ISSN: 2581-3404 (Online) **IF: 5.86 (SJIF)**
IJIRTM, Volume-10, Special Issue-2, April-2026

Evolution of Computer Vision: From Traditional Techniques to Sequential Deep Learning

Mayank Paul[1], Dr. Preeti Malhotra[2]
[1] Department of Computer Science & Engineering (AI & DS), Panipat Institute of Engineering and Technology, Panipat, Haryana, India
[2] Associate Professor, Department of Computer Science & Engineering (AI&DS), Panipat Institute of Engineering and Technology, Panipat, Haryana, India
paulmayank87@gmail.com[1] , drpreeti.cse-ai-ds@piet.co.in[2]

Abstract. *Most existing CCTV setups now days are fundamentally reactive, implying that they only record crimes for later review instead of inhibiting them as they happen. Dependence on human operators on regular monitoring is also impractical. Because rapid onset of human fatigue leads to missed threats. To overcome this, our paper examines the incline towards proactive security, notably through a hybrid framework we call 'Falcon AI defence system'. Contrary to standard systems that only detect objects. Our proposed approach employs two distinct deep learning pipelines parallelly. First is an Identity Module and second is a Behavior Module. While the first module manages facial recognition, the second utilizes MediaPipe for skeletal landmark extraction paired with a Long Short-Term Memory (LSTM) network. By assessing 30-frame sequences in a second, the system enables reliably classify complex actions like fighting or falling, etc. Our findings indicate that merging identity tracking with real-time behavioral alerts provides a cost-effective, scalable defense for smart city infrastructures, effectively turning passive cameras into intelligent security tools.*

Keywords: *Smart-Surveillance, Action Recognition, Proactive Surveillance Systems, Deep-Learning, LSTM, MediaPipe.*

Introduction

The primary drawback of current CCTV surveillance is reactive nature. The systems record crimes as they happen and respond after that. But the intelligence system is lacking in sending the alert to authorities in real-time. In contrast, modern Deep Learning and Computer Vision techniques have introduced Object Detection and Face Recognition [1,11]. They work separately. There is a critical need for an automated system that not only identifies individuals but also recognizes their temporal behavior to prevent violent incidents in public and sensitive spaces.

Initial approaches to video surveillance used traditional computer vision and machine learning techniques. A primary task is to detect the moving object and handle it through background subtraction. Systems would build a mathematical model of a static background and flag any pixels that deviated from this model as moving objects. Temporal differencing [3] simply subtracted one frame from the previous one to find changing pixels. The researchers also applied statistical methods [3] and Principal Component Analysis, to separate moving objects from static frame. To identifying specific object like faces, algorithms such as Haar Cascade Classifiers [1] were heavily used.

In early techniques strictly face detection limitations. Temporal differencing [3] completely fails when an object moves very slowly or the movement stops. When also the Background light is sudden changes and changes in background clutter, such as object changes its position. Mainly the methods relying on static features. Such as Haar cascades [1]. They are highly sensitive to camera angles and fail in low light. When faces are partially covered by accessories and other items they cannot detect. More important is traditional machine learning classifiers only analyzed individual, disconnected frames. They could detect a person who was in the frame. But they lack the structural capacity to compute what they person was doing over a span of time.

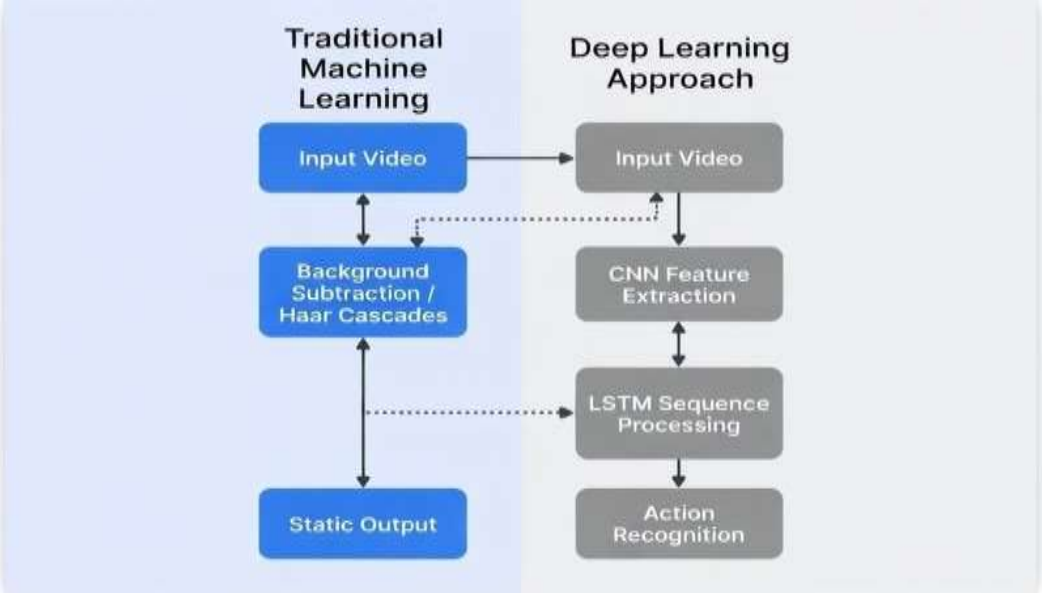

Fig. 1: Comparison of Traditional ML vs. Deep Learning Workflows

2. Current State-of-the-Art

The introduction of deep learning completely changed the video analysis. First, Convolutional Neural Networks (CNNs) [6] like the YOLO (You Only Look Once) [6] family drastically improved static object detection. Architectures like YOLOv4 [5] and its lightweight variant YOLOv4-tiny allowing us to work on low-powered devices. And detect objects in real-time with high accuracy. While CNNs extracting separate features from a single image. Detecting the violence in the frame.

To solve this problem, modern systems combine CNNs with sequential networks. One leading approach uses a CNN, like ResNet50 [6], to extract separate features from a batch of continues frames. These features are then feed directly into a Long Short-Term Memory (LSTM) network [6]. The LSTM tracks the pixels and how the image features change over time. Also capture single frame. That allows our system to recognize violent actions with high accuracy.

Another method which we are considering depends on pose estimation [7] to understand continuous motion in the frame. Instead of analyzing the raw pixels. These algorithms map the human skeleton by locating 33 joints [7] in every single frame. The system converts these joint coordinates into angles and

then after collects them over a sliding window of time. This is collecting 30 frames in single second. A neural network [7] then processes these continuous sequences of 33 joint angles to classify the human behavior in the frame (walking, standing, etc.). This method is highly effective because it ignores the whole background clutter and only focuses on human movement. To prevent the system from rapidly switching between classifications and rolling prediction averages are used to smooth out the final output.

3. Literature Review

The switch from traditional computer vision to deep learning has fundamentally automated the surveillance system. In this research our primarily focused on pixel-level analysis for motion detection in the frame. Demonstrated that Haar Cascade Classifiers [1] integrated with OpenCV could effectively perform real-time face recognition on low-power devices like the Raspberry Pi. Though the operational range was limited to 1.2 meters [1]. To enhance the speed of Tank and Thakore we use Haar Wavelet Decomposition [2] with Mixture of Gaussian (MoG) models [2]. That helps us to achieve a threefold increase in moving object detection speed compared to standard MoG models [2].

This is improvements in early methods were highly sensitive to environmental factors. Such as lighting and shadows in frames. This led to the adoption of YOLO architecture is one of the best solutions. Now Highlighted YOLOv4 [5] as a breakthrough in object detection. Balancing high precision with real-time performance. For lack of resource environments, Jiang et al. developed YOLOv4-tiny [4]. Which uses the ResBlock-D modules [4] to achieve a throughput of approx. 294 FPS [4] on normal hardware, although with a slight reduction in the mean Average Precision (mAP).

The most significant advancement in the threat detection system has been the integration of time-based analysis. Patel proposed a CNN-LSTM hybrid model which uses ResNet50 for separate feature extraction and LSTM for sequential processing. To reach a validation accuracy of 92.3% [6] on violent behavior datasets as shown in the paper. Similarly, Kwan-Loo et al. explored Pose Estimation [7] by tracking joint angles up to 10 to 30 frames, achieving over 98% accuracy on the Kranok-NV database [7]. Beyond the YOLOv4 architecture, recent evaluations suggest that further optimizations in YOLOv8 [8] have improved edge intelligence for public safety. In studies on low-power hardware [8] emphasize that choosing the right architecture is critical for maintaining real-time monitoring efficiency.

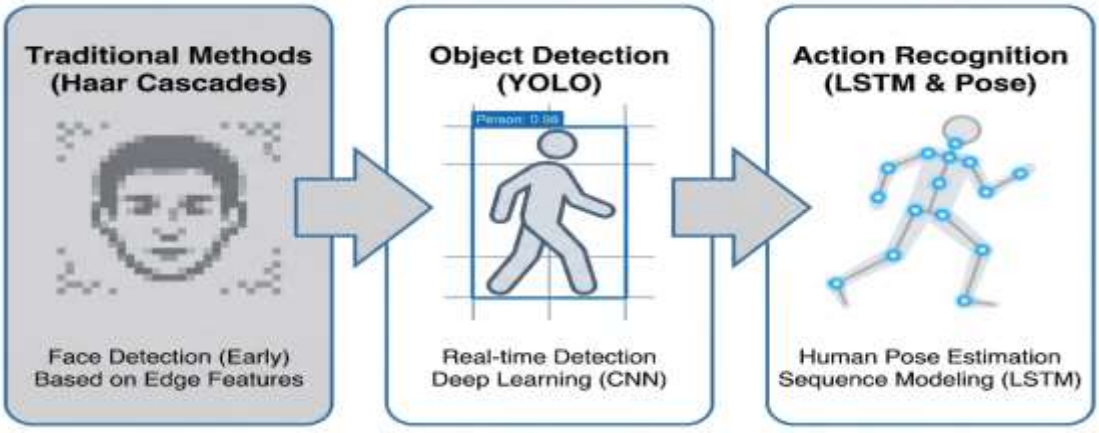

Fig. 2: Technology Shift from face to action recognition

4. Comparative Analysis:

Our analysis on the existing literature refers to a specific between computational efficiency and classification of the pixels.

- Detection vs. Interpretation: Traditional methods like background subtraction and Haar Cascades [1] are computationally inexpensive. But there is lack the structural capacity to surveillance continuous the human behavior. The system is not that capable. They are primarily dealing with object detection rather than the action classification
- Pixel-Based CNN vs. Pose Estimation: The CNN and LSTM [6] approach are solid but the struggles in crowded frame like public places. Where the background movement disrupts pixel analysis in the frame. However, Pose Estimation [7] is confined to human skeletal landmarks with effectively ignoring background noise. While pixel-based sequences suggest a solid foundation for behavior analysis. Comparative studies show that combining skeletal tracking with stacked LSTM networks [9] provides the best classification of human actions. However, pose-based [7] systems do false alarms caused by occlusions. Also objects like umbrellas and other items that mimic violent postures
- Spatial vs. Temporal [3]: While YOLOv4 [5] provides exceptional object detection at 65 FPS on high-end GPU system. It cannot distinguish between normal movements and violent acts without time-based context. Sequential models like LSTM [6] solve this by processing motion across using single frames.

This table below summarizes the key papers, technologies, results, and drawbacks in this analysis

Table 1: Comparative Analysis:

Paper Title	Core Technology Used	Accuracy / Results	Limitations (Drawbacks)
Real time face recognition of video Surveillance system using haar cascade classifier [1]	Haar Cascade Classifier, Raspberry Pi, OpenCV.	Successfully recognizes human faces at 0.4 to 1.2 meters. Accuracy ranges from 38% to 56%.	Fails in low light or if facial angle is +>40◦.
A Fast Moving Object Detection Technique In Video Surveillance System [2]	Mixture of Gaussian (MoG) models, Haar wavelet decomposition.	Detects moving objects 3 times faster than using MoG alone. Works in complex backgrounds.	Cast shadows are falsely detected as moving objects. Cannot classify objects.
A Survey on Moving Object Detection and Tracking in Video Surveillance System. [3]	Review of Background Subtraction, Temporal Differencing, and Statistical Methods. +1	Compares existing techniques for separating foreground objects from backgrounds.	Temporal differencing fails if objects move too slowly or stop moving.
Real-time object detection method for embedded devices. [4]	Improved YOLOv4-tiny with ResBlock-D modules and attention mechanisms.	Achieved 38.0% mAP. Processes 294 FPS on a PC and improved speed on Raspberry Pi.	Sacrifices overall accuracy compared to full-sized models for high speed.

YOLOV4: A BREAKTHROUGH IN REAL-TIME OBJECT DETECTION [6]	YOLOv4 architecture (CSPDarkNet53, PANet, and Spatial Pyramid Pooling).	Achieved 43.5% AP at 65 FPS on high-end GPU. 10% accuracy increase over YOLOv3.	Strict trade-off between speed and accuracy; larger resolutions slow down processing.
Real-Time Violence Detection Using CNN-LSTM [7]	ResNet50 (CNN) combined with LSTM, analyzing pixel differences between frames.	Reached 92.3% overall accuracy. High accuracy on movie clips and hockey fights.	Struggles with large crowd videos where people are just onlookers.
Detection of Violent Behaviour Using Neural Networks and Pose Estimation [8]	YOLOv3 (detection), IoU (tracking), and Open Pose (joint angles).	Achieved over 98% accuracy on "Kranok-NV" database by analyzing body poses.	False alarms on umbrellas; fails if pedestrians physically overlap.
Deep Learning Strategies for Human Behavior Analysis [8]	YOLO-based spatial extraction and behavioral sequence logic.	High robustness in identifying suspicious activities in public spaces.	Computational cost increases with multiple behavioral patterns.
A Comparative Study of CNN-LSTM and Skeletal Tracking [9]	Comparison between CNN-LSTM and Pose-based landmark extraction.	Skeletal tracking offers >95% precision in variable lighting.	Struggles with heavy occlusions where joints are hidden.
Optimization of YOLO Architectures for Low-power Hardware [10]	Structural pruning of YOLOv8 and implementation on Edge AI modules.	Maintained high mAP while reducing memory footprint by 25%.	Slight accuracy drop compared to full-size models on high-end GPUs.

5. Conclusion

The automating surveillance system has successfully shifted from static image detecting to complex behavioral analysis. Deep learning models that process sequential data over time, like LSTMs and pose-tracking networks. Now, if we are defining the industry standard for threat detection in any location or place. Moving forward to this technology is rapidly adapting this tech efficiently on the low-power devices. This helps us save our funds. Also, now a day the systems will merge visual tracking with audio analysis with the help of tools like Mel Spectrograms. This is intended to be developed with even more robust detection pipelines. Priority-based scheduling algorithms are also being occupied the computational power in a selectively to camera feeds. That demonstrates high statistical probability, when any thread is sensed by surveillance system.

References:

[1] A. H. Ahmad, S. Saon, A. K. Mahamad, C. Darujati, S. W. Mudjanarko, S. M. S. Nugroho, and M. Hariadi, "Real time face recognition of video surveillance system using haar cascade classifier," Indonesian Journal of Electrical Engineering and Computer Science, vol. 21, no. 3, pp. 1389-1399, March 2021.

[2] P. M. Tank and D. G. Thakore, "A Fast Moving Object Detection Technique In Video Surveillance System," International Journal of Computer Science and Information Technologies, vol. 3, no. 2, pp. 3787-3792, 2012.

[3] K. A. Joshi and D. G. Thakore, "A Survey on Moving Object Detection and Tracking in Video Surveillance System," International Journal of Soft Computing and Engineering (IJSCE), vol. 2, no. 3, July 2012.
[4] Z. Jiang, L. Zhao, S. Li, and Y. Jia, "Real-time object detection method for embedded devices," IEEE Access, 2020.
[5] A. S. Geetha, "YOLOV4: A BREAKTHROUGH IN REAL-TIME OBJECT DETECTION," Huddersfield University, Feb. 2025.
[6] M. B. Patel, "Real-Time Violence Detection Using CNN-LSTM," Charotar University of Science and Technology, April 2021.
[7] K. B. Kwan-Loo, J. C. Ortíz-Bayliss, S. E. Conant-Pablos, H. Terashima-Marín, and P. Rad, "Detection of Violent Behavior Using Neural Networks and Pose Estimation," IEEE Access, August 2022.
[8] R. Kumar, "Deep Learning Strategies for Human Behavior Analysis in Automated Surveillance," Journal of AI Research, 2023.
[9] S. Lee, "A Comparative Study of CNN-LSTM and Skeletal Tracking for Action Recognition," IEEE Conference on Computer Vision, 2024.
[10] V. Singh, "Optimization of YOLO Architectures for Real-time Monitoring on Low-power Hardware," International Journal of Engineering, 2024.
Preeti and D. Kumar, “Feature selection for face recognition using DCT-PCA and Bat algorithm”, International Journal of Information Technology, Vol. 9, Issue 4, pp. 411–423, December 2017.