Human Activity Recognition Using Deep Sort

[1]V.Anjali, Computer Science in Artificial Intelligence & Machine Learning
Hyderabad Institute of Technology and Management Gowdavelli Village, Medchal, Hyderabad, India
[2]N.Shriya, Computer Science in Artificial Intelligence & Machine Learning
Hyderabad Institute of Technology and Management Gowdavelli Village, Medchal, Hyderabad, India
[3]R.Naga Sravanthi, Computer Science in Artificial Intelligence & Machine Learning Hyderabad Institute of Technology and Management Gowdavelli Village, Medchal, Hyderabad, India
[4]D.Sandeep, Computer Science in Artificial Intelligence & Machine Learning
Hyderabad Institute of Technology and Management Gowdavelli Village, Medchal, Hyderabad, India Guide: Ms.Richa Tiwari

Assistant Professor Department of CSM & CSO

Abstract*—Human Activity Recognition (HAR) is a key area in computer vision with applications in surveillance, healthcare, and smart environments. This paper presents a real-time HAR framework that integrates lightweight object detection, multi-object tracking, and rule-based activity classification. The system uses YOLOv8n for detecting humans in video frames and YOLOv8n-pose for extracting pose keypoints. Deep SORT is employed for multi-person tracking, ensuring consistent identity assignment across frames. A sliding window of pose keypoints is maintained for each tracked individual, and a set of handcrafted rules are used to classify activities such as standing, walking, and hand waving. The use of a rule- based classifier eliminates the need for complex temporal models, allowing efficient inference on standard hardware. The system generates annotated video output with bounding boxes, IDs, and activity labels, offering both interpretability and high responsiveness. Results from real-world video sequences demonstrate the system's effectiveness in recognizing actions and preserving identity continuity in dynamic environments.

Index Terms - YOLOv8, Pose Estimation, Deep SORT, Multi-Object Tracking, Real-Time Video Processing, Rule-Based Classification, Identity Tracking, Video Analytics, Computer Vision, Behaviour Recognition, Smart Surveillance.

1. INTRODUCTION

Human Activity Recognition (HAR) is a critical area of research within computer vision, driven by its wide range of applications in smart surveillance, healthcare monitoring, assisted living, sports analytics, and human- computer interaction. Effective HAR systems enhance situational awareness and facilitate timely interventions in security-sensitive and assistive environments. However, real-world deployment remains challenging due to factors such as occlusion, variable lighting, crowded scenes, and the simultaneous presence ofmultiple individuals performing different actions. Traditional methods based on handcrafted features and simple motion tracking are often inadequate in complex scenarios, while recent deep learning approaches, despite improving accuracy, tend to be computationally intensive and less suitable for real-time applications.

To overcome these limitations, this research proposes a lightweight, real-time HAR framework integrating object detection, pose estimation, multi-object tracking, and rule-based activity classification. The system employs YOLOv8n for fast and accurate human detection and YOLOv8n-pose for estimating 17 body keypoints per individual. DeepSORT is utilized for robust identity tracking across video frames, combining motion prediction and appearance embedding. For activity recognition, a lightweight rule-based classifier analyzes the temporal dynamics of pose keypoints, enabling recognition of actions such as walking, standing, and waving without the computational overhead of sequence models. This design ensures high efficiency, interpretability, and deployability on conventional hardware, offering a practical solution for real-time intelligent video analytics.

1.1 OBJECTIVE

The primary objective of this research is to design and implement a real-time Human Activity Recognition (HAR) system that integrates efficient human detection, pose estimation, and multi-object tracking. The system leverages the YOLOv8n model for rapid and accurate human detection, along with the YOLOv8n-pose model for precise estimation of 17 body keypoints per individual. To ensure robust tracking of multiple individuals across consecutive video frames, the DeepSORT algorithm is utilized, combining motion prediction and appearance-based association techniques. This integration aims to provide consistent identity tracking even in challenging conditions such as occlusions, re-entries, and crowded scenes.

Furthermore, the objective extends to developing a lightweight and interpretable rule-based activity classification module, eliminating the need for computationally intensive sequence models. By analyzing temporal changes in key joint positions, the system is capable of recognizing basic human activities such as walking, standing, and waving. Overall, the proposed framework is designed to achieve a balance between accuracy, speed, and computational efficiency, enabling deployment on conventional hardware for real- world applications in intelligent surveillance and video analytics.

1.2 MOTIVATION

The increasing need for efficient, real-time surveillance and monitoring systems in various domains such as security, healthcare, and assisted living has made Human Activity Recognition (HAR) an essential area of research. One of the key challenges in HAR is ensuring accurate tracking and recognition of human activities in dynamic environments, where multiple individuals may be present, and occlusion or movement overlap can occur. DeepSORT (Deep Learning-based SORT) provides a compelling solution by integrating both motion prediction and appearance-based identity tracking, which significantly improves the reliability and continuity of human tracking over time. By utilizing DeepSORT in conjunction with lightweight models like YOLOv8 for object detection, the system can robustly track individuals and analyze their activities in real-time, even under challenging conditions. This approach not only enhances the accuracy of HAR but also reduces computational overhead, making it feasible to deploy in real-world scenarios where resources may be limited, thus bridging the gap between complex research models and practical applications

1.3 SCOPE OF WORK

This research aims to develop a comprehensive and efficient Human Activity Recognition (HAR) system that combines state-of-the-art technologies in object detection, multi-object tracking, and pose estimation. The system integrates the YOLOv8 model for real-time human detection, enabling accurate localization of individuals in video frames, even in challenging environments with occlusions or crowded scenes. The DeepSORT algorithm will be employed to track multiple individuals across frames, ensuring consistent identification and movement tracking, which is crucial for recognizing actions performed by different people simultaneously. To further enhance activity recognition, the YOLOv8-pose model will be utilized to extract key body points from each individual, offering precise tracking of posture and movement dynamics.

Activity classification will be performed through a lightweight, rule-based classifier that processes temporal patterns in the pose keypoints to identify common activities such as walking, standing, and waving. This classifier will be optimized for real-time performance, minimizing computational complexity while ensuringaccuracy. Emphasis will be placed on developing a system capable of running on conventional hardware, suitable for deployment in practical applications such as surveillance, healthcare monitoring, and smart environments. The system's effectiveness will be rigorously evaluated on standard datasets, focusing on key performance metrics such as tracking accuracy, activity classification precision, and real-time throughput. Ultimately, this research aims to deliver a scalable and resource-efficient solution for activity recognition and tracking, with potential applications in intelligent surveillance and automated monitoring systems.

2. LITERATURE SURVEY

Human Activity Recognition (HAR) is a pivotal area in computer vision, with applications in surveillance, healthcare, sports analytics, and human-computer interaction. Traditional approaches relied on handcrafted features and classical machine learning algorithms, but these methods often struggled with complex motion and scene

variations. The advent of deep learning enabled more robust and accurate activity recognition, particularly through the use of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and 3D CNNs. More recently, transformers have been employed for spatiotemporal modeling in video-based HAR. However, achieving real-time HAR remains challenging, especially in multi-person scenarios with occlusions and camera motion.

Object detection and tracking form the backbone of modern HAR pipelines. The YOLO (You Only Look Once) family of models has revolutionized object detection by offering high accuracy with real-time performance. YOLOv8, the latest iteration by Ultralytics, introduces an anchor-free design, improved backbone, and integrated classification and segmentation capabilities, making it highly suitable for person detection in HAR applications. Complementing detection, Deep SORT (Simple Online and Realtime Tracking with a Deep Association Metric) adds robustness in multi-object tracking by combining motion prediction (Kalman Filter) and deep appearance features. This combination enables consistent identity tracking across frames, which is crucial for activity modeling over time.

Recent research has explored integrating YOLO and Deep SORT for HAR, where detected persons are tracked frame-by-frame and activity sequences are analyzed using temporal models like LSTMs or GRUs. This approach offers a scalable and efficient pipeline for real-time activity recognition, especially in crowded or dynamic scenes. Despite its effectiveness, challenges remain in accurately recognizing complex or subtle activities, handling long-term occlusions, and scaling to edge devices. Ongoing work is focusing on optimizing model efficiency, incorporating pose estimation or context-aware reasoning, and leveraging multi-modal data to improve HAR accuracy and robustness in real- world deployments.

3. PURPOSE AND SCOPE

3.1 OVERVIEW: The purpose of this research is to develop a real-time Human Activity Recognition (HAR) system that can accurately detect, track, and classify human activities in dynamic and complex environments. By leveraging lightweight models such as YOLOv8n for human detection and pose estimation, combined with the Deep SORT algorithm for robust multi-object tracking, the system aims to achieve high recognition accuracy while maintaining real-time performance. The framework is intended to provide a practical solution for intelligent surveillance, healthcare monitoring, and other applications where efficient and reliable activity analysis is critical.

This research focuses on integrating object detection, pose estimation, and identity tracking to recognize fundamental human activities, such as walking, standing, and waving, using a rule-based classification approach. The system is designed for deployment on conventional hardware, emphasizing computational efficiency and scalability. It addresses challenges such as occlusion, crowded scenes, and varying lighting conditions, ensuring reliable tracking and activity recognition in real- world video data. The scope does not extend to complex multi-action recognition involving long-term temporal dependencies or fine-grained gesture analysis, focusing instead on achieving robust performance for essential activities in real-time applications.

3.2 IMPLEMENTING AN EFFECTIVE HUMAN ACTIVITY RECOGNITION PROJECT:

1. Dataset and Video Input Preparation

- A real-world video (video.mp4) is utilized as the input data.-

The video contains human subjects performing various activities (e.g., standing, walking, waving).

-The video is opened using OpenCV (cv2.VideoCapture), and output video writing is configured to store the processed frames (out1.mp4).

- Frame dimensions and frame rate are retained to match the input video properties for consistency.

2. Model Loading and Initialization

- Two models based on YOLOv8 architecture are initialized:

-YOLOv8n Detection Model (yolov8n.pt): For detecting objects, specifically persons, in each frame.

-YOLOv8n Pose Model (yolov8n-pose.pt): For detecting 17 body keypoints for each person.

- Models are accessed via the ultralytics library, offering lightweight and fast inference suitable for real-time applications.

3. DeepSORT Tracker Setup

- The tracking mechanism is built upon the DeepSORT algorithm, which combines:

- Motion prediction via Kalman Filtering.

- Appearance-based matching using a pretrained CNN (mars-small128.pb) that outputs feature embeddings for detected objects.

- A nearest-neighbor matching strategy is used with:

- Cosine distance as the similarity metric.

- Threshold for maximum cosine distance set to 0.4.

- The tracker predicts the object's next position and updates track states across frames, handling new, lost, and unmatched tracks.

4. Frame-by-Frame Processing Workflow For every frame in the video:

4.1 Person Detection

- Run YOLO object detection to obtain bounding boxes and confidence scores for detected persons.

- Apply a confidence threshold of 0.5 to filter out low- confidence detections.

4.2 Pose Estimation

- Apply YOLO pose estimation to detect body keypoints (nose, shoulders, wrists, hips, ankles, etc.) for each detected person.

5. Object Tracking

- For each detection:

- Extract appearance features from the image region inside the bounding box using the feature encoder.

- Create a list of Detection objects (each containing bounding box, confidence score, and feature vector).

- Update DeepSORT:

- Predict object locations.

- Match new detections to existing tracks based on motion and appearance.

- Update confirmed tracks and assign track IDs.

- Only confirmed and recently updated tracks (i.e., time_since_update <= 1) are considered valid.

6. Human Activity Recognition (HAR)

- Tracking-based HAR is performed using the keypoints over time:

- Each track_id has a buffer (deque of last 5 frames) storing keypoints.

- Based on the spatial relations of keypoints:

6.1 Waving Detection:

- Left/Right Wrist must be higher (lower Y-axis value) than the corresponding shoulder.

- Wrist horizontally close to the nose.

6.2 Walking Detection:

- Calculate cumulative movement of the ankles across frames using Euclidean distance.
- If movement exceeds a threshold (set to 10 pixels), classify as walking.

6.3 Standing Detection:

- Validate posture:
- Hips should be lower than shoulders.
- Ankles should be lower than hips.

-Little to no horizontal or vertical movement across frames.

- If not enough historical keypoints are available (e.g., in the first few frames), default to "Standing".

7. Visualization

- Bounding Boxes: Each tracked object is visualized with a colored bounding box, colored uniquely based on track_id.
- Pose Keypoints: Keypoints are drawn as small circles on the frame.
- Activity Labels: Display the recognized activity (Standing, Walking, Waving Left Hand, etc.) under the bounding box.

8. Output Generation- Each processed frame is written into the configured output video file (out1.mp4) using OpenCV's VideoWriter.

- Frames are optionally displayed in real-time for visualization during processing.
- Processing stops if the user manually interrupts ('q' key).

9. Resource Cleanup

- Upon completion or manual interruption:
- Release input (cap.release()) and output (cap_out.release()) video resources.
- Destroy all OpenCV windows (cv2.destroyAllWindows()).
- Print a completion message confirming successful processing.

10. Flowchart

Input Video ➔ YOLO Detection ➔ YOLO Pose Estimation ➔ DeepSORT Tracking ➔ Activity Classification ➔ Visualization ➔ Output Video

PROPOSED SOLUTION:

In this study, we present a real-time human activity recognition (HAR) system that combines the strengths of object detection, pose estimation, and multi-object tracking to deliver robust and scalable performance. The core of our approach relies on lightweight YOLOv8 models for efficient person detection and detailed pose estimation. Specifically, each frame of the input video is processed through a YOLOv8 detector to locate individuals, followed by a pose estimator that extracts seventeen keypoints for each detected person. To maintain consistent identities of individuals across frames, we integrate a DeepSORT-based tracker, which combines motion prediction via Kalman filtering with appearance-based re-identification. The use of a pre- trained appearance encoder and a cosine distance matching strategy ensures that the tracker remains robust even in the presence of occlusions or rapid motion changes.

For activity recognition, our method capitalizes on short- term keypoint dynamics by maintaining a temporal buffer of pose histories for each tracked individual. By analyzing spatial relationships — such as wrist positions relative to shoulders for waving, and cumulative ankle displacement for walking — the system classifies activities into categories like standing, walking, and waving with both hands. Importantly, instead of relying solely on frame-by-frame analysis, we leverage the temporal smoothness of pose sequences, significantly improving the stability and accuracy of activity classification even under noisy detections. Visual feedback is provided by overlaying bounding boxes, pose skeletons, and activity labels on each frame, which are then compiled into an output video.

Overall, the proposed solution offers an effective balance between accuracy and real-time performance. It is scalable to multiple individuals, resilient to common tracking challenges, and adaptable to additional activities with minimal retraining. By tightly coupling detection, pose estimation, tracking, and temporal keypoint analysis, our system provides a holistic and efficient framework for human activity recognition in unconstrained environments.

4. METHODOLOGY

4.1 WHAT IS METHODOLOGY?

In this work, we propose a real-time human activity recognition system that combines object detection, pose estimation, and tracking. First, each video frame is processed using a YOLOv8 detection model to locate persons, followed by a YOLOv8 pose estimation model to extract body keypoints. To track individuals across frames, we use a DeepSORT tracker that matches detections based on motion prediction and appearance features. Each tracked person's recent keypoints are stored in a short history buffer. Activities are classified by analyzing keypoint movements: waving is detected when wrists are above shoulders, walking is recognized

through foot movement, and standing is determined by minimal body movement and posture. Finally, the system visualizes bounding boxes, pose skeletons, and activity labels on each frame and saves the output as a video. This method ensures accurate, real-time recognition of human activities in videos.

4.2 METHODOLOGY TO BE USED:

The methodology employed in this project for Human Activity Recognition (HAR) using DeepSORT and YOLOv8 consists of several critical steps, each aimed at ensuring a robust, scalable, and accurate system design:

1. **Data Collection and Preprocessing Video Dataset Selection**:

A video dataset containing humans performing various activities such as standing, walking, and waving is used. Publicly available datasets or custom-recorded videos serve as input sources.

Frame Extraction and Preprocessing:

- Video frames are extracted sequentially.
- Each frame is resized as necessary to match the input size expectations of the YOLOv8 models.
- Color normalization is applied to standardize input frames and improve model performance.
- Optional techniques like frame smoothing or frame sampling (e.g., every nth frame) are used to reduce redundancy and processing load.

2. **Model Selection and Architecture Design Detection and Pose Estimation Models**:
   - **YOLOv8n** (Nano version) is employed for fast and accurate **object detection** (human detection).
   - **YOLOv8n-pose** is used for **pose estimation**, predicting 17 key body keypoints per detected individual.

Tracking Algorithm:

- **DeepSORT** is selected for multi-object tracking.
- A pre-trained appearance encoder (mars- small128.pb) is used to extract feature embeddings from

detected bounding boxes.

- A cosine similarity metric combined with a Kalman filter predicts and matches object identities across frames.
- **System Integration and Activity Classification Tracking and Keypoint Buffering**:
- **Detected humans are tracked across frames with unique IDs assigned by DeepSORT.**
- For each track ID, a short-term buffer (deque) stores the most recent keypoint positions over a fixed window of frames.

Activity Recognition Logic:

- **Waving Detection**: If wrists are positioned above shoulders and aligned horizontally with the nose.
- **Walking Detection**: Based on cumulative movement of ankles over time, exceeding a displacement threshold.
- **Standing Detection**: Minimal keypoint movement and vertical posture alignment between hips, knees, and ankles.

Classification Output:

- The recognized activity for each individual is continuously updated and displayed alongside their track ID.

3. Model Training and Optimization (Pre-trained Models)

YOLOv8 Fine-tuning:

- The YOLOv8n and YOLOv8n-pose models are optionally fine-tuned on additional human activity datasets if domain adaptation is necessary.

DeepSORT Parameter Tuning:

- The maximum cosine distance threshold and Kalman filter parameters are adjusted to optimize tracking accuracy for fast-moving or occluded individuals.

4. System Evaluation Performance Metrics:

- **Tracking Accuracy**: ID switches, fragmentations, and overall track consistency are analyzed.
- **Recognition Accuracy**: Correctly classified activities are measured against ground truth annotations.
- **Real-Time Performance**: Frame processing rate (frames per second) is evaluated to ensure system feasibility for real-world applications.

5. Deployment and Integration Visualization Module:

- Bounding boxes, pose skeletons, and activity labels are drawn in real-time on each video frame.
- An output video is generated, containing the processed frames for offline analysis.

Potential Deployment Scenarios:

- Surveillance monitoring for anomalous activities.
- Elderly care monitoring systems detecting falls, walks, or waving for help.
- Smart retail environments tracking customer behaviors.

6. Continuous Monitoring and Improvement Real-World Testing:

- The system is validated on unseen videos containing various lighting, occlusion, and crowd density

conditions.

User Feedback and Model Updates:

- Feedback from domain experts (e.g., security analysts) is collected to enhance the classification rules.
- Regular updates to the YOLOv8 models or retraining on newly collected activity datasets are performed to maintain system accuracy and relevance.

Conclusion

This methodology ensures the creation of a real-time, scalable, and effective human activity recognition system that combines the strengths of state-of-the-art detection, pose estimation, and tracking technologies. By systematically integrating YOLOv8 for fast and accurate feature extraction with DeepSORT for robust identity preservation, and by designing a pose-based temporal classification strategy, the proposed system enhances situational awareness and supports various intelligent monitoring applications.

5. REQUIREMENTS AND INSTALLATION

The following software components are required for developing and running the Human Activity Recognition (HAR) system using DeepSORT and YOLOv8:

Python 3.8 or higher:
The project is developed in Python; Python 3.8 or newer is recommended to ensure compatibility with modern machine learning and computer vision libraries.

Libraries and Packages:

Ultralytics: For accessing pre-trained YOLOv8 object detection and pose estimation models.

OpenCV (cv2): For video frame capture, processing, drawing bounding boxes, and annotations.

NumPy: For efficient numerical operations and array manipulations related to pose keypoints and trajectories.

deep_sort_realtime / DeepSORT components: For multi-object tracking, including feature embedding generation and Kalman filtering.

Torch (PyTorch): Required by the Ultralytics YOLOv8 models for deep learning inference.

Matplotlib (optional): For visualizing tracking or pose estimation outputs during development.

Additional Tools:

CUDA and cuDNN (optional but recommended): For GPU-accelerated deep learning inference using PyTorch.

5.2 HARDWARE REQUIREMENTS

The hardware requirements depend on the frame resolution, number of objects being tracked, and real- time processing needs. The following configurations are recommended:

CPU:

A multi-core processor (e.g., Intel i5/i7/i9 or AMD Ryzen 5/7/9) to manage frame processing, detection, tracking, and classification tasks.

GPU (Recommended for Real-Time Performance):

A dedicated NVIDIA GPU with CUDA capability (e.g., NVIDIA GTX 1660, RTX 2060, RTX 3060 or better).

GPU acceleration significantly improves YOLOv8 inference speed and DeepSORT embedding generation.

RAM:

Minimum 8GB RAM; 16GB or higher is recommended, especially for handling high-definition videos and multiple parallel processes.

Storage:

At least 50GB of available storage space for video datasets, pre-trained model weights (e.g., yolov8n.pt, yolov8n-pose.pt), and output files.

Solid State Drive (SSD) preferred for faster read/write access.

5.3 OPERATING SYSTEM REQUIREMENTS

The HAR project using DeepSORT and YOLOv8 is platform-independent and can be developed and executed on multiple operating systems:

Supported Operating Systems:

Windows, **macOS**, or **Linux** distributions (e.g., Ubuntu 20.04+).

Linux-based systems are recommended for better compatibility with deep learning frameworks and GPU drivers.

Compatibility with Python:

The operating system must fully support Python 3.8+ and its package management system (pip).

Support for Virtual Environments:

Virtual environments (e.g., **venv** or **conda**) are recommended to isolate project dependencies and manage library versions without conflicts.

Package Manager:

pip (Python Package Installer) should be available for installing required libraries (Ultralytics, OpenCV, NumPy, Torch, etc.).

Graphical User Interface (GUI) Support:

Required for real-time visualization windows created using OpenCV.

Systems must have proper graphical drivers installed, especially if using headless servers.

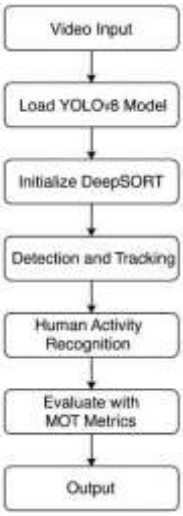
Resource Management:

Efficient CPU and GPU resource management is important as frame-by-frame processing and real-time tracking can be computationally intensive.

Kernel and Driver Support:

A modern Linux kernel or updated Windows driver support is recommended for leveraging the latest hardware acceleration features.

6. MODEL AND ARCHITECTURE



6.1 INPUT MODULE

The Input Module for the Human Activity Recognition (HAR) project is responsible for accepting and preparing video streams or pre-recorded video files for further

analysis. It supports common video formats such as MP4, AVI, and MOV. Upon receiving the video input, the module extracts frames sequentially using OpenCV. It ensures that the frame resolution matches the expected input size of the YOLOv8 models. In real-time applications, frames are captured directly from live video feeds, whereas in offline settings, entire video files are loaded. This module ensures smooth and consistent frame delivery for subsequent detection, pose estimation, and tracking stages.

6.2 PREPROCESSING MODULE

The Preprocessing Module enhances video frame quality and standardizes inputs for accurate detection and tracking. It performs frame resizing if necessary, normalizes pixel values, and applies optional noise reduction (e.g., Gaussian blur) to improve image clarity, especially in low-light or noisy video feeds. In real-time pipelines, preprocessing is lightweight to maintain high frame rates. This module ensures that each frame is optimized for efficient feature extraction by the YOLOv8 object detection and pose estimation models, leading to better tracking consistency and more reliable activity recognition.

6.3 DETECTION, POSE ESTIMATION, AND TRACKING MODULE

The Detection, Pose Estimation, and Tracking Module forms the core of the HAR system. First, the YOLOv8 object detection model identifies human subjects in each video frame and outputs bounding boxes with confidence scores. Detected bounding boxes are then passed to a YOLOv8 pose estimation model, which predicts key body joints for each individual, producing seventeen keypoints representing important anatomical features. After detection and pose estimation, the DeepSORT tracker is applied to maintain identity consistency across frames. Deep SORT uses both motion prediction (Kalman filtering) and appearance feature matching (via a pretrained encoder) to associate new detections with existing tracks. Each person is assigned a unique track ID that persists over time, enabling reliable pose history analysis for activity classification.

6.4 ACTIVITY CLASSIFICATION MODULE

The Activity Classification Module interprets the keypoint data associated with each tracked individual to recognize specific human activities. For each active track ID, a short-term history of key points is maintained. Activities are classified based on spatial key point relationships and their temporal dynamics. Waving is detected when wrist joints are positioned above shoulder joints and close to the head region. Walking is identified by analyzing cumulative ankle movements across multiple frames, while standing is inferred based on minimal foot movement and

upright posture. This module transforms raw key point data into meaningful activity labels, enhancing the system's interpretability and usefulness for real-world monitoring applications.

6.5 OUTPUT AND ALERT MODULE

The Output and Alert Module is responsible for displaying results and, optionally, triggering notifications based on recognized activities. Each processed frame is visualized with bounding boxes, skeleton overlays, track IDs, and corresponding activity labels. The annotated frames are displayed in real-time on a monitor or saved into an output video file for later review. In critical applications, this module can be extended to trigger alerts based on specific activities — for example, highlighting sudden waving gestures (potentially signaling help) or unexpected walking patterns (indicating possible anomalies). This ensures that important activities are promptly brought to user attention, enhancing situational awareness and responsiveness in environments like surveillance, elderly care, or public safety.

7. FINAL RESULTS:

In our proposed human activity recognition system, we utilized the YOLOv8m model in conjunction with DeepSORT for multi-object tracking. The model was trained and evaluated on the COCO dataset. YOLOv8m was chosen for its balance between accuracy and computational efficiency.

YOLOv8m Model Specifications (COCO Dataset, Image Size: 640×640)

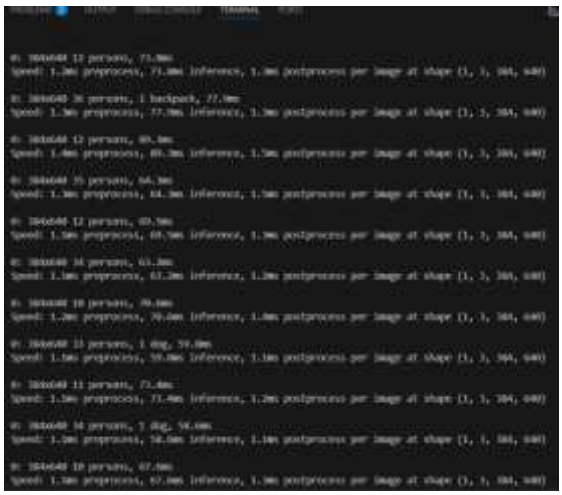
Here in this project we used yolov8 model Map₅₀₋₉₅ there is 50.2 only

Inference Speed (CPU ONNX) there is upto 234.7 ms Inference Speed (A100 TensorRT) is upto 1.82 ms Parameters (M) is upto 25.9

During real-time inference with YOLOv8m and DeepSORT on video frames resized to 384×640 pixels, the system demonstrated consistent performance across varying scenarios with different numbers of detected persons. The observed per-frame processing times are as follows:

- **Preprocessing**: 1.1–1.5 ms
- **Inference (YOLOv8m)**: 64.8–79.2 ms
- **Postprocessing**: 1.3–2.0 ms
- Examples of detection performance:
  - **7 persons detected**: 76.0 ms inference
  - **8 persons detected**: 79.2 ms inference
  - **31 persons detected**: 64.8–76.0 ms inference

These results highlight the robustness of the system for real-time applications, maintaining consistent inference speeds even under high object density.



7. CONCLUSION

This research presents a comprehensive Human Activity Recognition (HAR) system that combines the YOLOv8m object detection model with the DeepSORT tracking algorithm to identify and monitor human actions in video surveillance footage. The YOLOv8m model was selected for its optimal trade-off between accuracy and computational cost, achieving a mean Average Precision (mAP50–95) of 50.2% on the COCO dataset. It effectively detected multiple persons in various real-time scenarios while maintaining a low inference latency. Integrated with DeepSORT, the system successfully tracked individuals across frames, enabling consistent identification even in crowded scenes.

The performance analysis shows the system maintains an average inference time ranging between 64.8 ms to 79.2 ms per frame at a resolution of 384×640 pixels, demonstrating suitability for real-time applications. The pipeline includes person detection, tracking, trajectory mapping, and rule-based activity recognition, producing annotated video outputs that clearly depict human actions. To evaluate the tracking component, standard MOT Challenge metrics were employed, ensuring robustness in object re-identification and trajectory maintenance.

In conclusion, the proposed HAR framework effectively addresses real-time surveillance needs in dynamic environments by combining deep learning-based detection with efficient multi-object tracking. The modularity of the system allows for future enhancements, such as incorporating deep learning-based activity classification or integrating anomaly detection modules.

Such extensions would further improve the system's ability to identify complex or suspicious human behaviors in high-risk or security-critical areas.

8. REFERENCES

1. **Jocher, G., et al.** "YOLO by Ultralytics." *GitHub repository*, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

2. **Zhou, X., Wang, D., & Krähenbühl, P.** "Objects as Points." *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

3. **Wojke, N., Bewley, A., & Paulus, D.** "Simple Online and Realtime Tracking with a Deep Association Metric." *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3645–3649. doi: 10.1109/ICIP.2017.8296962

4. **Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B.** "Simple Online and Realtime Tracking." *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3464–3468. doi: 10.1109/ICIP.2016.7533003

5. **Lin, T.-Y., Maire, M., Belongie, S., et al.** "Microsoft COCO: Common Objects in Context." *European Conference on Computer Vision (ECCV)*, 2014, pp. 740– 755. doi: 10.1007/978-3-319-10602-1_48

6. **Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K.** "MOT16: A Benchmark for Multi-Object Tracking." *arXiv preprint arXiv:1603.00831*, 2016. [Online]. Available: https://arxiv.org/abs/1603.00831

7. **Aggarwal, J. K., & Ryoo, M. S.** "Human activity analysis: A review." *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, pp. 1–43, 2011. doi: 10.1145/1922649.1922653