Computer Vision Techniques in Violence Detection: An Overview

P-ISSN: 1680-9300
E-ISSN: 2790-2129
Vol. (26), No. (2)
pp. 143-157

[1]Fatimah A. Jasim, [2] Ielaf O. AbdulMajjed
[1,2] Computer Science Department, University of Mosul, Iraq.

Abstract:

The current research paper discusses the importance of computer vision systems in every facet of life including security, health, commerce, and human-machine interaction. The paper shows that there are numerous computer vision methods of tracking human behavior and detecting violence in video. These techniques are convolutional neural networks (CNNs) to extract spatial features, deep time networks, including RNNs and LSTMs, to analyze the time sequence of events, transformer-based models, including ViViT and TimeSformer, to analyze long sequences of spatial and temporal data, and active optical flow methods to detect sudden changes between images. It is also used in the detection of violence cases, which is based on the posture of people and YOLO or object detection. Moreover, audio analysis and multimodal fusion are used to improve the algorithm of violence detection and supervisory self-learning, advanced spatiotemporal analysis, and matrix analysis are used to present more information about individual behavior. The pros and cons of computer vision-based violence detection algorithms are also compared.

Keywords: Computer Vision, Violence Detection Techniques, CNN, RNN, LSTM, YOLO.

1. Introduction

The adoption of smart monitoring systems in higher education institutions has gained increasing importance as a strategic measure to ensure a safe and stable learning environment. These technological systems facilitate the proactive detection of inappropriate or aggressive behavior, leading to timely corrective action, which significantly reduces incidents and safeguards the institution's security and reputation.

The importance of monitoring human behavior using computer vision as following:

I. Enhancing Security and Preventing Incidents.

"Detecting abnormal human behaviors in surveillance videos is crucial for various domains, including security and public safety" (Wastupranata et al., 2024).

II. Enhancing Human-Machine Interaction.

"Scientists are developing hand gesture recognition systems to improve authentic, efficient, and effortless human–computer interactions without additional gadgets, particularly for the speech-impaired community, which relies on hand gestures as their only mode of communication" (Chang et al., 2023).

Journal of Prospective Researches
Vol. (26), No. (2)
The paper was received on January 17, 2026; accepted on February 11, 2026; and published on June 18, 2026.
Corresponding author's e-mail: fatma.24csp59@student.uomosul.edu.iq

10.61704/pr.567

III. Worker Behavior Analysis.

" Using computer vision for workplace monitoring offers promising benefits, including enhanced workplace security, improved worker safety, assistance in monitoring productivity, and misconduct prevention." (Sebastian et al., 2025).

IV. Uses of Computer Vision in Healthcare and Rehabilitation.

Computer vision can be used to enhance the quality of healthcare delivery and increase the efficiency of the rehabilitation program, as it will be used to monitor patient movements, prevent stumbles or falls, and help to improve patient behavior by revealing and analyzing his or her movements (Gaya-Morey et al., 2024).

V. Monitoring Consumer Behavior During Shopping

Computer vision technologies allow analysis of buyer behavior, allowing those retailers with awareness into foot traffic, result preferences, and customer reaction with promotions, which helps enhance the shopping experience and tailor offers to their interests (Paolanti et al., 2020).

VI. User Behavior Analysis in Digital Applications

In order to improve the user experience by monitoring their interactions with devices and providing intelligent and appropriate responses (Elhoseny et al., 2025).

Computer vision decision is important because it converts the facial expression into precise quantitative data, and the deep model sorts the features according to the quantitative data by eliminating the possibility of using biased or untrustworthy human judgment (Ibrahim & Ramo, 2023).

VII. Supporting Scientific Studies in The Field of Behavior

Computer vision is used to help researchers study human behavior more accurately by analyzing movements and tracking gestures, which helps better understand human behavior (Du & Ikenaga, 2025).

VIII. Enabling Robots to See Their Surroundings and Interact with Humans

"In the present and future, Human-Robot Interaction (HRI) will be a prominent area of robotics. The primary purpose of robots in this domain is to serve as helpful assistants that can effectively aid humans in performing various tasks. For this, robots must be able to interact and communicate well with people." (Abdullah & Mohammed, 2022). This text highlights the importance of computer vision as a key component that enables robots to interact and communicate with humans.

IX. Improve image quality, automate image analysis, and recognize patterns and similarities between them (Raouf & Aldabbagh, 2024).

X. Steganography

Computer vision is important in steganography because it allows images and videos to be processed in a precise manner that allows data to be hidden within them without loss of quality or easy detection (Al-Hayali, 2018).

I. Analyzing Biometric Images

The importance of computer vision in analyzing biometric images (Palm print) to accurately identify a person (Ibrahim & Ramo, 2023).

2. Computer Vision Methods

Methods can usually be classified into several main categories based on how data is processed (video, images, audio) and the type of models used.

2.1 Computer vision-based methods

A. The VGG-19 model is used with Conv2D layers to extract spatial features from videos and images for the purpose of detecting violence using convolutional neural networks (CNNs)(Calderon-Vilca et al., 2025).
B. Convolutional neural networks (CNNs) are combined with long-term and short-term memory networks (LSTMs) to analyze video sequences and detect violence over time (Calderon-Vilca et al., 2025).
C. The ViViT model, a type of transformative video network, is used to analyze videos and understand

changes in images and motion over time (Dilek & Dener, 2025).

2.2 Methods based on facial and body expressions (Human Behavior):

A. Micro-expressions analysis techniques were used to detect signs of anger or potential violence (Yildirim et al., 2023).
B. Pose estimation was used to analyze hand or leg movements to detect fighting or assault (Garcia-Cobo & SanMiguel, 2023).

2.3 Methods based on motion and optical flow (MOF)

A. Optical Flow Analysis: Optical flow was used to detect sudden movement between successive frames in the video (Rendón-Segador et al., 2023).
B. Optical flow was combined with neural networks (CNN + LSTM) to analyze motion and spatial and temporal features in the video (Rendón-Segador et al., 2023).

2.4 Audio-based methods

A. Audio and Emotion Analysis: Neural networks were used to detect emotional sounds such as screaming (Negre et al., 2024).
B. Multimodal Vision and Audio Fusion: Video and audio analysis were combined to improve the accuracy of violence detection (Negre et al., 2024).

2.5 Multimodal Approaches

Combine all types of information, such as images, audio, motion, and facial expressions, to achieve more accurate results in violence detection (Jin et al., 2025). Additionally, A multi-track AI model is used, where each track processes a specific type of data, and then the results are combined to detect violence (Swapnil et al., 2024).

2.6 Conventional forms of Machine Learning

Prior to the development of deep learning, scholars used conventional machine learning algorithms like support vector machines, random forests and decision trees. These schemes involved manual selection of features and pre-analysis then training the model. Conversely, people do not have to perform the manual task of finding features since deep learning models like CNN and LSTM can do this using data (Anwar et al., 2023; Faouzi & Colliot, 2023).

3. Overview of Violence Detection Methods in Computer Vision

Table 1. Summary of Video Violence Detection Techniques, Advantages, and Limitations.

Method Type	Techniques Used	Advantages	Disadvantages	Practical Examples
Convolutional Neural Networks (CNNs)	VGG-19, ResNet, Inception	It proved effective in extracting spatial features (Jain et al., 2023)	It is slow and requires a very large dataset to train the processing (Jain et al., 2023)	Violence in Videos based on CNNs (Jain et al., 2023)
Temporal Deep Learning (RNN, LSTM)	LSTM, GRU	It examines time-based data, enabling it appropriate for tasks of a sequential nature (Kaur & Singh, 2025)	Non-sequential data which is hard to manage, may be slow to process (Kaur & Singh, 2025)	Video violence prediction with LSTM (Kaur & Singh, 2025)
Transformers	ViViT, TimeSformer	Good at processing spatial and time data,	Expensive compute, complicated	Violence detection in videos with ViViT

		appropriate to long videos (Singh et al., 2022)	implementation (Singh et al., 2022)	(Singh et al., 2022)
Motion Analysis (Optical Flow) (MOF)	Farneback, Lucas-Kanade	Identifies frame motion, which is appropriate in sudden motion (Park et al., 2024)	Light sensitive, can be untrue in complicated scenes (Park et al., 2024)	Video violence detection based on Optical Flow (Park et al., 2024)
Pose Estimation	OpenPose, HRNet PoseNet	The system can also be fast and more efficient when using the skeletal points as inferred by PoseNet, which is more useful in a real-time application (Omarov et al., 2022)	The system sometimes gets confused with the identification process of individuals especially on long videos or crowded scenes (Omarov et al., 2022)	Violence detection in videos using Pose Estimation (Garcia-Cobo & SanMiguel, 2023)
YOLO – Real-time Object Detection	YOLOv5, YOLOv7, YOLOv8, YOLOv9	Very fast, detecting people/objects in real-time, can be combined with pose tracking for better detection (Maji et al., 2022)	Less accurate in distinguishing normal vs violent actions without additional techniques, needs extensive training (Salehin et al., 2024)	Detecting fights in schools/universities, crowd monitoring, smart security systems in public places (Senthilkumar et al., 2025)
Audio Analysis	MFCC, Spectrograms	Picks up the sounds of violence, i.e., screaming or loud (Durães et al., 2023)	Can be inaccurate in noisy environments (Durães et al., 2023)	In security surveillance systems and early detection of violent behavior in videos (Durães et al., 2023)
Multimodal Fusion	Combining video, audio, and motion (Ahmad et al., 2025)	Improves detection accuracy, Model attains robust performance on various and challenging violent-scene datasets (Ahmad et al., 2025)	Multi-modal fusion demands a lot of processing power and adds complexity to the model (Ahmad et al., 2025)	Violence detection in videos using Multimodal Approaches (Ahmad et al., 2025)
Self-supervised Learning	Contrastive Learning, Autoencoders (Wu et al., 2024)	"do not require labeled data, making them suitable for unlabeled datasets."(Wu et al.,	These methods perhapes smaller and more accurate compared to supervised approaches and require	Violence detection in videos using Self-supervised Learning (Nardelli & Comminiello, 2024)

		2024)	huge training datasets in order to obtain goodness performance (Wu et al., 2024)	
Spatio-temporal Analysis	GNNs capture interactions and spatio-temporal relationships in videos (Huang & Jiang, 2025)	On multiple video datasets the model achieves high detection (Huang & (Jiang, 2025	High computational cost, complex implementation (Park et al., 2024)	Violence detection using spatiotemporal graphs and 3D-CNN (Park et al., 2024)
Matrix-based Analysis	Grey Level Co-occurrence Matrix (GLCM) (Lohithashva et al., 2020)	Detect patterns in frames, suitable for detecting abnormal activities (Lohithashva et al., 2020)	Sensitive to complicated or dark scenes (Lohithashva et al., 2020)	Violence detection in videos using GLCM (Lohithashva et al., 2020)

3.1 Convolutional Neural Networks

It starts with the intake of a picture contained in a Human Action Recognition (HAR) dataset. The objects in the image are then extracted using a Mask R-CNN model and a segmentation mask is generated. Original image and the segmentation mask are then fed to a separate VGG19 network to obtain visual features. The features obtained are formed into a sequence layer and then forwarded to a fully connected layer (FC layer) to do final classification of the action or behavior in the image (Slade et al., 2022).

3.1.1 The Technique's Advantages

Convolutional neural networks (CNNs) are characterized by their ability to adapt to any new task through the learning process, as the filters in them acquire the ability to automatically extract distinctive and appropriate features from the dataset, making them more effective than traditional methods that rely on manually extracting features (Slade et al., 2022).

3.1.2 The Technique's Drawbacks

Despite advanced CNN models frequently have intricate structures and a lot of parameters, requiring a lot of processing power and memory during training and deployment (Cheng et al., 2024).

3.2 Temporal Deep Learning (RNN and LSTM)

This form of learning is initiated by extracting spatial features of raw data through breaking down images or frame into numbers that capture the most significant features of the image or frame. The analysis is then performed on sequences of events or motions with networks like RNNs and LSTMs. Such networks update information that has been developed over time by recalling what was transmitted in the previous steps.

RNN memory is straightforward that facilitates information transfer through time points, but it can be ineffective in long sequences. The better variant of LSTM is the introduction of gates to assist in deciding what information to control and what to discard and hence it is more effective in acquiring time-dependent long-term relationships in time-dependent data (Anwar et al., 2023).

3.2.1 The Technique's Advantages

- Improved detection: The CNN-RNN models show better results in detecting violence than the traditional models (Hanson et al., 2018).
- Facility to work with Multiple Videos: The combined models are able to deal with different types of

videotapes, including movies, live-action footage, and CCTV footage (Hanson et al., 2018).

3.2.2 The Technique's Drawbacks

- Computational Challenge: The computational resources that are required to combine CNN and RNN models are quite considerable and can limit the ability to process data in real-time (Halder & Chatterjee, 2020).
- Large Data Training Requirement: Models require a lot of data to be meaningfully trained, a fact that can be difficult in certain applications (Halder & Chatterjee, 2020).
- Challenges of interpretation: CNN + RNN models may be difficult to interpret, and it is also difficult to understand how the model decides (Hanson et al., 2018).

3.3 Video Vision Transformer (ViViT)

The ViViT model uses spatiotemporal attention to determine whether a video is violent or not. Spatiotemporal patches (tubes) composed of miniature films are linearly combined to create codes containing spatial and temporal information. These codes are routed through layers of the researcher's encoder to model the dependence of subjective attention on time and space. A classification head is then used to classify the video. In training, data enhancement is used to improve generalization and reduce the weak inductive bias of the transformers on smaller datasets (Singh et al., 2022).

3.3.1 The Technique's Advantages

This technique is distinguished by its ability to detect violent actions, threatening movements, and seizures in video footage. When trained on small datasets, this technique overcomes weak inductive bias by using data amplification techniques. It has demonstrated good results compared to modern methods used on complex standard datasets. As a comprehensive deep learning framework, the detection process is efficient and automated (Singh et al., 2022).

3.3.2 The Technique's Drawbacks

Complex computations and complex implementation (Singh et al., 2022).

3.4 Farneback algorithm Motion analysis technique (optical flow)

The Farneback optical flow algorithm is one of the most important algorithms for estimating motion between two consecutive frames.

- **How does it work?**

  A second-order polygraphed function that approximates the light intensity distribution is used to estimate the local area of each pixel in the two frames. The displacement can be determined by studying how the coefficients of these functions change.

  Applying this process results in a dense optical flow field that is calculated across the entire image, indicating the movement of each point.

  To capture large movements and improve accuracy, the algorithm uses a multi-level approach (image pyramid) (Yakar et al., 2025; Elek et al., 2020; Farnebäck, 2003).

3.4.1 The Technique's Advantages

The Farnback method has been described as being able to allow for a dense visualization of the field flow and improve the visual perception of subtle differences between adjacent pixels (Kim & Gratchev, 2021).

3.4.2 The Technique's Drawbacks

It may be poorly affected by very large displacements or scenes containing homogeneous, texture-free areas (Elek et al., 2020).

3.5 The Technology of Human Posing (HPE)

Deep learning-based HPE methods typically employ convolutional neural networks (CNNs) to obtain features from input images or videos. These networks master elaborate patterns and Depictions that allow the model to identify the positions of human joints correctly. The output is a framework representation of the person, which can be applied for motion analysis, motion gesture detection, and human-computer

interaction applications (Arenas et al., 2025).

3.5.1 The Technique's Advantages

The benefits of the deep learning-based human posture estimation methods are that they are more accurate in the positioning of the joints, less sensitive to the partial obstructions, and can generalize the findings to other postures and bodies. Such techniques prove better than traditional feature-based ones in complicated cases (Arenas et al., 2025).

3.5.2 The Technique's Drawbacks

Although models of human posture estimation using deep learning are effective, they are highly demanding of computing resources, large volumes of labeled data, and might not be capable of operating in extreme impediments or special stances (Arenas et al., 2025).

3.6 YOLO - Object Detection Technology in Real Time

Underlying Feature Extraction.YOLOv8 removes multi-scale picture features successfully based on CSPNet.

- **Neck of Feature Aggregation**
  It helps to detect both the large and small objects as it allows multi-scale detection through a combination of FPN and PAN.
- **Prediction Head**
  It streamlines the training process and makes it more accurate by predicting bounding boxes and class probabilities directly using an anchor-free method (Hussain, 2024).
- **Optimization and Training**
  In order to enhance the robustness and efficiency of training, it utilizes such modules as C2f and EfficientRep, as well as such strategies as CIoU and DFL (Hussain, 2024).

3.6.1 The Technique's Advantages

station for average climate data 1971-2000,

The good thing about the nano model is that it trains quickly and is quick to run when optimized to solve image processing tasks and videos (inference) even when implemented on a hardware platform with no special GPU (Hermens, 2024).

According to the results, YOLOv8 has improved the performance of the previous YOLO versions with an ability to detect objects in low-light conditions and in real-time (Hermens, 2024).

3.6.2 The Technique's Drawbacks

the capability of the detector to extrapolate the same item with invalid backdrops is wanton (Hermens, 2024). Although YOLOv8x has the highest and the highest mean average precision, it was even slower than YOLOv8l (Hermens, 2024). This has proved that even though YOLOv8x has the highest accuracy scores, it requires the use of high-speed object detection, which are not suitable to it because of its complexity (Elhanashi et al., 2024).

3.7 Audio Analysis Technology

A. Phase one: The audio stream is transformed into spectrum representations (MFCC and log-Mel spectrograms) a representation of all the important information in the audio and the features are then accessed (Haque & Afsha, 2022).
B. Phase Two: neural networks analysis of data. The audio features that have been retrieved are checked in this step with deep learning models. One approach utilizes a lightweight Convolutional Neural Network (CNN), whereas another one is applied to a Recurrent Neural Network (RNN). The two algorithms are both trained using real world audio data to identify patterns indicating aggression (Bakhshi et al., 2023). Both algorithms are trained using a dataset of real audio signals to seek suggestive patterns of aggressive behavior (Nardelli & Comminiello, 2024).

3.7.1 The Technique's Advantages

They asserted that their best lightweight approach on the Real-Life Violence Situations dataset had an accuracy of 96.59%. This is an example of how easily lightweight deep

neural networks can detect aggressive behavior in audio samples in the real world (Bakhshi et al., 2023).

3.7.2 The Technique's Drawbacks

Even though the proposed method is practical and lightweight, it may not be efficient in a noisy environment where noises can overlap (Bakhshi et al., 2023).

3.8 Multimodal Fusion Technique

It uses three streams of features, namely: RGB video stream, optical flow stream and audio stream whereby Spatial and Temporal features are extracted in a complementary nature in the form of one another feature using an enhanced attention module. The feature syncretisation takes advantage of the strengths of each individual modality and creates a multimodal feature representation that significantly boosts the accuracy and robustness of anomaly detection (Shin et al., 2024).

3.8.1 The Technique's Advantages

The multimodal fusion concatenation, by integrating the merits of both modalities, generates a complete set of features that highly enhances the strengths and accuracy of anomaly detection of three datasets (Shin et al., 2024).

3.8.2 The Technique's Drawbacks

The model has performed well and conversely, it is said to be complex due to its reliance on the quality of the multi-media data and hence its performance will be compromised in situations where one of the media is noisy or even lost (Shin et al., 2024).

3.9 Self-Supervised Learning Technology

Self-supervised learning has been applied to detect violence in videos using auxiliary tasks, in which human labels are not required. As illustrated in (Yu et al., 2022), contrastive learning helps the model to distinguish similar and dissimilar samples. To determine the latent patterns, autoencoders are used to generate compressed forms of data as shown in (Renugadevi & Kalpana, 2025). Together, these methods can be used to detect violent actions in the video streams that do not have labeled data efficiently.

3.9.1 The Technique's Advantages

The most salient benefit of the SSL is that it minimizes the reliance of labeled information. According to the pre-training and fine-tuning approach, a limited quantity of labelled information can be used to a very high level of performance (Renugadevi & Kalpana, 2025).

3.9.2 The Technique's Drawbacks

In practice, it might not be easy to access much unlabeled data, and, therefore, self-supervised learning is less effective in this case, as explained in (Yang et al., 2025).

3.10 Spatiotemporal Analysis Technology

Advanced spatio-temporal analysis, with the help of spatio-temporal attention processes, integrates both spatial locally based change information, particularly, the location of individuals or objects in a frame, and temporal information, namely how these objects vary with time. This allows the model to identify the salient areas and spots in a video (Varghese et al., 2025). In others Graph neural networks (GNNs) are used to model human spatiotemporal relationships between skeletal points (bones and joints) and individuals and their associations to identify violent behavior in some studies (Mahmoodi & Nezamabadi-pour, 2024).

3.10.1 The Technique's Advantages

Pertinent temporal and spatial analysis at any time and area of interest "The model can be used to realize more effectual analysis of violent behavior through the dynamical analysis of the location, time, and duration to pay attention in a video, using the dynamical prediction networks of three-dimensional attention (Varghese et al., 2025).

3.10.2 The Technique's Drawbacks

Interpretability Problems “These design choices introduce structural biases, add computational complexity, and often limit interpretability." (Alisetti et al., 2025).
The cost of computation is high in graph neural networks.

"The complexity and lack of interpretability of data-driven methods can also impact the accuracy of traffic state prediction." (Guo et al., 2025).

3.11 Matrix-based Analysis method

The technique derives local patterns around each pixel based on the conversion of each frame of the video into grayscale followed by LBP. The frequency of pixel pairs of gray color is then calculated at a specific distance and direction as an output of GLCM. Four texture metrics are extracted out of GLCM, homogeneity, contrast, correlation, energy, After extracting LBP and GLCM features per frame, the features are combined into a single vector and then fed to supervised classifiers e.g. SVM or decision tree to classify the video as a violent or non-violent, In conclusion (post-processing) After extracting LBP and GLCM features in each frame, the results are combined into a single representation and then fed into the supervised classifiers e.g. SVM or decision tree to classify the video as a violent or non-violent (Lohithashva et al., 2020).

3.11.1 The Technique's Advantages

The combination of LBP and GLCM properties yields better detection of violent video events and GLCM properties are computationally cheap and efficient in texture description (Lohithashva et al., 2020).

3.11.2 The Technique's Drawbacks

Acknowledging illumination, intricate background, scale variations and slow-motion in video, they still present the challenge and cannot be resolved by GLCM alone without post-processes or integration of features to represent long-term time variations, as this paper says (Lohithashva et al., 2020).

Table 2. Violence Detection Studies Using Computer Vision Techniques.

#	Authors & Year	Algorithm(s) / Methods Used	Objective of the Work	Dataset Used	Results	Performance Metrics	Contributions	Limitations
1	Beddiar et al., 2020	Survey of HAR methods	Review vision- based HAR approaches	Data acquired in previous researches	Not available	Not available	The paper includes a comparison of human activity recognition techniques and the future issues with it	Obsolete to progress in 2021-2025
2	Ye et al., 2021	CNN + temporal modeling	Identify violent crime scenes on campuses	Campus surveillance datasets	Good domain-based detection	Accuracy, Precision, Recall, F1	Adjusts to special campus conditions	Domain-specific; increscent generalization
3	Zeng et al., 2021	JRL backbone + YOLO	Detection of dangerous objects in real-time with reuse of more features	Public nuisance object subsets; information in paper	mAP improvements over YOLO; faster convergence	mAP, Precision, Recall, FPS	Small- object detection is supported by novel JRL backbone design.	Additional complexity dataset-sensitive
4	Omarov et al., 2022	Systematic review	Summarize violence detection Methods	Literature	Not available	Not available	Comprehensive review	No experiments

5	Singh et al., 2022	Video Vision Transformers (ViViT)	Transformer models that model long-range dependencies can be used to detect violence	UCF-Crime, Hockey Fight, Violent Scenes Dataset	Transformers competitive vs CNNs	Accuracy, F1, mAP, AUC	Adds transformer s in to violence detection.	Heavy compute, Demands extensive datasets for effective training.
6	Omarov et al., 2022	Skeleton + temporal classifier (GCN/ LSTM)	Campus violence detection via skeleton trajectories	Annotations	Robust to clothing/background	Accuracy, F1	Using skeleton trajectories instead of appearance or background areas makes the model more factual and objective and minimizes the effects of external factors	Skeleton extraction fails under occlusion
7	Garcia-Cobo & SanMiguel, 2023	Skeleton-based detection, Change detection	Efficient violence detection	Skeleton-based Datasets Extracted from RWF-2000, Hockey, Movies, and Crowd videos	Reduces false Positives	Accuracy, False alarms	Combines skeletons & change detection	Limited dataset validation
8	Haque et al., 2023	Optical Flow + Mobile Net + Bi- LSTM	Violence detection through lightweight spatio-temporal model	Custom annotated dataset; possibly public ones	Good accuracy vs baselines	Accuracy, Precision, Recall, F1	Combines motion (optical flow) and lightweight solution using (optical flow) with temporal Bi-LSTM	Custom dataset constrained generalization
9	Sahithi et al., 2023	YOLO v 8, OC-SORT	Enhance detection & tracking	Public surveillance datasets	Higher mAP, improved IDF1 Fewer ID switches	mAP, IDF1, IDSW	combining YOLOv8 with tracking algorithm is more successful	Brief experiments only
10	Ullah et al., 2023	Survey	Overview of the vision- based violence detectors	Literature datasets	Not available	Not available	Comprehensive review	No new method

11	Uganya et al., 2023	Tiny-YOLO	Detection of crime scene objects in edge.	5254 images, 8327 images, Internet, GitHu b, IMFD B, Image Net	Acceptable accuracy, high FPS	mAP, FPS	Lightweight, edge- ready	Lower accuracy vs full YOLO
12	Mohammadi & Nazerfard, 2023	Semi-supervised + hard attention	Violence recognition with fewer Labels	RWF, Hockey, Movies	Better localization, data-efficient.	Accuracy, Localization error	Reduces labeling need	Training Complexity
13	Sagunthala et al., 2023 (duplicate/in complete)	Tiny-YOLO variant	Crime scene detection	Similar to #11	Similar results	–	Edge suitability	Duplicate
14	Salehin et al., 2024	YOLO v9-based abnormal detection	Identification of abnormal/violent activity	Public datasets	Improved detection vs YOLOv8	Accura cy, Precision, Recall	Early YOLOv9 evaluation	public release is required in validation
15	Chakraborty et al., 2024	Multi- model ensemble (CNNs, temporal models)	Reduce false positives in violence detection	Likely standard public datasets	A combination of several models enhances the power and stability of the system and the numbers given are detailed in the paper.	Accura cy, Precision, Recall	Demonstrates the advantage of a combination of models using in noisy surveillance videos	Demands more power to compute, and testing is restricted due to it being a conference paper
16	Elmir et al., 2024	Frame Subtraction, YOLOv9	Record only upon activity being detected	Simulated or sample streams	Storage decreased greatly	Storage saved %, Recall, False alarms	Real-life bandwidth/ storage optimization	Risk of missed events

4. limitations

From table 2, scientific, technical and ethical restrictions can be extracted.

4.1 Scientific Limitations

The Small or heterogeneous data sets. Various works used non-specific or small data sets (Beddiar et al., 2020; Ye et al., 2021; Omarov et al., 2022; Sallam et al., 2025).Ad hoc or unpublished data was used in other studies (Haque et al., 2023; Sallam et al., 2025), and it is difficult to extrapolate the results.

4.2 Technical Limitations

Depending on complex or computationally intensive methods, e.g. models which exploit sophisticated architectures such as transformers, or models which are multi-model (Singh et al.,

2022; Chakraborty et al., 2024; Nimma et al., 2025; Jaffrin et al., 2025),will require significant resources for training and operation, limiting their practical application in environments with limited capabilities. Another technical limitation is the impact of data quality, as poor video or image quality significantly degrades performance (Ye et al., 2021; Sahithi et al., 2023; Uganya et al., 2023).

Another Technical limitation include computational and time complications, as integrating multiple models or intensive deep learning increases inference time and resource consumption (Chakraborty et al., 2024; Nimma et al., 2025; Jaffrin et al., 2025).

4.3 Ethical Restrictions

Most studies used surveillance video or personal data (Ye et al., 2021; Sahithi et al., 2023; Jaffrin et al., 2025; Sallam et al., 2025),which may raise legal and ethical concerns about unauthorized surveillance and recording. Therefore, researchers must be made aware of the importance of maintaining privacy and warned of the legal consequences.

5. Conclusion

Overall, this review covered the history of computer vision methods to detect violence, such as the shift to traditional, manually designed feature detection methods to deep learning and hybrid methods. Comparison of CNN, RNN, Transformer, and multimedia integration frameworks indicates that the current systems are getting more and more effective in identifying violent acts in real-time. Nevertheless, existing models continue to have a number of weaknesses related to the asymmetry of the data on which they are trained, their minimal applicability to other settings, and the issue of ethics of surveillance and privacy. Future research would enhance the diversity of datasets, lightweight architecture that can be deployed on peripheral devices, and ethical AI principles to make sure that they are used responsibly. Moreover, it is possible to couple visual data with environmental ones (sound and motion) to strengthen detection models. To sum up, despite the fact that computer vision in violence detection has already achieved a technically high level, the possibility of attaining reliable, unbiased and scalable practical applications is an open question that needs to be addressed multidisciplinary.

References

Abdullah, D. B., & Alnuaimy, M. R. (2022). Real-time face tracking for service-robot. Technium: Romanian Journal of Applied Sciences and Technology, 4(9), 47–52. https://doi.org/10.47577/technium.v4i9.7330

Ahmad, B., Khan, M., & Sajjad, M. (2025). Gated fusion networks for multi-modal violence detection. AI, 6(10), Article 259. https://doi.org/10.3390/ai6100259

Al-Hayali, A. I. B. (2018). Hiding and retrieving encrypted data using LSB in an image based on RBF network. AL-Rafidain Journal of Computer Sciences and Mathematics, 1(1), 199-209 https://stats.uomosul.edu.iq/index.php/csmj/article/view/37053/36843

Anwar, S., Ullah, A., Rocha, Á., & Sousa, M. J. (Eds.). (2023). Proceedings of International Conference on Information Technology and Applications: ICITA 2022 (Lecture Notes in Networks and Systems, Vol. 614). Springer. https://doi.org/10.1007/978-981-19-9331-2

Alisetti, S. V., Kalagi, V., & Krishnagopal, S. (2025). Beyond Attention: Learning Spatio-Temporal Dynamics with Emergent Interpretable Topologies. arXiv preprint arXiv:2506.00770.

Arenas, R., Méndez, R., Pedraza, L., Castelló, E., & Flores, J. (2025). Human pose estimation solutions: a low cost tool for increasing natural interaction in virtual television sets. Universal Access in the Information Society, 24(4), 3257-3270.

Bakhshi, A., García-Gómez, J., Gil-Pita, R., & Chalup, S. (2023). Violence detection in real-life audio signals using lightweight deep neural networks. Procedia Computer Science, 222, 244-251.

Chang, V., Eniola, R. O., Golightly, L., & Xu, Q. A. (2023). An exploration into human–computer interaction: Hand gesture recognition management in a challenging environment. SN Computer Science, 4(5), 441.

Calderon-Vilca, D., Cuadros-Ramos, K., Valcarcel-Ascencios, S., & Aguilar-Alonso, I. (2025). Hybrid CNN Xception and Long Short-Term Memory Model for the Detection of Interpersonal Violence in Videos. International Arab Journal of Information Technology (IAJIT), 22(5).

Du, S., & Ikenaga, T. (2025). Human pose analysis: Deep learning meets human kinematics in video. Springer. https://doi.org/10.1007/978-981-97-9334-1

Dilek, E., & Dener, M. (2025). An overview of transformers for video anomaly detection. Neural Computing and Applications, 1-33. https://link.springer.com/article/10.1007/s00521-025-11218-1

Durães, D., Veloso, B., & Novais, P. (2023). Violence detection in audio: evaluating the effectiveness of deep learning models and data augmentation. DOI: https://doi.org/10.9781/ijimai.2023.08.007

Elhoseny, M., Lydia, E. L., Sree, S. R., Akhmetshin, E., & Shankar, K. (2025). Lightweight Convolutional Neural Network Based Computer Vision Model for Human Behaviour Analysis on Consumer Internet of Things Devices. IEEE Transactions on Consumer Electronics. https://ieeexplore.ieee.org/abstract/document/10980367/

Elhanashi, A., Dini, P., Saponara, S., & Zheng, Q. (2024). TeleStroke: real-time stroke detection with federated learning and YOLOv8 on edge devices. Journal of Real-Time Image Processing, 21(4), 121. https://link.springer.com/article/10.1007/s11554-024-01500-1

Elek, R., Károly, A., Haidegger, T., & Galambos, P. (2020, January). Towards optical flow ego-motion compensation for moving object segmentation. In Proceedings of the International Conference on Robotics, Computer Vision and Intelligent Systems. https://doi.org/10.5220/0010136301140120

Faouzi, J., & Colliot, O. (2023). Classic machine learning methods. Machine learning for brain disorders, 25-75. https://doi.org/10.1007/978-1-0716-3195-9_2

Farnebäck, G. (2003, June). Two-frame motion estimation based on polynomial expansion. In Scandinavian conference on Image analysis (pp. 363-370). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45103-X_50

Gaya-Morey, F. X., Manresa-Yee, C., & Buades-Rubio, J. M. (2024). Deep learning for computer vision-based activity recognition and fall detection of the elderly: a systematic review: GM F. Xavier et al. Applied Intelligence, 54(19), 8982-9007.

Garcia-Cobo, G., & SanMiguel, J. C. (2023). Human skeletons and change detection for efficient violence detection in surveillance videos. Computer Vision and Image Understanding, 233, 103739.

Guo, Q., Tan, Q., Peng, Y., Xiao, L., Liu, M., & Shi, B. (2025). Model-enhanced spatial-temporal attention networks for traffic density prediction. Complex & Intelligent Systems, 11(1), 19.

Huang, H., & Jiang, Q. (2025). IDG-ViolenceNet: A Video Violence Detection Model Integrating Identity-Aware Graphs and 3D-CNN. Sensors, 25(20), 6272.

Hanson, A., Pnvr, K., Krishnagopal, S., & Davis, L. (2018). Bidirectional convolutional lstm for the detection of violence in videos. In Proceedings of the European conference on computer vision (ECCV) workshops.

Halder, R., & Chatterjee, R. (2020). CNN-BiLSTM Model for Violence Detection in Smart Surveillance: R. Halder, R. Chatterjee. SN Computer science, 1(4), 201.

Hussain, M. (2024). Yolov1 to v8: Unveiling each variant–a comprehensive review of yolo. IEEE access, 12, 42816-42833.

Hermens, F. (2024). Automatic object detection for behavioural research using YOLOv8. Behavior research methods, 56(7), 7307-7330.

Haque, M. A. H. M. U. D. U. L. (2022). Detection and classification of sensitive audio-visual content for automated film censorship and rating.

Ibrahim, R. T., & Ramo, F. M. (2023). Hybrid intelligent technique with deep learning to classify personality traits. International Journal of Computing and Digital Systems, 13(1), 231-244.

Jin, W., Zhu, L., & Sun, J. (2025). Aligning First, Then Fusing: A novel weakly supervised multimodal violence detection method. Knowledge-Based Systems, 322, 113709.

Jain, D. M., Bora, M. S., Chandnani, S., Grover, S., & Sadwal, S. (2023). Comparison of VGG-16, VGG-19, and ResNet-101 CNN models for the purpose of suspicious activity detection. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 121-130.

Kaur, G., & Singh, S. (2025). An ensemble based approach for violence detection in videos using deep transfer learning. Multimedia Tools and Applications, 84(12), 11001-11025.

Dong-Hyun, K., & Gratchev, I. (2021). Application of optical flow technique and photogrammetry for rockfall dynamics: A case study on a field test. Remote Sensing, 13(20), 4124.

Lohithashva, B. H., Aradhya, V. N., & Guru, D. S. (2020). Violent Video Event Detection Based on Integrated LBP and GLCM Texture Features. Revue d'Intelligence Artificielle, 34(2). https://doi.org/10.18280/ria.340208

Mahmoodi, J., & Nezamabadi-pour, H. (2024). A spatio-temporal model for violence detection based on spatial and temporal attention modules and 2D CNNs. Pattern Analysis and Applications, 27(2), 46.

Maji, D., Nagori, S., Mathew, M., & Poddar, D. (2022). Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2637-2646). https://doi.org/10.48550/arXiv.2204.06806

Negre, P., Alonso, R. S., González-Briones, A., Prieto, J., & Rodríguez-González, S. (2024). Literature review of deep-learning-based detection of violence in video. Sensors, 24(12), 4016.

Nardelli, P., & Comminiello, D. (2024). Josenet: A joint stream embedding network for violence detection in surveillance videos. arXiv preprint arXiv:2405.02961. https://doi.org/10.48550/arXiv.2405.02961

Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., & Khassanova, M. (2022). A skeleton-based approach for campus violence detection. Computers, Materials & Continua, 72(1).

Paolanti, M., Pietrini, R., Mancini, A., Frontoni, E., & Zingaretti, P. (2020). Deep understanding of shopper behaviours and interactions using RGB-D vision. Machine Vision and Applications, 31(7), 66.

Park, J. H., Mahmoud, M., & Kang, H. S. (2024). Conv3D-based video violence detection network using optical flow and RGB data. Sensors, 24(2), 317. https://doi.org/10.3390/s24020317

Raouf, N. N., & Aldabbagh, M. T. (2024). Developing a Third-Party API to Enhance Image Documents at the University of Mosul Data Center. International Journal of Computing, 23(4), 692-701.

Rendón-Segador, F. J., Álvarez-García, J. A., Salazar-González, J. L., & Tommasi, T. (2023). Crimenet: Neural structured learning using vision transformer for violence detection. Neural networks, 161, 318-329.

Renugadevi, P., & Kalpana, S. (2025). A self-supervised lightweight CNN-LSTM network for violence detection in low-labelled surveillance video. International Journal of Scientific Research and Engineering Development, 8(4), 914–918. https://doi.org/10.5281/zenodo.16411318

Sebastian, R. A., Ehinger, K., & Miller, T. (2025). Do we need watchful eyes on our workers? Ethics of using computer vision for workplace surveillance. AI and Ethics, 5(4), 3557-3577.

Swapnil, A. L., Peris, M. D., Nihal, I. H., Khan, R., & Matin, M. A. (2024). Multimodal Deep Learning for Violence Detection: VGGish and MobileViT Integration with Knowledge Distillation on Jetson Nano. IEEE Open Journal of the Communications Society, 6, 2907-2925.

Singh, S., Dewangan, S., Krishna, G. S., Tyagi, V., Reddy, S., & Medi, P. R. (2022). Video vision transformers for violence detection. arXiv preprint arXiv:2209.03561. https://doi.org/10.48550/arXiv.2209.03561

Salehin, S., Rahman, S., Nur, M., Asif, A., Bin Harun, M., & Uddin, J. I. A. (2024). A Deep Learning Model for YOLOv9-based Human Abnormal Activity Detection: Violence and Non-Violence Classification. Iranian Journal of Electrical & Electronic Engineering, 20(4).

Senthilkumar, S., Agarwal, G., Shirish, A., & Kolte, S. (2025, February). Real Time Violence Detection System Using YOLOv7 and Deep Learning Techniques. In 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) (pp. 1447-1454). IEEE. https://doi.org/10.1109/IDCIOT64235.2025.10914712

Slade, S., Zhang, L., Yu, Y., & Lim, C. P. (2022). An evolving ensemble model of multi-stream convolutional neural networks for human action recognition in still images. Neural computing and applications, 34(11), 9205-9231.

Shin, J., Miah, A. S. M., Kaneko, Y., Hassan, N., Lee, H. S., & Jang, S. W. (2024). Multimodal attention-enhanced feature fusion-based weakly supervised anomaly violence detection. IEEE Open Journal of the Computer Society, 6, 129-140.

Varghese, E. B., Elzein, A., Yang, Y., & Qaraqe, M. (2025). A temporal–spatial deep learning framework leveraging dynamic 3D attention maps for violence detection. Neural Computing and Applications, 37(32), 26689-26709.

Wastupranata, L. M., Kong, S. G., & Wang, L. (2024). Deep Learning for Abnormal Human Behavior Detection in Surveillance Videos—A Survey. Electronics (2079-9292), 13(13).

Wu, P., Pan, C., Yan, Y., Pang, G., Yan, Q., Wang, P., & Zhang, Y. (2026). Deep learning for video anomaly detection: A review. IEEE Transactions on Neural Networks and Learning Systems.

Yu, J., Liu, J., Cheng, Y., Feng, R., & Zhang, Y. (2022, October). Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. In Proceedings of the 30th ACM international conference on multimedia (pp. 6278-6287). https://doi.org/10.1145/3503161.3547868

Yildirim, S., Chimeumanu, M. S., & Rana, Z. A. (2023). The influence of micro-expressions on deception detection. Multimedia Tools and Applications, 82(19), 29115-29133. https://link.springer.com/article/10.1007/s11042-023-14551-6

Yang, Z., Du, H., Niyato, D., Wang, X., Zhou, Y., Feng, L., ... & Qiu, X. (2025). Revolutionizing wireless networks with self-supervised learning: A pathway to intelligent communications. IEEE Wireless Communications.

Yakar, İ., Kuçak, R. A., Bilgi, S., Ferhanoglu, O., & Akinci, T. C. (2025). A Hybrid Deep Learning and Optical Flow Framework for Monocular Capsule Endoscopy Localization. Electronics, 14(18), 3722.

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 57(4), 99.