AI Evaluation: Staff Identification and Tracking Solution

Author: Hin Jian Heng Date: 23/9/2025

1. Executive Summary

This document outlines a modular, multi-stage pipeline for identifying and tracking staff in top-down video streams. This solution is designed to address common real-world challenges such as scale variation, occlusion, and appearance changes. A key principle of this work is strict adherence to a zero-data-leakage policy, achieved by fine-tuning a model using a synthetically generated dataset. The final output provides continuous tracks and frame-by-frame coordinates for all identified staff.

The pipeline can be summarised into four main parts:

- **Detection and Tracking**: The YOLOv8L model (enhanced with SAHI for improved small object recall) generates initial object detections. A tracking algorithm then processes these to form initial track segments.
- **Trajectory Clustering**: An innovative clustering algorithm merges track segments based on a multi-factor scoring model to form a single, coherent track for each person.
- **Staff Identification**: A specialised YOLO detector, fine-tuned on a purpose-built synthetic dataset, is used to identify which tracks belong to staff members.
- **Trajectory Completion**: Template matching and interpolation are used to fill any remaining gaps in the final staff tracks.

2. Detailed Solution Architecture

Phase 1: Detection and Tracking

- Problem: The standard YOLO detector suffers from low recall for small, distant, or partially occluded objects.
- Decision: To improve detection recall, we implemented the **SAHI method**. By performing inference on overlapping, high-resolution patches within the frame, this method significantly increases the probability of detecting small objects that might otherwise be lost during downscaling. This approach was chosen over simply lowering the confidence threshold to minimise the introduction of noisy false positives.
- Output: An initial set of fragmented object tracks.

Phase 2: Track Clustering

- Problem: Initial tracking produces fragmented tracks due to long-term occlusions, significant appearance changes, and YOLO detector limitations (such as low recall for the top-down view).
- Decision: A custom clustering algorithm was developed to merge these fragments. The scoring function evaluates track pairs based on three weighted heuristics: **temporal continuity, spatial proximity of transition points, and overall appearance similarity** (using ResNet50 embeddings). This multi-factor approach provides more robust merging decisions than relying on any single metric. However, for complex cases like drastic appearance changes, tracks that remain fragmented may require manual review.
- Output: A coherent set of long-term tracks is generated for each unique individual.

Phase 3: Staff Identification

- Problem: Due to the lack of a pre-existing training set, directly using images from the video frames to train an identifier would result in data leakage and an unreliable performance evaluation.
- Decision: A **synthetic data pipeline** was created. top-view person images were obtained from the **P-DESTRE dataset** to serve as backgrounds. A clear staff name tag was obtained by taking screenshots from the sample examples. Positive samples were generated by **augmenting the tag and pasting it onto images of external individuals.** To force the model to learn the correct features, hard negative samples were also generated by pasting similar but incorrect "distractor" templates. A **YOLOv8 model** was then fine-tuned on this purely synthetic dataset. This approach strikes a balance between the need for a specialised detector and the strict requirement of zero data leakage. The fine-tuned detector was applied to the final tracks, and the staff/non-staff classification was performed through a voting mechanism.
- Output: A list of trajectory IDs corresponding to staff members.

3. Limitations and Future Work

This section outlines the current system's known limitations and potential routes for future improvement.

Limitation 1: Atypical Person Detection

- Observation: Even with SAHI, the pre-trained YOLOv8L model has difficulty detecting people in certain atypical poses. This video contains many challenging conditions, such as a steep top-down perspective and a wide-angle camera view, which result in some individuals failing to be detected in many frames.
- Proposed Mitigation: As demonstrated in Phase 3, targeted fine-tuning can address this "domain gap." By collecting more related training sets and fine-tuning, the detector's robustness to these conditions can be significantly improved.

Limitation 2: Incomplete Track Generation in Some Cases

- Observation: Due to the detection limitations mentioned above, many frames were not correctly detected, resulting in numerous empty frames. Therefore, even though our clustering method already considers spatial, temporal, and appearance factors, it still struggles to merge tracks. We also attempted to address this issue using backtracking and forward tracking. However, the issue can’t be completely solved due to the significant appearance change
- Recommended Mitigation: The system could be augmented with multi-modal tracking features. Incorporating non-appearance-related biometrics, such as gait analysis, can provide a secondary signal to maintain identity continuity despite drastic changes in appearance.

4. Conclusion

The developed pipeline provides a robust and well-reasoned solution for the staff identification and tracking problem. By breaking down the problem into modular stages and making thoughtful and rational technical decisions at each step, the system effectively handles a variety of complex real-world situations. The identified limitations and planned future work provide a clear roadmap for further improvement.