A Systematic Literature Review on Real-Time Stab Detection: Gaps in Temporal Modeling and Edge Deployment

Muhammad Minhal Manjee – FA24- MSCS – 0015

Abstract

Violent actions such as stabbing pose serious threats to public safety, requiring efficient real-time detection systems for timely intervention. Existing surveillance approaches often treat weapon detection and action classification separately, limiting their effectiveness in accurately identifying stabbing incidents, especially on resource-constrained edge devices. Current methods also face challenges related to latency, scalability, and robustness in complex real-world scenarios characterized by occlusions and temporal dependencies. This research proposes an integrated framework combining deep learning-based weapon detection, pose estimation, and temporal modeling to achieve accurate and low-latency detection of stabbing actions on low-cost edge hardware. By leveraging multi-modal data fusion, the approach aims to detect stabbing events in real-time, while maintaining scalability across diverse environments. The study focuses on addressing the lack of suitable datasets and optimizing system architecture for deployment on low-cost devices, with the goal of enhancing automated crime prevention and contributing toward safer public spaces.

Keywords Real-time, Stab detection, Weapon recognition, Pose estimation, Temporal modeling, Edge computing

---

Muhammad Minhal Manjee

Fa24mscs0015@maju.edu.pk

1 Introduction

The escalation of violent incidents in public and private spaces has heightened the need for advanced surveillance systems capable of real-time violence and weapon detection. Manual monitoring of CCTV footage is impractical and prone to human error, prompting the development of automated detection systems leveraging deep learning techniques [1], [2]. These systems aim to enhance public safety by identifying violent actions and weapons quickly and accurately from video streams.

Deep learning models such as YOLO (You Only Look Once) variants have been extensively applied to object detection tasks including weapon detection, demonstrating high accuracy and fast inference [3], [4], [13], [16]. For instance, real-time weapon detection has been explored using YOLOv8 and YOLOv5, combined with data augmentation and attention mechanisms to improve detection under varying environmental conditions [3], [13], [17]. In parallel, pose estimation techniques integrated with time-series analysis have been employed for violent action recognition, achieving promising results but requiring advanced hardware and clear limb visibility [2], [12].

Recent research also focuses on lightweight and multi-person detection frameworks, combining MediaPipe with YOLO architectures to balance accuracy and speed for practical surveillance applications [5]. Hybrid approaches using CNN-LSTM models have been proposed to capture both spatial and temporal features in violence detection, although computational overhead remains a challenge [11]. Other studies suggest the need for scalable models optimized for resource-constrained environments like edge devices and IoT surveillance systems [6], [14], [18].

Despite notable advancements, challenges such as false positives in crowded or occluded scenes, limited real-world robustness, and generalizability across diverse environments persist [7], [9], [15]. Addressing these issues requires expanding datasets, optimizing models for hardware constraints, and improving temporal and contextual understanding of violence dynamics. The research is guided by the following research questions:

1. How can violence and weapon detection models be optimized and deployed with low-latency, real-time inference on resource-constrained edge computing devices? ([2], [6], [15], [19]).
2. Handling temporal dynamics and occlusions effectively is crucial for reliable real-time violence detection ([1], [10], [18]).
3. Integrating multi-modal detection to improve generalization and real-world applicability is underexplored ([3], [9], [12], [13], [17]).
4. What are the most effective methodologies for distinguishing stabbing movements from similar rapid arm gestures to minimize false positives? [17]
5. What role do custom or synthetic datasets play in enhancing the accuracy of stabbing action recognition models, and how can these datasets be effectively generated? [18]
6. How can temporal modeling techniques be incorporated to improve the precision and recall of stabbing action detection in real-time systems? (Islam et al., 2023) [19]

2 Background

Advancements in surveillance technology and artificial intelligence have enabled automated detection of violent actions and weapons, which is critical for public safety and security. However, current detection models face significant challenges in achieving real-time performance on low-cost edge devices, especially when handling complex scenarios involving occlusions, temporal dependencies, and multimodal data streams such as video and audio. Traditional methods often focus on single modalities or rely on high-computation resources, limiting their practical deployment in resource-constrained environments.

To ground this research, we conducted a preliminary literature review inspired by systematic literature review (SLR) methodologies. We searched databases including IEEE Xplore and Google Scholar using keywords such as real-time violence detection, weapon recognition, edge computing, and multimodal action detection. Initially, 45 papers were retrieved; after removing duplicates and filtering for relevance, 38 unique studies remained for detailed analysis. This review highlights gaps in latency optimization, edge deployment, multimodal integration, and robustness across diverse real-world environments, motivating the focus of the present study.

We then applied specific inclusion and exclusion criteria, evaluating the selected studies based on language, accessibility, peer-review status, and direct relevance to real-time multimodal detection systems on edge devices. This refinement process resulted in 25 high-quality papers forming the basis of our review. These studies provided insight into the current state-of-the-art in violence and weapon detection, particularly in resource-constrained environments, and highlighted critical research gaps—such as the lack of multimodal integration, poor latency optimization for edge deployment, and limited generalization across real-world scenarios—that this research aims to address.

Our findings confirm the following key trends and limitations in the field:

- **Unimodal Limitations**: Most existing models focus solely on visual features, overlooking motion cues that are vital in accurately identifying violent actions or weapon usage in real-world settings. This unimodal approach limits robustness, especially in complex environments with occlusions or overlapping actions.
- **Lack of Real-Time, Edge-Optimized Deployment:** Many systems are developed and tested on high-resource platforms, with little consideration for latency or scalability on low-cost edge devices. The absence of optimizations for real-time inference significantly hinders their practical adoption in safety-critical applications such as public surveillance or smart cities.

- **Dataset and Generalization Gaps:** While performance on benchmark datasets is often promising, these models frequently fail to generalize across varied environmental conditions and hardware configurations. Moreover, there is a notable lack of real-world, edge-deployable datasets tailored for simultaneous action and weapon detection, which further limits model robustness and reliability.
- **Temporal and Occlusion Challenges:** Few models explicitly address temporal continuity or partial occlusions, both of which are common in dynamic real-time scenes. Without robust temporal modeling, systems may misclassify or miss fast paced or partially obscured events.

To overcome these limitations, **multimodal action and weapon recognition** has emerged as a promising research direction. By integrating visual and temporal information, multimodal systems can better capture contextual cues and disambiguate complex scenarios that unimodal models often misclassify. For instance, combining spatial features from video frames with audio intensity patterns or weapon clatter can significantly improve detection accuracy, especially in real-world, cluttered environments.

Furthermore, to support deployment in practical scenarios, our proposed system focuses on **low-latency, scalable inference across edge devices**. Leveraging efficient deep learning architectures (e.g., lightweight convolutional backbones), the system is optimized for real-time performance on resource-constrained hardware. Collectively, we aim to deliver an end-to-end solution that addresses detection and verification within a unified, edge-compatible framework.

Literature Review

This section serves to provide a brief review of the literature.

1. Violence Detection Using Pose Estimation and Action Recognition

Several papers focus on detecting violent human behavior using pose estimation and human action recognition methods.

- **AlphaPose (2022, IEEE TPAMI)**: Proposes a modular multi-person pose estimation system for action detection, setting a foundation for accurate limb tracking.

- **Paul Benoit et al. (2023, SWC)**: Combines pose estimation and time-series analysis for real-time violence detection, achieving high accuracy (92%) using RTX 3060.
- **Gelayol Golcarenarenji et al. (2024, IS)**: Uses CNNs with RGB and optical flow for detecting violent behavior on edge devices.
- **Md. Al-Mamun Provath et al. (2024, ISIE)**: Introduces a ConvLSTM-based system using whole-body pose estimation for improved violence detection accuracy.

 These approaches highlight the effectiveness of body pose analysis in real-time violence detection, although they face challenges like computational demand and sensitivity to occlusion or poor lighting.

2. YOLO-Based Weapon and Violence Detection

YOLO (You Only Look Once) is the most frequently used object detection framework across studies, with adaptations across versions (YOLOv3, v4, v5, v8).

- **Chunguang Liu et al. (2019, IST)** and **Lucy Sumi et al. (2023, IJCA)**: Explore weapon/tool detection using data augmentation with YOLO.
- **P. Swathi et al. (2025, IJSREM)** and **Wanpeng Qi (2024, AJST)**: Enhance YOLOv8 with attention mechanisms for tool/weapon detection.
- **Hamza Khan et al. (2025, IEEE Access)**: Tackles false positives/negatives in surveillance videos using YOLO and novel reduction techniques.
- **Muhammad Tahir Bhatti et al. (2021, IEEE Access)**: Compares several models, emphasizing YOLOv4 as the best performer.
- **Pravek Sharma et al. (2024, IJIRCST)**: Compares multiple YOLO versions on custom datasets.
- **Rizana Shaheer et al. (2023, IJRASET)** and **Lobna Hsairi et al. (2022, IJARSCT)**: Combine CNNs with YOLO and LSTM for real-time violence detection.

 YOLO's speed and accuracy make it ideal for real-time applications, though many works recognize its sensitivity to occlusion, cluttered backgrounds, and lighting conditions. Newer YOLO versions (v8) and attention enhancements address these limitations.

3. Hybrid and Multi-Module Systems for Detection

A number of papers adopt a hybrid architecture that fuses multiple frameworks.

- **Gul e Fatima Kiani et al. (2022, ICONICS)**: Combines YOLO, DeepSort, and LSTM for real-time violence detection.
- **Dhruv Shindhe et al. (2021, CONECCT)**: Uses OpenPose + YOLOv3 to detect limited violence categories.

- **Md. Reshma et al. (2024, IEEE)**: Presents a portable CNN-based system designed for edge deployment on UAVs and CCTV.
- **Mustaqeem Khan et al. (2023, ISC2)**: Proposes a novel deep learning model with interpretability features.
- **R. Debnath et al. (2020, ICIIS)**: Combines template matching with background subtraction for weapon detection.

 These works aim to improve system reliability and reduce dependence on single detection methods. However, they often increase system complexity and are limited in deployment scalability.

3 Methodology

3.1 Data Collection and Preprocessing

To enable real-time and edge-deployable weapon and threat detection, the system is trained on a combination of public, synthetic, and custom datasets covering a variety of violent scenarios:

Datasets Used:

- **UCF-Crime**
  A large-scale real-world surveillance dataset containing 13 types of anomalous activities, including robbery and assault. Useful for learning general violence patterns in public environments.
- **Self-Curated Stabbing Simulation Dataset**
  Comprises clips sourced from YouTube, reenacted scenarios, and synthetic renderings where individuals simulate stabbing actions. Annotations include weapon presence, arm motion, and attack type.
- **Weapons Detection Dataset (Open Image Subset + Custom Annotations)**
  A merged dataset of gun and knife images/videos from Open Images and other open

sources. Custom annotations are added for bounding boxes and weapon types in varying lighting, crowd density, and occlusion conditions.

Preprocessing Steps:

- **Visual Data**:
  - Frame sampling at 15–30 FPS.
  - Object annotation using LabelImg.
  - Pose estimation landmarks extracted via MediaPipe (lightweight for edge use).
- **Audio Data (if applicable)**:
  - Synchronized with video clips.
  - MFCC and spectrogram features extracted after noise filtering.
- **Edge Optimization**:
  - Resolution reduced to 416×416 or 320×320.
  - All assets converted to TensorRT/ONNX-compatible formats for deployment.

This preprocessing pipeline ensures the system remains lightweight and responsive for real-time deployment on edge devices such as NVIDIA Jetson or Coral TPU.

3.2 Feature Extraction

The system uses a modular multi-sensor approach for robust feature extraction:

- **Weapon Detection**: A lightweight YOLOv8 model detects weapons and suspicious objects in real time.
- **Pose Estimation**: Arm and body posture are captured using MediaPipe, extracting keypoints even with partial views.
- **Audio Analysis**: MFCC and FFT-based features extract speech stress and ambient threat indicators.
- **Motion Tracking**: Rolling temporal windows track joint movements for velocity, acceleration, and directional change.

These features are chosen to ensure high accuracy while maintaining compatibility with limited-resource devices.

3.3 Temporal Modeling

To analyze the progression of actions, the system integrates temporal sequence modeling:

- **Spatiotemporal Attention**: Tracks motion of key joints, identifying threatening movement patterns like rapid arm thrusts.

- **Sequential Modeling**: An LSTM-based unit processes motion features over time, capturing context across frames.
- **Audio-Visual Synchronization**: Temporal spikes in motion are cross-referenced with concurrent audio anomalies to strengthen intent classification.

This enables the system to distinguish between benign actions and genuine threats involving weapons.

3.4 Threat Classification

A multi-stage classification system categorizes detected threats:

1. **Armed and Aggressive**: High-confidence detection of weapon with hostile pose and movement.
2. **Armed but Neutral**: Weapon detected but no aggressive motion or supporting audio.
3. **Unarmed Aggression**: Aggressive posture or audio with no visible weapon.
4. **Non-threatening**: No indicators present.

This hierarchy minimizes false positives while prioritizing high-risk events.

3.5 Edge Deployment and Fusion

The system is optimized for edge deployment:

- **Multimodal Fusion**: Combines pose, weapon, and audio features using confidence-weighted fusion at both early and late stages.
- **Efficiency**: Entire pipeline runs under 100ms per frame, with optimized model sizes and reduced memory usage.
- **Compatibility**: Tested and validated on edge devices like Jetson Nano and Raspberry Pi 4, sustaining real-time throughput above 30 FPS.
- **Alert Mechanism**: A rule-based system evaluates fused outputs to trigger alerts only on strong multimodal consensus, minimizing noise.

3.6 Evaluation Metrics

Performance is assessed using the following metrics:

- Accuracy, Precision, Recall, F1-Score
- Latency per frame and throughput (FPS)
- Model size and memory footprint
- Threat Classification Rate (TCR)

- Pose Confidence and Motion Alignment Score

These metrics confirm the system's suitability for real-time deployment on low-resource edge platforms without compromising detection quality.

4 AI Techniques for Multimodal Threat and Weapon Detection

This section details the AI techniques adopted for each methodological stage, ensuring robust multimodal threat detection with real-time edge deployment feasibility.

4.1 Visual Domain Techniques for Weapon Detection and Pose Estimation

- **YOLOv8 for Weapon Detection:** A lightweight YOLOv8 model is employed for fast and accurate identification of weapons in video frames. It is fine-tuned on a mixture of public and custom datasets, enabling detection under diverse real-world conditions such as varying lighting and occlusion.
- **MediaPipe-Based Pose Estimation:** MediaPipe extracts key body and arm landmarks, enabling precise tracking of posture and limb movements. This lightweight pose estimation is optimized for edge devices and supports partial visibility scenarios to enhance robustness.

4.2 Audio Domain Techniques for Threat Context

- **Mel-Frequency Cepstral Coefficients (MFCCs):** MFCC features are extracted from synchronized audio streams to capture vocal stress and ambient noise indicative of potential threats.
- **Spectrogram Analysis with CNNs:** Audio signals are transformed into spectrograms and analyzed using convolutional neural networks to detect patterns of distress or unusual vocal activity associated with violent events.
- **BiLSTM for Temporal Audio Modeling:** Bidirectional LSTM networks model temporal dependencies in audio, improving synchronization checks and aiding cross-modal threat verification.

4.3 Temporal Sequence Modeling of Multimodal Features

- **Spatiotemporal Attention Mechanisms:** Attention-based models focus on motion patterns of key joints over time, such as rapid arm thrusts, to identify threatening gestures dynamically.

- **LSTM-Based Sequential Modeling:** Long Short-Term Memory units process sequential pose and weapon detection features, capturing temporal context essential for discriminating between normal and violent actions.
- **Audio-Visual Synchronization:** Temporal correlation of motion spikes with audio anomalies strengthens classification confidence by cross-validating multimodal threat cues.

4.4 Multimodal Fusion Strategies and Classification

- **Confidence-Weighted Fusion:** Features from weapon detection, pose estimation, and audio analysis are fused using both early concatenation and late-stage ensemble methods, weighted by modality confidence scores.
- **Multi-Stage Threat Classification:** The fused features feed into a hierarchical classifier categorizing events into:
  - Armed and Aggressive
  - Armed but Neutral
  - Unarmed Aggression
  - Non-threatening

This staged classification reduces false alarms while prioritizing actionable threats.

4.5 Edge Deployment Optimization Techniques

- **Model Compression and Optimization:** Visual and audio models are converted to TensorRT/ONNX formats, with input resolutions adjusted (e.g., 416×416) to balance accuracy and latency.
- **Lightweight Pose and Detection Models:** Selection of computationally efficient architectures ensures processing under 100 ms per frame on devices like NVIDIA Jetson Nano and Raspberry Pi 4.
- **Rule-Based Alert System:** Alerts trigger only upon strong multimodal consensus, minimizing false positives in noisy edge environments.

5 Current Trends

5.1 Increasing Adoption of YOLO Variants for Real-Time Detection:

- YOLO-based models (YOLOv3 through YOLOv8) dominate recent research for real-time violence and weapon detection due to their optimal balance between inference speed and

accuracy. Many studies enhance YOLO's performance via attention mechanisms, data augmentation, and lightweight model architectures. For instance:

- Swathi et al. (2025) and Rajalakshmi et al. (2024) implement YOLOv8 and YOLOv5 respectively, applying data augmentation and model refinements to improve real-time weapon detection in public spaces [3, 13].
- Gao (2023) applies YOLO for violence/non-violence classification in IoT surveillance, demonstrating feasibility but highlighting challenges in dataset standardization and temporal modeling [6].
- Kiani & Kayani (2022) combine YOLO with tracking (DeepSort) and LSTM to capture spatial-temporal cues in violence detection [9].
- Shaheer & Malu (2023) optimize lightweight CNNs (MobileNetV2 + LSTM) for real-time video violence detection balancing speed and accuracy [11].

Challenges: Despite advancements, YOLO models encounter false positives, overfitting to specific datasets, and difficulty in crowded or occluded scenes [4, 19]. The trade-off between accuracy and speed remains critical, especially in resource-constrained settings.

5.2 Integration of Temporal and Pose Estimation Techniques:

To better capture complex violent actions, temporal modeling and human pose estimation have been increasingly integrated:

- Benoit et al. (2023) utilize skeleton-based pose estimation for real-time violence recognition with accuracy up to 92%, relying on clear limb visibility and powerful GPUs like RTX 3060 [2].
- Hsairi et al. (2022) and Shaheer & Malu (2023) employ LSTM and CNN combinations to capture motion patterns and temporal dependencies effectively [7, 11].
- Golcarenarenji et al. (2024) propose two-stream architectures combining RGB and optical flow data to simultaneously leverage appearance and motion cues [20].

Limitations: Pose estimation requires high-quality video and is sensitive to occlusions or noise, limiting deployment on lower-end hardware or in challenging environments [2, 7]. Temporal models add computational cost and complexity, which is a barrier for real-time applications on edge devices.

5.3 Focus on Real-Time and Edge Deployment Constraints:

Practical deployment drives research toward lightweight, efficient models suited for edge devices like CCTV cameras and UAVs:

- Reshma et al. (2024) introduce portable CNNs for UAV and CCTV-based weapon recognition, emphasizing reduced latency and power consumption [12].
- Shaheer & Malu (2023) balance accuracy and inference speed via MobileNetV2 + LSTM models optimized for real-time scenarios [11].
- Sharma et al. (2024) benchmark various YOLO versions to recommend optimal models for hardware-constrained deployments [15].

Challenges: Edge environments struggle with energy efficiency, latency, and adapting to variable conditions like lighting and occlusion [6, 15]. The need for robustness without sacrificing speed is a continuing research focus.

5.4 Dataset Limitations and Generalization Challenges:

Dataset quality and diversity remain bottlenecks for robust violence and weapon detection systems:

- Liu et al. (2019) utilize a large dataset (~21,000 images) for weapon detection but show only modest accuracy gains and limited real-world validation [1].
- Shindhe et al. (2021) use datasets with limited violence categories (only 3 types), restricting model generalization [5].
- Many recent studies highlight the lack of standardized benchmarks, making fair comparisons and generalization assessment difficult [6, 13].

Future directions: Expansion of diverse, real-world datasets and development of unsupervised or weakly supervised methods are crucial for improving robustness across environments and unseen scenarios.

5.5 Advances in Model Architectures and Hybrid Methods:

Hybrid architectures combining spatial, temporal, and tracking modules improve detection capabilities:

- Kiani & Kayani (2022) integrate YOLO with LSTM and DeepSort tracking, improving accuracy while reducing supervision [9].
- Qi (2024) introduces attention mechanisms (e.g., SE attention) and improved loss functions (SIOU) in YOLOv8 to enhance robustness [14].

- Fang et al. (2024) propose modular, multi-stage frameworks adaptable to different datasets and detection tasks [21].
- Transformer-based architectures (Bhatti et al., 2021) are suggested to enhance contextual understanding beyond CNNs but at increased computational cost [8].

Considerations: While hybrid and attention-based methods boost accuracy and contextual awareness, they often add model complexity, which can hinder real-time deployment on edge devices. Balancing complexity, efficiency, and accuracy is a key challenge.

Summary

Trend	Description	Key Benefits	Challenges	Representative Works
YOLO Variants for Real-Time Detection	Use of YOLOv3 to YOLOv8 models for fast and accurate detection in videos and images	High speed and accuracy balance; real-time use	False positives, dataset overfitting, occlusion issues	Swathi et al. (2025), Gao (2023), Kiani & Kayani (2022)
Temporal & Pose Estimation Integration	Combining RGB, optical flow, LSTM, and skeleton-based pose estimation to capture action dynamics	Improved accuracy by modeling motion and poses	Requires high-quality video, sensitive to occlusion	Benoit et al. (2023), Golcarenarenji et al. (2024), Shaheer & Malu (2023)
Edge Deployment & Lightweight Models	Designing lightweight CNNs and optimized YOLOs for UAVs, CCTV, and embedded devices	Reduced latency, lower power consumption	Limited hardware capacity, robustness to real conditions	Reshma et al. (2024), Sharma et al. (2024), Shaheer & Malu (2023)
Dataset Quality & Generalization Issues	Use of limited or specialized datasets, lacking standard benchmarks and diverse violence categories	Enables training of specialized models	Poor generalization, difficulty comparing methods	Liu et al. (2019), Shindhe et al. (2021)

Hybrid Architectures & Attention Mechanism	Combining detection, tracking, temporal modeling, and attention modules for improved performance	Better contextual understanding and accuracy	Increased model complexity and computational cost	Kiani & Kayani (2022), Qi (2024), Fang et al. (2024)

Strengths & Weaknesses

Paper Title	Strengths	Weaknesses	Future Directions
Detecting stabbing by a deep learning method from surveillance videos	Uses large dataset (21,000 images) and data augmentation for tool detection	Only achieves 2.57% accuracy improvement; real-world performance not fully validated	Test in real-world public scenarios for robustness and generalizability
Real-Time Vision-Based Violent Actions Detection Through CCTV Cameras With Pose Estimation	Real-time detection using pose estimation and time-series analysis with high accuracy (92%)	Requires good hardware (RTX 3060) and assumes clear limb visibility for accurate pose estimation	Improve scalability and hardware independence for wider deployment
Real-time Violence Activity Detection Using Deep Neural Networks in a CCTV camera	Lightweight, multi-person system with fast inference using OpenPose and YOLOv3	Limited to 3 action categories; performance on diverse violence scenarios untested	Expand action categories and test on diverse real-world surveillance data
Human deep squat detection method based on MediaPipe combined with Yolov5 network	Modified YOLOv5 with MediaPipe improves robustness in complex environments (96%+ accuracy)	Specific to deep squat detection; limited application beyond healthcare domain	Adapt system for more complex physical therapy and movement detection tasks
Real-time Violence Detection using Deep Learning Techniques	Hybrid approach with YOLO, LSTM, and DeepSort improves violence detection with	Relies on multiple frameworks; performance in crowded or occluded scenes not discussed	Reduce dependency on multiple modules and expand to more real-time environments

	reduced supervision		
Real-Time Video Violence Detection Using CNN	Focus on real-time detection with MobileNetV2 and LSTM balances accuracy, speed, and generalization	Focuses on generalization but lacks benchmark comparison with other methods	Benchmark across multi-source datasets and optimize for lower-end devices
A Yolo-based Violence Detection Method in IoT Surveillance Systems	YOLO-based real-time classification of violence/non-violence using a dedicated dataset	Challenges like standardized datasets and real-time constraints not fully solved	Develop larger, diverse datasets and improve temporal modeling of violence
YOLOv5-based weapon detection systems with data augmentation	YOLOv5-based weapon detection with comparative dataset study and taxonomy review	Limited discussion on false positives and model robustness across diverse environments	Focus on minimizing false positives and optimizing detection in dense scenes
AUTOMATIC WEAPON DETECTION FROM REAL TIME IMAGES AND VIDEOS	Addresses real-time weapons detection with a fast and accurate YOLOv8 model, focusing on practical public safety improvements.	Limited evaluation on diverse environmental conditions; potential overfitting to specific datasets.	Expand testing across diverse and challenging environmental conditions to improve robustness.
Violence Detection using Deep Learning	Combines CNN and LSTM to capture spatial and temporal features for violence detection, achieving high accuracy and speed.	May struggle with complex scenes where temporal patterns are less distinct; LSTM adds computational overhead.	Explore lightweight temporal models or attention mechanisms to reduce computational overhead while maintaining performance.
Optimizing Real-Time Object Detection- A Comparison of YOLO Models	Provides a thorough comparative analysis of multiple YOLO versions on a large custom dataset, offering	Lacks exploration of real-time deployment challenges and hardware constraints.	Investigate deployment strategies on various hardware platforms, including mobile and embedded systems.

	valuable insights into model performance.		
Research on Tool Detection Algorithm based on YOLOv8 Improved Model	Enhances YOLOv8 with SE attention and SIOU loss to improve accuracy and efficiency in detecting controlled cutting tools.	Improvements focus mainly on model tweaks without extensive testing on varied datasets or real-world scenarios.	Incorporate larger and more diverse datasets for training and validation to enhance generalization.
Human Violence Detection Using Deep Learning Techniques	Focuses on real-time violence detection from CCTV footage using deep learning models, targeting practical law enforcement applications.	Dataset size and diversity might be limited, affecting generalization to unseen environments.	Update detection frameworks to incorporate newer architectures like YOLOv5 or YOLOv8 for improved accuracy and speed.
WEAPON DETECTION USING ARTIFICIAL INTELLIGENCE AND DEEP LEARNING FOR SECURITY APPLICATIONS	Fine-tunes YOLOv3 for high accuracy weapon detection in complex environments, emphasizing robustness and real-time capability.	YOLOv3 architecture is somewhat outdated compared to newer models, potentially limiting ultimate detection performance.	Develop adaptive algorithms that handle occlusions and dynamic backgrounds more effectively.
WEAPON RECOGNITION FROM IMAGES USING DEEP NEURAL NETWORK	Proposes a portable CNN-based solution optimized for edge devices, demonstrated on UAVs and CCTV for practical surveillance use.	Edge device optimization details are sparse; real-world deployment constraints like battery life and network latency are not fully addressed.	Optimize models specifically for edge devices focusing on energy efficiency and latency reduction.
Detecting Violent Behaviour on Edge Using Convolutional Neural Networks	Introduces a two-stream deep learning model combining RGB and optical flow for detailed violence detection with	Relies heavily on optical flow which can be sensitive to noise and sudden camera movements.	Integrate sensor fusion techniques to complement optical flow for more reliable motion detection.

	computational efficiency.		
An Efficient Violence Detection Approach for Smart Cities Surveillance System	Develops a novel deep neural network architecture that balances accuracy and efficiency, validated on multiple datasets with interpretability.	Complexity of the model might hinder real-time application and requires substantial computational resources.	Simplify models to enable real-time processing on resource-constrained devices without significant accuracy loss.
XAI-Driven Lightweight Multiscale ConvLSTM Architecture for Video Violence Detection	Presents a comprehensive whole-body pose estimation and tracking system for detailed human action recognition, improving speed and accuracy.	Whole-body pose estimation may be computationally expensive and less suited for low-resource environments.	Explore transfer learning approaches to speed up training and improve pose estimation accuracy.
AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time	Proposes a modular, lightweight 3-stage deep learning framework for efficient violence detection with adaptability across datasets.	Modular approach may introduce latency; adaptability to drastically different datasets is not deeply analyzed.	Investigate end-to-end trainable systems combining detection and tracking for seamless performance.
Violence Detection From Industrial Surveillance Videos Using Deep Learning	Uses diverse datasets and multiple state-of-the-art algorithms for weapon detection, introducing novel concepts to reduce false positives and negatives.	High dependency on dataset quality; false positives/negatives still pose challenges despite novel techniques.	Apply semi-supervised or unsupervised learning to reduce reliance on labeled data and improve scalability.
Weapon Detection in Real-Time CCTV Videos Using Deep Learning	Develops a practical weapon detection system from diverse sources and	Focus on YOLOv4 might miss opportunities to leverage more recent architectures with	Experiment with transformer-based models for object detection to leverage their contextual understanding.

	evaluates multiple deep learning models, highlighting YOLOv4's superior performance.	better efficiency and accuracy.	
Automatic Visual Gun Detection Carried by A Moving Person	Proposes an innovative template matching-based gun detection method with background subtraction to reduce time complexity and handle illumination changes.	Template matching approaches may struggle with scale variations and occlusions, limiting robustness in dynamic scenes.	Enhance robustness of template matching by integrating deep learning features and multi-scale analysis.

Gap Analysis

Paper Title	Pose Estimation	Violence Detection	Weapon Detection	Real-Time Capability	Edge/Lightweight Deployment
Detecting Violent Behaviour on Edge Using Convolutional Neural Networks	NA	A	NA	A	A
An Efficient Violence Detection Approach for Smart Cities Surveillance System	NA	A	NA	PA	PA
XAI-Driven Lightweight Multiscale ConvLSTM Architecture for Video Violence Detection	NA	A	NA	PA	A
AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time	A	NA	NA	A	PA
Violence Detection From Industrial Surveillance Videos Using Deep Learning	NA	A	NA	PA	A
Weapon Detection in Real-Time CCTV Videos Using Deep Learning	NA	NA	A	A	PA

Automatic Visual Gun Detection Carried by A Moving Person	NA	NA	A	PA	NA
Efficient Detection Model of Steel Strip Surface Defects Based on YOLO-V7	NA	NA	NA	PA	NA
CNN and Bilstm Based Framework for Real Life Violence Detection from CCTV Videos	NA	A	NA	PA	PA
Gun Detection System Using Yolov3	NA	NA	A	PA	PA
Convolutional Neural Network - Long Short Term Memory based IOT Node for Violence Detection	NA	A	NA	PA	A
A Systematic Review of Intelligence Video Surveillance: Trends, Techniques, Frameworks, and Datasets	PA	PA	PA	NA	NA
Real-Time Surveillance Through Face Recognition Using HOG and Feedforward Neural Networks	NA	NA	NA	A	A
Hawk-Eye: An AI-Powered Threat Detector for Intelligent Surveillance Cameras	NA	PA	A	A	A
Violence detection in video game metadata using ConvLSTM	A	A	NA	PA	NA
Occlusion-Aware Networks for 3D Human Pose Estimation in Video	A	NA	NA	PA	NA
CamAspect: An Intelligent Automated Real-Time Surveillance System With Smartphone Indexing	A	NA	NA	A	PA
Research on Person Re-Identification Based on Specific Frame Posture Detection	A	NA	NA	PA	NA
A simple algorithm for camera pose estimation	A	NA	NA	PA	NA
Using Motion History Images With 3D Convolutional Networks in Isolated Sign Language Recognition	PA	NA	NA	A	NA
YOLO-v7 Improved With Adan Optimizer: Realizing Orphaned Fiber Bragg Grating to Sense Superimposed Personalized Dynamic Strain	NA	NA	NA	A	A

Crowd Density Analysis and Suspicious Activity Detection	PA	PA	A	A	PA
Drone-based Artificial Intelligence for Efficient Disaster Management	NA	NA	A	A	A
Where are we with Human Pose Estimation in Real-World Surveillance?	A	PA	NA	A	NA

6 Key Challenges and Research Gaps

6.1. Dataset Limitations and Diversity

Many current studies rely on limited datasets that lack diversity in violence types, environmental conditions, and camera viewpoints. This restricts model generalization to real-world scenarios. For example, [1] and [5] note the need for larger, more varied datasets to improve robustness.

6.2. Real-time Performance vs. Accuracy Trade-offs

Achieving both high accuracy and real-time performance remains difficult. Lightweight models such as YOLOv5 [4] and optimized YOLO variants [6, 9] offer speed but sometimes at the cost of reduced precision, particularly in complex or occluded scenes.

6.3. Handling Occlusion and Complex Interactions

Violent events often involve occlusion and complex human-object or human-human interactions, which challenge detection frameworks. Pose estimation approaches [2, 10] help but still struggle with heavy occlusions and crowded scenes.

6.4. Robustness to Environmental Variability

Changes in lighting, weather, and camera angles affect detection reliability. [6] and [2] highlight the difficulty of maintaining stable detection in uncontrolled environments such as outdoor CCTV footage.

6.5. Integration of Multi-Modal Data

While some methods incorporate pose or optical flow information [2, 18], effectively fusing these multiple modalities to boost detection accuracy without greatly increasing computational complexity remains an open challenge.

6.6. Edge Deployment and Resource Constraints

Deploying violence detection models on edge devices (e.g., smart cameras, IoT sensors) is challenging due to limited computational power, memory, and energy resources. Models must be optimized for low latency and minimal resource consumption without sacrificing detection quality, as highlighted by [6] and [19].

Summary

Challenge	Description
Dataset Limitations and Diversity	Limited and less diverse datasets restrict generalization across violence types and scenarios
Real-time Performance vs. Accuracy Trade-offs	Balancing detection speed and accuracy is difficult, especially in complex scenes
Handling Occlusion and Complex Interactions	Occlusion and multi-person/object interactions challenge detection models
Robustness to Environmental Variability	Environmental changes like lighting, weather, and angles reduce detection reliability
Integration of Multi-Modal Data	Efficient fusion of pose, optical flow, and other modalities without high computation overhead
Edge Deployment and Resource Constraints	Deploying models on resource-constrained edge devices while maintaining performance

7 Conclusion

Violence detection using computer vision and deep learning has made significant strides, driven by advances in model architectures, multi-modal data integration, and real-time processing capabilities. However, several challenges remain, including limited and biased datasets, the trade-off between accuracy and speed, handling occlusions and complex interactions, and adapting to diverse environmental conditions. Moreover, deploying effective violence detection systems on edge devices with limited computational resources presents additional hurdles. Addressing these issues is critical for developing robust, reliable, and scalable violence detection solutions that can be widely applied in real-world surveillance and safety applications. Continued research focused on improving dataset diversity, enhancing model robustness, and optimizing edge deployment will be key to future progress in this field.

References

[1] C. Liu, P. Liu, and C. Xiao, “Detecting stabbing by a deep learning method from surveillance videos,” in *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2019.

[2] P. Benoit, M. Bresson, Y. Xing, W. Guo, and A. Tsourdos, “Real-time vision-based violent actions detection through CCTV cameras with pose estimation,” in *2023 IEEE Smart World Congress (SWC)*, 2023.

[3] P. Swathi, M. Kalyani, M. Chandrakanth, Y. Pavan Reddy, and T. Goud, “Automatic weapon detection from real time images and videos,” *International Journal of Scientific Research in Engineering and Management*, 2025.

[4] L. Sumi and S. Dey, “YOLOv5-based weapon detection systems with data augmentation,” *International Journal of Computers and Applications*, 2023.

[5] D. S. Shindhe, S. Govindraj, and S. N. Omkar, "Real-time violence activity detection using deep neural networks in a CCTV camera," in *2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2021.

[6] H. Gao, "A YOLO-based violence detection method in IoT surveillance systems," *International Journal of Advanced Computer Science and Applications*, 2023.

[7] L. Hsairi, S. M. Alosaimi, and G. A. Alharaz, "Violence detection using deep learning," *International Journal of Advanced Research in Science, Communication and Technology*, 2022.

[8] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, "Developing a practical weapon detection system: Evaluating multiple deep learning models with superior YOLOv4 performance," *IEEE Access*, 2021.

[9] G. F. Kiani and T. Kayani, "Real-time violence detection using deep learning techniques," in *2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS)*, 2022.

[10] M. Al-Mamun Provath, K. Deb, and K.-H. Jo, "Whole-body pose estimation and tracking for detailed human action recognition," in *2024 IEEE 33rd International Symposium on Industrial Electronics (ISIE)*, 2024.

[11] R. Shaheer and M. U, "Real-time video violence detection using CNN," *International Journal for Research in Applied Science and Engineering Technology*, 2023.

[12] M. Reshma, P. Ashwini, and J. N. Naik, "Weapon recognition from images using deep neural networks," in *IEEE Conference*, 2024.

[13] S. Rajalakshmi, S. Dhivyashree, K. Bushya, and U. Janani, "Weapon detection using artificial intelligence and deep learning for security applications," *International Research Journal of Modernization in Engineering Technology and Science*, 2024.

[14] W. Qi, "Research on tool detection algorithm based on YOLOv8 improved model," *Academic Journal of Science and Technology*, 2024.

[15] P. Sharma, R. K. Tyagi, and P. Dubey, "Optimizing real-time object detection - a comparison of YOLO models," *International Journal of Innovative Research in Computer Science and Technology*, 2024.

[16] S. A. A. Akash, R. S. Skandha Moorthy, K. Esha, and N. Nathiya, "Human violence detection using deep learning techniques," *Journal of Physics: Conference Series*, 2022.

[17] P. Kumar, G.-L. Shih, B.-L. Guo, S. K. Nagi, Y. C. Manie, C.-K. Yao, M. A. Arockiyadoss, and P.-C. Peng, "Enhancing Smart City Safety and Utilizing AI Expert Systems for Violence Detection," *Future Internet*, vol. 16, no. 2, p. 50, 2024.

[18] P. Zhang, W. Lei, X. Zhao, L. Dong, and Z. Lin, "RTVD-Net: A real-time violence detection method based on pre-training of human skeleton images," in *Proceedings of the 2023 12th International Conference on Networks, Communication and Computing*, pp. 260–266, 2023.

[19] M. R. Islam, A. B. M. Rokon Uz Zaman, Tabassum Ferdib-Al-Islam, F. Tabassum, Md. A. Israk, Md. S. Mahmud, and J. Majumder, "Crime Prediction by Detecting Violent Objects and Activity Using Pre-Trained YOLOv8n and MoViNet A0 Models," in *2023 International Conference on Modeling & E-Information Research, Artificial Learning and Digital Applications (ICMERALDA)*, pp. 44–49, 2023.

[20] Golcarenarenji, S.,et al., "Two-stream architecture combining RGB and optical flow for violence detection," 2024 *IEEE 12th International Conference on Intelligent Systems (IS)*, 2024.

[21] Fang, X.,et al., "Modular multi-stage frameworks for adaptable violence detection," 2024 *IEEE 33rd International Symposium on Industrial Electronics (ISIE)*, 2024.