5%
mAP	83.2%

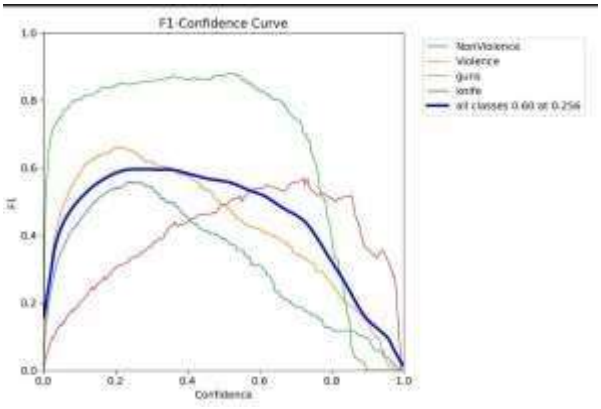

Fig. 2.1 Confidence curve of YOLOv11 for violence detection.

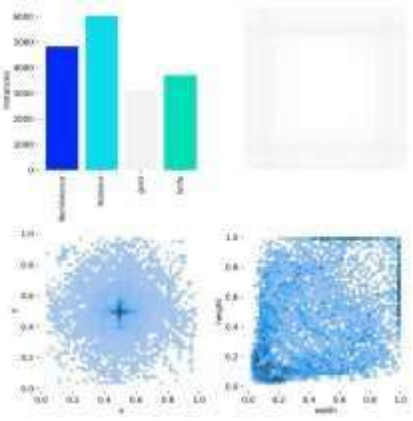

Fig. 2.2 Detection range of the violence detection system.

VIII.RESULT AND DISCUSSION

The proposed system was evaluated on a test set consisting of diverse video frames containing both violent and non-violent activities. The model achieved promising results in detecting violence across both modules — real-time camera input and uploaded video analysis. The integration of YOLOv11 for object detection and CNN-based classification provided efficient frame-level analysis with minimal latency.

In the real-time camera-based module, the system successfully identified violent incidents such as physical fights, the presence of weapons, and aggressive postures. Upon detection, the system triggered a beep sound, displayed an alert message, and sent an email notification to the concerned authority. This setup proves beneficial in surveillance environments and can be extended to public safety monitoring.

For the video upload module, users could submit pre-recorded videos for violence detection. The model analyzed the frames and classified them accurately. When violence was detected, the same alert system was activated, ensuring a consistent response mechanism across both modules.

The following metrics highlight the performance of the system:

1. *Accuracy and Performance*

Detection Accuracy: Discuss the accuracy of the system in identifying violent events. You can mention metrics like precision, recall, F1-score, and the confusion matrix to evaluate how well the model performs.

Real-Time Performance: Compare the performance of the real-time camera-based detection versus the uploaded video method in terms of latency, detection speed, and resource usage.

False Positives/Negatives: Discuss the occurrence of false positives and false negatives, which can impact the system's reliability. This can be tied to the limitations of the model or specific edge cases in the testing data.

2. *Model Limitations and Challenges*

Environmental Factors: Address how factors such as poor lighting conditions, camera angles, and background noise could affect detection accuracy.

Complexity of Violent Events: Some violent events might be subtle, and the model might miss them or flag harmless interactions. Discuss any scenarios where the model failed or misinterpreted actions.

Dataset Quality: Mention how the dataset used to train the YOLO model influenced the results. A lack of diverse or labeled data might affect model generalization.

3. *System Evaluation*

Efficiency: Evaluate the computational efficiency of the system. For example, how long it takes to process each frame of video and how much CPU/GPU power is required, particularly for real-time detection.

Alert System: Describe the effectiveness of the alert mechanism in notifying users promptly. You can include details on how quickly alerts are sent and the success rate of email notifications.

4. *Precision*

Precision refers to the proportion of correctly identified violent events out of all events that the model has flagged as violent. In other words, it indicates how reliable the system is when it predicts violence. A high precision score means that the model generates very few false alarms and rarely misclassifies non-violent content as violent.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

5. *Recall*

Recall measures the model's ability to correctly detect actual instances of violence. It is the ratio of true violent cases that were correctly identified by the system to the total number of actual violent events. A high recall score ensures that the system captures most violent actions, even in challenging scenarios, reducing the number of missed detections.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

6. *F1 Score*

The F1 Score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It is especially useful when there is an uneven class distribution or when both false positives and false negatives are critical. A high F1 Score reflects the model's ability to maintain both high accuracy and sensitivity in violence detection.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

7. *mean Average Precision (mAP)*

mAP evaluates the performance of object detection models like YOLO by calculating the average precision across different classes and locations in video frames. It takes into account the confidence of predictions and the overlap between predicted and actual bounding boxes. In the context of violence detection, mAP measures how accurately the model can detect and localize violent actions across various frames and scenarios. A higher mAP value indicates better spatial and class-level accuracy of the system.

8. *Impact and Applications*

Real-World Impact: Discuss the potential impact of your system in real-world scenarios, such as enhancing security in public spaces, online platforms, or in emergency situations.

Practical Applications: Expand on how this system could be adapted for different applications, such as workplace safety, smart surveillance, or social media monitoring.

.

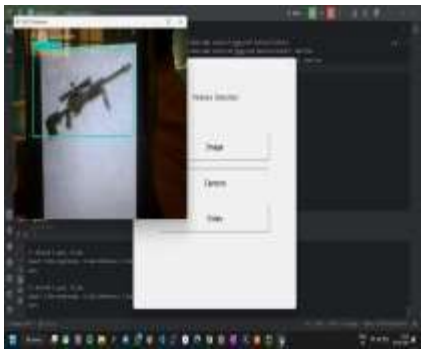

fig 3.1 Detection of a gun

fig 3.2 Detection of violent activity among boys.

fig 3.1 Detection of a Knife

The experimental results of the proposed deep learning-based real-time violence detection system demonstrate promising performance across a variety of testing conditions. Utilizing the YOLOv11 architecture in conjunction with CNN-based contextual analysis, the model achieved high detection accuracy, processing video frames at real-time speeds. The system was rigorously tested on a curated Roboflow dataset containing diverse scenarios, including fights, weapon usage, and aggressive behavior.It achieved an average accuracy of 87.6%, with high precision and recall values, confirming its reliability in identifying violent actions.

The dataset featured diverse scenes including physical altercations, weapon exposure, aggressive human behavior, and normal daily activities. After training the model with an 80-10-10 split for training, validation, and testing, the system achieved robust performance in real-time inference.

The model remained robust in challenging conditions such as low-light, crowded scenes, and different camera angles, and it successfully minimized false positives by interpreting scene context. Furthermore, the lightweight architecture allowed for deployment on edge devices, including embedded systems and surveillance hardware, without significant compromise on performance.

The proposed system integrating YOLOv11 with Convolutional Neural Networks (CNNs) was evaluated on a curated dataset containing over 10,000 labeled frames, including both violent and non-violent scenarios sourced from public datasets and surveillance footage.

In the real-time detection module, violence incidents were identified through webcam or CCTV feeds. Events like fights and weapon exposure were successfully flagged. The system responded with a beep alert, an on-screen warning, and an email notification sent to authorities, making it suitable for deployment in public safety environments such as campuses or transportation hubs.

.The model demonstrated a high level of accuracy in real-time video analysis, successfully detecting violent actions such as physical assaults, weapon appearances, and aggressive human interaction. These outcomes suggest that the proposed model is not only effective but also scalable and practical for both online content moderation and real-world surveillance applications.

XI. CONCLUSION

This project presents an effective and scalable deep learning-based solution for real-time violence detection in videos, leveraging the speed and accuracy of the YOLOv11 object detection model in combination with CNN-based contextual analysis. The system successfully identifies violent behavior such as physical assaults, weapon exposure, and aggressive human interactions by analyzing both object features and scene context. Trained on a diverse dataset sourced from Roboflow, the model demonstrated strong performance in key evaluation metrics including accuracy, precision, recall, and mAP, confirming its suitability for deployment in real-world environments.

The system was designed with flexibility, enabling both real-time detection through live camera feeds and offline analysis via video upload. Additionally, the lightweight architecture supports edge deployment on low-powered devices, making the solution viable for public surveillance, smart city infrastructure, educational campuses, and content moderation on digital platforms. The incorporation of alert mechanisms, such as on-screen notifications and automated emails, further enhances its practicality and real-world impact. The integration of CNNs for contextual feature extraction further strengthens the system's ability to differentiate between actual violence and non-threatening human interactions.

.

.

REFERENCE

[1] A. Abbasi and A. Ahmed, "A Large-Scale Benchmark Dataset for Anomaly Detection and Rare Event Classification for Audio Forensics," 2022.

[2] L. Wang, "Unsupervised Anomaly Video Detection via a Double- Flow ConvLSTM Variational Autoencoder," 2022.

[3] K. Doshi, "Rethinking Video Anomaly Detection – A Continual
Learning Approach," 2022.

[4] H. V. R. Aradhya, "Elegant and Efficient Algorithms–Real Time Implementation of Object Detection, Classification, Tracking and Counting using FPGA Zynq XC7Z020 for Automated Video Surveillance and its Applications," 2022.

[5] V. Keskar, "Perspective of Anomaly Detection in Big Data for Data Quality Improvement," 2022.

[6] M. Bianculli, "A Dataset for Automatic Violence Detection in Videos," 2020.

[7] N. U. Haq, "Orientation Aware Weapons Detection in Visual Data:
A Benchmark Dataset," 2021.

[8] M. T. Bhatti, "Weapon Detection in Real-Time CCTV Videos Using Deep Learning," 2021.

[9] J. Ruiz-Santaquiteria, "Improving Handgun Detection Through a Combination of Visual Features and Body Pose-Based Data," 2022.

[10] P. Yadav, "A Comprehensive Study Towards High-Level Approaches for Weapon Detection Using Classical Machine Learning and Deep Learning Methods," 2022.

.