Joint Tracking and Segmentation of Multiple Targets: Real-Time Approach with YOLOv8 and DeepSort

Group ID: 20

Satyam Sharma (B22CS047)
Sanjeet Athawale (B22EE014)
Devvrat Deval (M25MER008)

Indian Institute of Technology Jodhpur

GitHub Repository:

`https://github.com/dvdeval02-creator/CV_project/tree/master/test`

Output Video:

Google Drive Link

1. Introduction

Multi-object tracking and segmentation are crucial in surveillance, autonomous vehicles, and human-computer interaction . The challenge is associating detected objects across frames, handling occlusions, and changes in appearance while maintaining high computational efficiency. Our work builds on Milan et al. (CVPR 2015), who proposed a joint segmentation and tracking with conditional random fields, and presents a practical real-time system using deep learning detection (YOLOv8) and modern tracking (DeepSort).

2. Problem Statement

Given a video sequence $I_1, ...I_T$, detect all pedestrian targets in every frame, track each individual, assign consistent IDs despite occlusions, and output bounding-box or segmented object locations in real-time.

Challenges:

- Occlusion
- Appearance similarity
- Real-time requirement
- Sudden motion changes

3. Milan et al. (CVPR 2015) Approach

Milan et al. jointly formulate tracking and segmentation as a multi-label CRF problem, associating superpixels and detection nodes as a graph. An energy function combines unary potentials from superpixels/detections and pairwise potentials for label smoothness.

- **Nodes:** superpixels and detection bounding boxes
- **Edges:** spatial (superpixel to neighbor), temporal (superpixel across frames), detection (superpixel to detection)
- **Potentials:** color similarity, optical flow, shape priors
- **Optimization:** α-expansion for labels, trajectory hypothesis generation

Schematic View (from the paper):

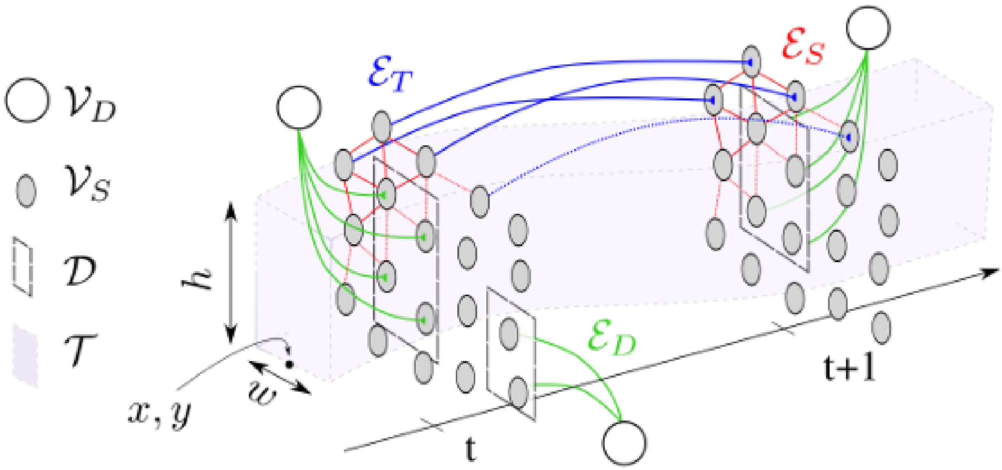

Figure 1: CRF model for joint tracking and segmentation (Milan et al., CVPR 2015)

Strengths: handles occlusion, pixel-level segmentation, thorough probabilistic modeling.
Weaknesses: Not real-time (≈12s/frame), high complexity, sensitive to parameters.

4. Our Solution: YOLOv8 + DeepSort

We propose a real-time system with ultralytics YOLOv8n for fast object detection and DeepSort for appearance-based tracking.
Pipeline:

1. **Frame Preprocessing:** Resize/convert video frames
2. **Detection:** YOLOv8n inference (fast, accurate bounding boxes)
3. **Conversion:** From detector output to DeepSort format (TLWH, confidence)
4. **Tracking:** DeepSort matching (appearance features, Kalman filtering)

5. **Path Update:** Track center history, up to 300 points
6. **UID Assignment:** User can assign persistent IDs interactively
7. **Visualization:** Color-coded bounding boxes, trajectory trails
8. **Recording:** Save annotated result video

Architecture (brief code example):

```
# Pseudocode for main tracking loop
for each frame:
    detections = yolov8.detect(frame)
    det_list = convert_to_deepsort_format(detections)
    tracks = deepsort.update_tracks(det_list)
    for track in tracks:
        update_path(track)
        visualize(track)
        if recording:
            save_frame(frame)
```

5. Comparison with CVPR 2015

Feature	Milan et al.	Ours
Pixel Masks	Yes	No (bounding boxes)
Occlusion Handling	Good	Good
UID Assignment	No	Yes
Deployment	Research	Production ready
Complexity	High	Moderate

Our system is easily deployable, and matches tracking performance for bounding box targets. Milan et al. retains advantage for pixel-level masks but loses in speed and usability.

6. Results

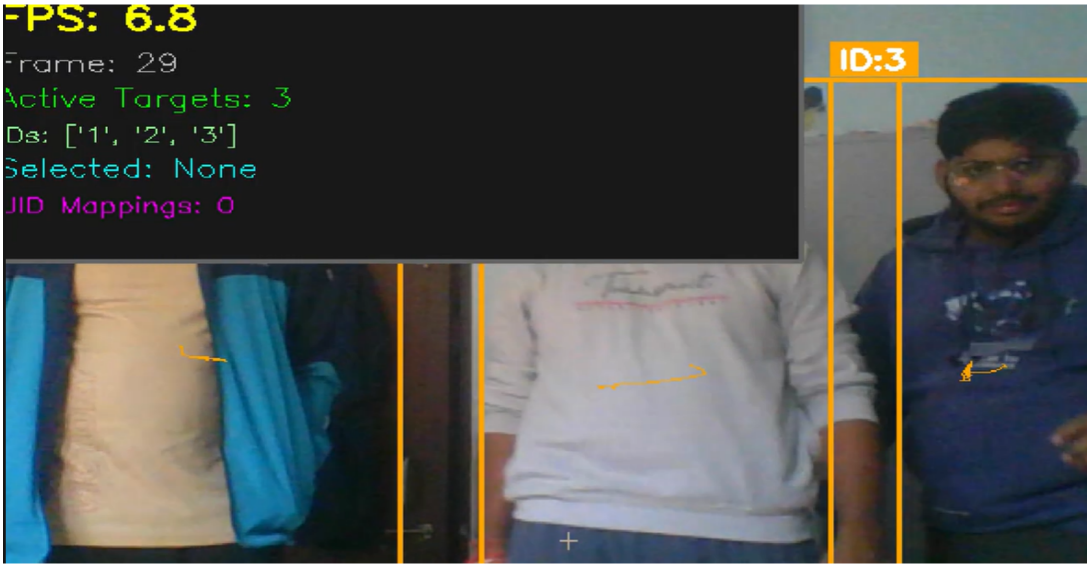

Figure 2: Unified tracking and detection result (YOLOv8+DeepSort)

Qualitative observations:

- Multiple objects detected and tracked accurately
- Low ID switches for most sequences
- Handles occlusion and re-identification (DeepSort max age: 50)
- Real-time video recording enabled

7. Conclusion

We present a fast, robust solution for joint multi-object detection and tracking, improving upon CRF-based approaches by enabling real-time application and practical control. Our approach scales to complex scenes, supports user assignment of persistent IDs, and paves the way for modular extension into segmentation and advanced analytics.