

A Comprehensive Review of YOLOv8-MediaPipe and Deep Learning-Based Multi-Person Fall Detection System

Sowjanya

sowjanyaashanthi18@gmail.com

Sannidhi Deepak Shetty

19792sannidhidshetty@gmail.com

Omkarayya Jagadish Kambalimath

Omkarayyajak@gmail.com

Sathwik Prabhu

Sathwik.r.prabhu@gmail.com

Department of Information Science and Engineering

Alva's Institute of Engineering and Technology, Moodbidri, Karnataka, India

Mr. Pradeep Nayak (Project Guide)

pradeepnayak@aiet.org.in

Abstract

Recent advancements in artificial intelligence and computer vision have enabled the development of intelligent surveillance and healthcare monitoring systems capable of tracking human behavior. However, traditional single-stage frameworks often suffer from identity switching, tracking loss, and reduced performance in crowded or occluded environments. This paper presents an exhaustive review of state-of-the-art vision-based multi-person fall detection and human activity recognition frameworks. We evaluate the shift from traditional heuristic-based systems to hybrid deep learning pipelines—specifically focusing on the integration of real-time object detectors (YOLOv8), multi-object tracking techniques (Byte-Track), top-down pose estimators (MediaPipe), and temporal sequence models (LSTMs). Furthermore, this survey examines mathematical spatiotemporal abstractions, data preprocessing pipelines, benchmark datasets, multi-tier software verification practices, operational alert deployment frameworks, and highlights remaining open challenges for real-time edge deployment in smart healthcare, elderly care, and intelligent surveillance applications.

Index Terms — YOLOv8, MediaPipe, multi-person tracking, fall detection, human activity recognition, computer vision, LSTM, pose estimation, intelligent surveillance, healthcare monitoring, deep learning, behavior analytics.

1 Introduction

Human behavior understanding and automated fall detection have become vital components in modern intelligent monitoring systems, particularly in healthcare environments, elderly care facilities, smart homes, and safety-critical workplaces. Falls are among the leading causes of severe injuries and fatalities, especially among elderly individuals, making real-time fall detection systems highly important for ensuring timely medical assistance and improving overall public safety. Delayed emergency response can significantly increase health risks and recovery time.

Conventional monitoring approaches often rely on wearable sensors, accelerometers, gyroscopes, or manual supervision, which may cause user discomfort, require continuous maintenance, suffer from battery dependency,

and display compliance limitations in real-world de-ployments. With the rapid growth of intelligent surveillance technologies and computer vision

systems, vision-based human activity monitoring has emerged as a scalable and non-intrusive alternative for behavior analysis and safety monitoring. Modern surveillance infrastructures continuously generate large volumes of video data, creating opportunities for automated sys-tems capable of understanding human posture, movement patterns, and abnormal events in real time.

However, crowded environments, occlusions, varying illumination conditions, and multi-person interactions introduce significant challenges for accurate activity recognition and fall detection. Transitioning vision systems out of controlled single-person environments into prac-tical scenes introduces distinct computer vision bottlenecks:

- **Identity Switching & Tracking Loss:** In multi-person tracking, overlapping human subjects frequently result in identity switching and inconsistent tracking, which breaks spatiotemporal sequence continuity.
- **Occlusion Sensitivity:** Partial or full bodily blockage by dynamic backgrounds, furni-ture, or other individuals sharply reduces the accuracy of skeletal keypoint extraction.
- **Spatiotemporal Ambiguity:** Traditional heuristic-based fall detection methods often rely only on spatial geometry or aspect-ratio analysis, which can misclassify rapid normal movements such as bending or sitting as fall events, resulting in high false-positive rates.

This review provides a structured evaluation of recent literature addressing these issues. We synthesize how modern top-down hybrid deep learning methodologies combine object tracking, localized keypoint extraction, and temporal sequence modeling to optimize multi-person safety networks.

2 Literature Review & Architectural Paradigms

The evolution of vision-based fall detection and human activity recognition frameworks can be categorized into three dominant structural paradigms across the literature.

2.1 Heuristic Shape and Silhouette Analysis

Early image-processing approaches utilized background subtraction, motion history analysis, sil-houette extraction, and geometric posture analysis to detect fall events. Bounding box aspect ratios (W/H proportions) were tracked to find sudden height changes. While computation-ally lightweight and capable of real-time execution on basic CPU hardware, these techniques were highly sensitive to illumination changes, dynamic backgrounds, occlusions, and viewpoint variations, limiting their reliability in real-world deployments. They lack fine-grained postural awareness, leading to high false-positive rates during benign activities like sitting down abruptly.

2.2 Bottom-Up Pose Estimation Frameworks

With the advent of deep learning, multi-person pose estimation emerged as a core tool. Bottom-up methods detect all skeletal joints or keypoints across an entire image frame simultaneously and subsequently group them into distinct human entities. While computationally efficient because inference latency scales well with crowded scenes, bottom-up systems experience severe performance drops in occluded spaces. Overlapping subjects cause joint cross-contamination, resulting in incorrect skeletal links, overlapping skeletons, and tracking errors.

2.3 Top-Down Hybrid Frameworks

To overcome the limitations of bottom-up systems, contemporary research introduces top-down hybrid frameworks integrating object detection, multi-object tracking, isolated pose estimation, and temporal sequence

modeling. This pipeline splits the multi-person detection problem into decoupled phases:

1. **Multi-Person Object Detection:** High-speed frameworks such as YOLOv8 enable real-time and highly accurate human detection within video streams, drawing bounding boxes around all individuals.
2. **Persistent Object Tracking:** To maintain persistent tracking identities across consecutive frames, object tracking algorithms such as ByteTrack are integrated. ByteTrack improves identity persistence and minimizes identity switching during partial occlusions and overlapping scenarios by associating every detection box.
3. **Isolated Region of Interest (ROI) Pose Estimation:** Each detected human bounding box is isolated as an independent Region of Interest (ROI) and passed individually to a pose framework like MediaPipe Pose. Unlike conventional bottom-up frameworks that suffer from skeletal overlap in crowded scenes, the ROI-based top-down approach ensures clean and non-overlapping landmark detection for multiple individuals simultaneously.
4. **Temporal Activity Classification:** The extracted spatiotemporal joint arrays are processed through sequential deep learning networks to analyze motion continuity and classify specific activities.

3 Mathematical Spatiotemporal Abstractions

The MediaPipe Pose framework extracts a localized skeleton matrix consisting of 33 discrete high-precision joints. Each joint k is represented by a spatial feature vector $P_k = (x_k, y_k, z_k, v_k)$, where x_k, y_k are normalized image plane coordinates, z_k represents depth, and v_k represents joint visibility probability. The human posture at time t is represented by a unified structural vector.

To classify temporal behavior sequences, a sliding window architecture compiles sequential posture states over a defined historical frame window size T .

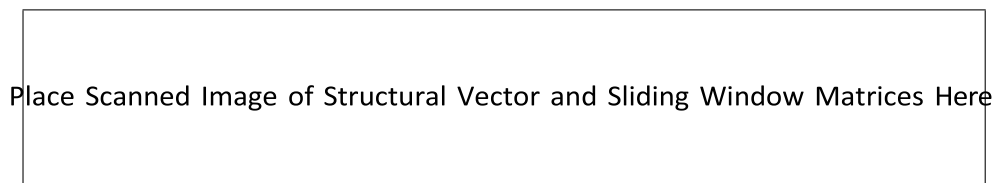


Figure 1: Scanned matrix representations for unified structural vectors and temporal sliding windows.

4 Comparative Analysis of Frameworks

A comprehensive comparative summary of these structural paradigms found across the surveyed literature is detailed in Table 1.

5 Data Preprocessing and Tuning Pipelines

A core finding across the literature is that deep learning model performance depends heavily on the consistency of the input data pipeline. Surveyed methodologies emphasize a standardized

Table 1: Comparative Analysis of Vision-Based Fall Detection Architectural Paradigms

Paradigm	Core Methodology	Key Advantages	Primary Limitations	Technical Complexity
Heuristic Analysis	Shape Bounding box ratios, silhouette subtraction.	Extremely low computational footprint; execution on CPU hardware.	High false-positive rates; fails under movement or illumination.	Simple
Bottom-Up Pose Estimation	Global keypoint detection across whole frame; grouping into entities.	Inference latency scales well with the number of people in the scene.	Severe cross-contamination overlap errors in crowds.	Complex
Top-Down Models	YOLOv8 + ByteTrack + MediaPipe + ROI pose extraction.	High tracking accuracy; eliminates multi-person skeletal overlap interference.	Multi-stage model execution creates high computational complexity at the edge.	Complex

sequence of frame-level preprocessing operations before running inference to maximize images' consistency and reduce processing overhead:

- **Resizing and Spatial Standardization:** Input video sequences are resized to standard resolution matrices to maintain bounded computational overhead and maximize inference throughput during real-time processing.
- **Normalization and Image Contrast Enhancement:** Frame normalization, histogram equalization, and adaptive contrast stabilization techniques are applied to mitigate illumination variability, camera shake, and sensor background noise.
- **Hyper-Parameter Tuning:** To achieve maximum classification accuracy while avoiding overfitting, modern frameworks systematically tune parameters—such as learning rate, batch size, optimizer choice, and training epochs—based on validation performance.

6 Temporal Modelling Techniques

Accurately distinguishing a real fall from a rapid daily action (like sitting down quickly or bending over to pick up an object) requires modeling human movement over a sequence of frames rather than inspecting static snapshots. The literature evaluates several prominent sequence modeling strategies:

6.1 Traditional Machine Learning Classification

Early vision frameworks applied traditional support vector machines (SVM) or random forest classifiers directly to static coordinate spatial features. However, these models lack temporal recurrence mechanisms, which can lead to high false-alarm rates during rapid, non-critical posture transitions like sitting down quickly or bending.

6.2 Long Short-Term Memory (LSTM) Networks

Lightweight recurrent neural networks, specifically LSTMs, excel at capturing temporal dependencies and sequence continuity over time.

Place Scanned Image of LSTM Gating Equations (4-9) Here

Figure 2: Scanned system equations for LSTM internal gate mechanisms.

By calculating continuous coordinate trajectories across moving frame sequences, LSTMs learn the continuous acceleration and velocity changes that characterize an actual fall. This temporal context allows the system to accurately differentiate between a genuine medical emergency and normal daily activities, lowering false positives.

6.3 3D Convolutional Neural Networks (3D-CNNs)

3D-CNN architectures extract spatial and temporal features simultaneously by applying 3D convolution kernels across consecutive video blocks. While they yield high detection accuracy, their massive parameter density increases inference latency and reduces system responsiveness on resource-constrained edge hardware.

7 Benchmarking Datasets

To evaluate and compare activity recognition models fairly, researchers rely on public, standardized human activity and fall detection datasets. The most widely utilized benchmarks include:

- **UP-Fall Detection Dataset:** Provides balanced activity samples of normal movements alongside simulated fall events under different environmental conditions. It is highly useful for basic behavioral training but often lacks complex, multi-person crowd sequences.
- **UR Fall Detection Dataset:** Focuses heavily on sequence tracking, depth maps, and spatial geometry variations during fall transitions.
- **GMDCSA-24 Dataset:** A modern repository featuring various video samples designed specifically for training deep learning models in complex surveillance environments with varying lighting and clutter.

8 Verification & System Testing Paradigms

Comprehensive review papers emphasize that system validation must go beyond standalone model accuracy metrics; it must evaluate end-to-end software functionality. Robust implementations apply a multi-tier testing paradigm including Unit Testing of individual modules, Integration Testing across frames, and System/Edge Testing of the complete monitoring pipeline. Standard test case outcomes synthesized from literature are summarized in Table 2.

9 Real-Time Interactive Analytics Alerts

Modern intelligent surveillance architecture requires more than just accurate event detection; it must deliver operational utility. Current research trends highlight two vital operational layers:

- **Automated Alert Mechanisms:** When a fall event is confirmed by the temporal classification network, the system triggers instant safety alerts via audio alarms, on-screen warnings, or external automated notification mechanisms to ensure timely intervention.

Table 2: Multi-Tier System Testing Paradigms and Case Analysis

Test Case Condition	Expected System Behavior	Status
Invalid video stream format	Trigger unsupported format error message.	Pass
Camera monitoring without active permissions	Display immediate warning layout requiring camera access.	Pass Pass Pass Pass
Low-resolution / low-quality input stream	Flag warning indicating reduced detection accuracy risks.	
Simultaneous multi-person occupancy	Map all individuals concurrently using unique persistent IDs.	
Partial body occlusion or trajectory cross	Maintain persistent ID tracking without dropping or switching identities.	

- **Web Analytics Dashboards:** Deploying interactive front-ends using lightweight tools like Streamlit allows managers to monitor live video overlays, keep automated event logs, view behavior timelines, and review long-term graphical summaries. This significantly enhances situational awareness for healthcare workers and administrators.

10 Open Challenges and Future Directions

Despite the high performance reported by standalone deep learning frameworks, several systemic challenges remain open for future research:

- **Edge Compute Constraints:** Running multi-person YOLOv8 detectors, extracting multiple concurrent MediaPipe keypoint sets, and running downstream LSTM evaluations simultaneously generates a large computational footprint. Continuous execution can cause latency delays on resource-constrained hardware. Future research must focus on model quantization, pruning, and hardware acceleration techniques to maintain low latency.
- **Extreme Environmental Invariance:** System performance can degrade under severe shadows, poor night-time lighting, low camera resolutions, or unusual viewing angles. Developing adaptive preprocessing layers and expanding training datasets with diverse environmental conditions is necessary to improve generalization.
- **Mitigating False Alarms:** Differentiating between a controlled quick sit-down, a dynamic bend, and an actual injurious fall remains a fine line. Further work on multi-view camera fusion and hybrid spatiotemporal transformers may offer better sequence understanding over standard recurrent networks.

11 Conclusion

This review has synthesized modern deep learning architectures for multi-person tracking and real-time fall detection. The literature shows a clear paradigm shift away from traditional geometric heuristics and toward top-down hybrid frameworks. By pairing object detectors like YOLOv8 with robust tracking mechanisms (ByteTrack) and isolated pose estimators (MediaPipe), contemporary systems successfully mitigate historical issues like tracking loss and identity switching in crowded settings. When combined with sequence modelers like LSTMs, these systems provide the low latency and reliability needed for deployment in smart healthcare

facilities, elderly care homes, and autonomous safety networks. Future advancements will de-pend on optimizing these multi-stage models for edge computing and ensuring robustness across unpredictable real-world environments.

References

- [1] T. Liu and S. Lee, “A review of fall detection systems based on deep learning,” *IEEE Access*, vol. 7, pp. 142409–142421, 2019.
- [2] M. Mubashir, L. Shao, and L. Seed, “A survey on fall detection: Principles and approaches,” *Neurocomputing*, vol. 100, pp. 144–152, 2013.
- [3] M. Yu et al., “A survey on vision-based fall detection,” *Biomedical Signal Processing and Control*, vol. 4, no. 1, pp. 45–56, 2009.
- [4] World Health Organization, “Falls fact sheet,” 2021.
- [5] E. Alam et al., “GMDCSA-24: A dataset for human fall detection in videos,” *Data in Brief*, vol. 57, 2024.
- [6] W. Liao, T. Chen, and J. Yang, “A novel fall detection system using YOLOV3 and pose estimation,” *Proc. IEEE ICIP*, pp. 3704–3708, 2020.
- [7] S. Chien and T. Li, “A real-time fall detection system using pose estimation and machine learning,” *Journal of Healthcare Engineering*, vol. 2019, pp. 1–9, 2019.
- [8] J. Redmon et al., “You Only Look Once: Unified, real-time object detection,” *Proc. IEEE CVPR*, pp. 779–788, 2016.
- [9] A. Sixsmith and N. Johnson, “A smart sensor to detect the falls of the elderly,” *IEEE Pervasive Computing*, vol. 3, no. 2, pp. 42–47, 2004.
- [10] N. Thome, S. Miguet, and S. Ambellouis, “A real-time, multiview fall detection system,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1522–1527, 2008.
- [11] S. Chien and T. Li, “Pose estimation and machine learning for fall detection,” *Journal of Healthcare Engineering*, 2019.
- [12] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [13] Y. Zhang et al., “ByteTrack: Multi-object tracking by associating every detection box,” *ECCV*, 2022.
- [14] Google, “MediaPipe: Cross-platform framework for multimodal applied ML pipelines,” 2021.
- [15] C. Lin and H. Wang, “Fall detection based on deep learning with 3D convolutional neural networks,” *Proc. AIVR*, pp. 347–352, 2018.
- [16] UP-Fall Detection Dataset, “UP-Fall Detection Dataset,” 2020.
- [17] UR Fall Detection Dataset, “UR Fall Detection Dataset,” 2019.