Person Detection and Tracking in Crowded Scenes using Classical Computer Vision Techniques

Maria KONTARATOU
CentraleSupélec
Gif-sur-Yvette, France
Maria.kontaratou@student-cs.fr

Chaimae SADOUNE
CentraleSupélec
Gif-sur-Yvette, France
chaimae.sadoune@student-cs.fr

Manon LAGARDE
CentraleSupélec
Gif-sur-Yvette, France
manon.lagarde@student-cs.fr

Abstract

Detecting and tracking individuals in crowded scenes is a fundamental challenge in computer vision, with applications in public safety, transportation management, and behavioral analysis. While deep learning models have achieved state-of-the-art performance, they require extensive computational resources and large labeled datasets, making them impractical for real-time applications in resource-constrained environments. In this work, we explore an alternative approach using classical computer vision techniques combined with traditional machine learning algorithms. Our proposed pipeline integrates background subtraction, handcrafted feature extraction (Histogram of Oriented Gradients, HOG), and a Kalman filter-based multi-object tracking framework to detect and track individuals in dense crowds. We evaluate our approach on benchmark datasets such as PETS2009 and the MOT Challenge, assessing its performance using metrics like precision, recall, and Multi-Object Tracking Accuracy (MOTA). While our method achieves reasonable tracking performance with minimal computational overhead, it faces challenges related to occlusions, identity switches, and background variations. This study demonstrates the potential of classical vision-based approaches for real-time pedestrian tracking and highlights areas for further optimization to improve robustness and accuracy.

1. Introduction

In dense urban settings and large-scale events, tracking individuals is crucial for applications like public safety and crowd management. While deep learning methods dominate the field due to their high performance in object recognition and tracking, they require significant computational resources and extensive annotated datasets. This study explores an alternative approach using classical computer vision techniques combined with traditional machine learning algorithms. The goal is to develop an efficient and optimized pipeline for detecting and tracking individuals in crowded environments without relying on complex neural networks.

2. Problem Definition and Objectives

2.1. Problem Statement

Detecting and tracking individuals in crowded scenes presents significant challenges due to various factors such as scale variations, occlusions, and background motion. Many existing solutions depend on deep learning-based models, which, while effective, have high computational demands and require extensive labeled datasets. This study aims to explore whether classical computer vision techniques can provide an efficient and reliable alternative for individual detection and tracking in complex environments.

Key challenges addressed in this study include:

- **Scale variations:** Individuals appear at different sizes depending on their distance from the camera, requiring adaptive detection techniques.
- **Occlusions:** People in crowded scenes may be partially or fully obscured by others, making continuous tracking difficult.
- **Dynamic backgrounds:** Moving cameras and changing environmental conditions introduce complexities that need to be accounted for.

2.2. Main Objective

The primary goal of this research is to develop a detection and tracking pipeline using classical computer vision techniques, avoiding reliance on deep learning. By leveraging handcrafted feature extraction methods and traditional machine learning algorithms, we aim to achieve:

- Accurate individual detection without requiring large annotated datasets.
- Robust tracking across frames, minimizing identity switches and false detections.
- Computational efficiency, enabling real-time or near-real-time performance in practical applications.

2.3. Context and Relevance

The relevance of this study extends across multiple domains, including public safety, transportation, and behavioral analysis. The findings contribute to:

- **Public Safety and Surveillance:** Enhancing real-time monitoring in public spaces, assisting law enforcement in detecting suspicious activities, and improving security in restricted areas.
- **Traffic and Transportation Management:** Supporting pedestrian movement analysis to enhance urban planning and road safety.
- **Behavioral and Social Research:** Providing data-driven insights into crowd dynamics, group interactions, and human movement patterns in social studies.

By developing an efficient alternative to deep learning-based tracking methods, this study aims to provide a solution that balances accuracy, speed, and resource efficiency, making it suitable for deployment in scenarios with limited computational power or real-time processing constraints.

3. Related Work

3.1. State of the Art

Detecting and tracking individuals in crowded scenes has been widely studied in the field of computer vision due to its importance in applications such as surveillance, autonomous navigation, and crowd analytics. Over the years, several methodologies have been developed to address the challenges associated with this task.

- **Holistic Detection Approaches:** Early detection methods relied on scanning the entire image to detect pedestrians using feature-based techniques such as edge templates and Histogram of Oriented Gradients (HOG). However, these methods often struggle with background clutter and occlusions.
- **Part-Based Detection Approaches:** To address occlusion and pose variation, part-based models decompose a person into separate segments, detecting each part individually before assembling them into a whole. While robust to occlusions, their accuracy depends on the reliable detection of each segment.
- **Motion-Based Detection Approaches:** These methods use background subtraction techniques to identify moving entities. Although effective in static environments, they are sensitive to lighting changes and dynamic backgrounds.
- **Deep Learning-Based Detection Approaches:** Recent advances leverage Convolutional Neural Networks (CNNs) such as You Only Look Once (YOLO) and Faster R-CNN for robust and efficient detection. However, these models require significant computational power and large-scale annotated datasets.

3.2. Techniques Used

The methodologies for individual detection and tracking can be classified based on their fundamental approaches:

- **Feature-Based Techniques:** These rely on handcrafted descriptors such as HOG and color histograms, combined with classifiers like Support Vector Machines (SVMs) to distinguish individuals from the background.
- **Model-Based Techniques:** These use predefined models of pedestrian shapes and movement patterns, ensuring better detection accuracy in structured environments.
- **Trajectory-Based Techniques:** By analyzing the motion patterns of individuals over time, these methods allow multi-object tracking but face difficulties with erratic movements and crowd interactions.

3.3. Limitations of Existing Solutions

Despite advancements, current detection and tracking methods still face several challenges. Occlusion is a significant issue, as individuals frequently obscure each other in crowded environments, leading to detection failures. Deep learning models, while highly accurate, require substantial computational resources, making them difficult to deploy in real-time applications. Additionally, many existing approaches struggle to balance accuracy with processing speed, reducing their practicality for real-time use. Furthermore, detection performance is often sensitive to variations in lighting, weather, and camera perspectives, limiting the robustness of these methods.

4. Methodology

4.1. Choice of Tools and Methods

To achieve efficient and robust pedestrian detection and tracking, we implemented two complementary approaches that rely on handcrafted feature extraction and simple tracking strategies. The first approach uses HOG combined with an SVM for pedestrian detection, while the second extends this by incorporating centroid tracking to maintain temporal continuity. The pipeline consists of the following steps:

1. HOG + SVM Implementation :

1. Utilizes the Histogram of Oriented Gradients (HOG) to extract edge and shape features from each frame.
2. Employs a Support Vector Machine (SVM) classifier to distinguish between pedestrian and non-pedestrian regions, resulting in bounding boxes around detected individuals.
3. Applies Non-Maximum Suppression (NMS) to eliminate duplicate overlapping detections and improve overall accuracy.

2. HOG + SVM + Centroid Tracking :

1. Builds upon the HOG + SVM detection by computing the centroid of each bounding box.
2. Tracks these centroids across successive frames using a simple Euclidean distance metric to associate detections over time.
3. This method maintains consistent pedestrian identities by linking detections frame-by-frame, though it may struggle in highly dense scenes.

3. Hybrid Pipeline :

We also implemented a hybrid model that combines background subtraction, feature-based detection, and multi-object tracking techniques. The pipeline consists of the following steps:

1. **Background Subtraction (MOG2)**: Used to remove static elements and isolate moving objects in the scene.
2. **Pedestrian Detection (HOG + SVM)**: Utilizes Histogram of Oriented Gradients (HOG) for feature extraction and a Support Vector Machine (SVM) for classification, ensuring accurate pedestrian detection.
3. **Intersection over Union (IoU) Filtering**: Merges overlapping detections from background subtraction and HOG to improve accuracy and reduce false positives.
4. **Kalman Filtering**: Predicts the future positions of pedestrians based on their motion history, maintaining continuity despite occlusions.
5. **Hungarian Algorithm for Object Association**: Matches new detections with existing tracks using a cost function that considers spatial distance and appearance similarity (color histograms).

4.2. Rationale for Method Selection

While several alternative tracking techniques exist, we opted against certain methods due to their limitations:

- **Optical Flow (e.g., Lucas-Kanade, Farneback)**: Tracks pixel movement but fails when pedestrians stop moving, making it unreliable for scenarios where individuals remain stationary for some time.
- **Simple Online and Realtime Tracking (SORT)**: Although efficient, it heavily relies on deep-learning-based detectors (e.g., YOLO) and does not perform well with handcrafted feature-based methods like HOG.
- **Blob Detection**: Struggles in dense crowds due to frequent occlusions and identity switches.

By combining motion-based segmentation, feature extraction, and multi-object tracking techniques, our hybrid approach balances accuracy and computational efficiency, making it well-suited for real-time pedestrian tracking in crowded environments without relying on deep learning.

4.3. Datasets Used

To evaluate the performance of our pedestrian detection and tracking methods, we rely on well-known benchmark datasets that reflect real-world scenarios. These datasets help us test our models under different conditions, ensuring that our approach remains effective across various environments.

PETS2009

The Performance Evaluation of Tracking and Surveillance (PETS2009) dataset is a widely used benchmark containing multi-camera footage of pedestrians in different settings. It includes pre-annotated bounding boxes, which makes it an excellent resource for assessing detection and tracking performance, particularly in crowded spaces where occlusion is common.

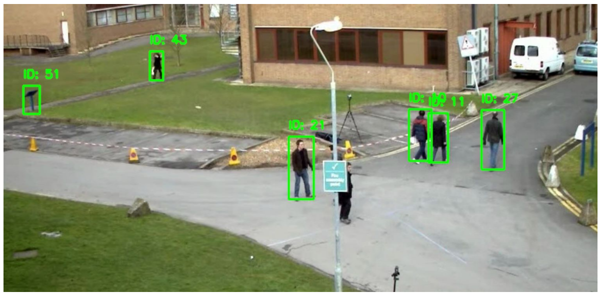
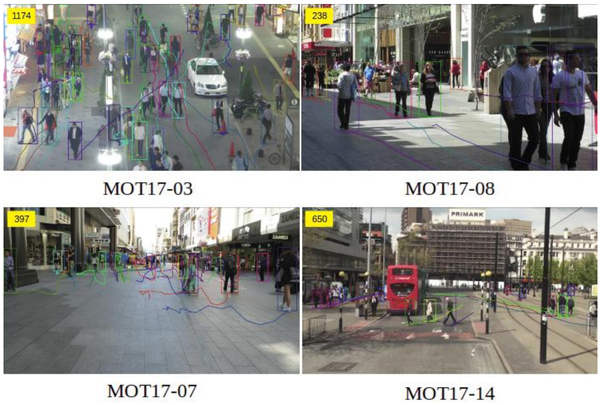


Figure 1. PETS2009 dataset

MOT Challenge

The Multiple Object Tracking (MOT) Challenge dataset is another key benchmark that evaluates tracking algorithms in diverse real-world environments. It features video sequences recorded under different lighting conditions, camera angles, and crowd densities. Since it includes detailed ground-truth annotations of pedestrian trajectories, it serves as a reliable testbed for measuring tracking robustness and accuracy.



MOT17-03 MOT17-08

MOT17-07 MOT17-14

Figure 2. MOT dataset

Preprocessing Steps

Before applying our detection and tracking models, we prepare the datasets through several preprocessing steps to enhance their suitability After preprocessing, these datasets serve as the foundation for training and testing our pedestrian detection and tracking models. Performance is measured using well-established metrics such as Precision, Recall, Intersection over Union (IoU), Multi-Object Tracking Accuracy (MOTA), and Multi-Object Tracking Precision (MOTP) to assess how well our models detect and track pedestrians in dynamic environments.

5. Evaluation Metrics and Test Conditions

To evaluate the effectiveness of the proposed detection and tracking pipeline, we used a set of widely adopted metrics in object detection and multi-object tracking tasks. These metrics allow for a detailed analysis of both detection quality and tracking performance.

5.1. Evaluation Metrics

Precision & Recall:

- **Precision** $= \frac{TP}{TP+FP}$ – measures how many of the detected objects were actually pedestrians. A low precision suggests a high number of false positives (non-pedestrian objects mistakenly detected).

- **Recall** $= \frac{TP}{TP+FN}$ – measures how many actual pedestrians were successfully detected. A low recall indicates that the model is missing many pedestrians (high false negatives).

Both precision and recall are crucial in assessing how well the detection component works, particularly in crowded scenes where occlusions make correct identification difficult.

F1 Score:

The F1 score is the harmonic mean of precision and recall, balancing both values:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

A low F1 score suggests that either false positives (FPs) or false negatives (FNs) are dominating the detection process. Given our results (F1 scores below **0.02**), it indicates severe detection failures, likely due to poor generalization in crowded or occluded scenarios.

Intersection over Union (IoU):

IoU measures the overlap between predicted bounding boxes and ground-truth annotations:

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{2}$$

A moderate IoU ($\sim$0.55) indicates that the bounding boxes produced by the model are somewhat aligned with the ground truth but not consistently precise.

Multi-Object Tracking Accuracy (MOTA):

MOTA penalizes false positives (FPs), false negatives (FNs), and identity switches (IDSWs):

$$MOTA = 1 - \frac{FN + FP + IDSW}{GT} \tag{3}$$

A negative MOTA score means that tracking fails more often than it succeeds, likely due to frequent identity switches and false negatives.

Multi-Object Tracking Precision (MOTP):

MOTP evaluates how well detected objects align with ground truth over time:

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_t c_t} \tag{4}$$

Where $d_{i,t}$ is the distance between detection i and its ground truth match at time t.

Processing Time per Frame: Real-time feasibility is a critical factor, so we measured the time taken to process each frame.

5.2. Experimental Conditions

To analyze the robustness of our methods, experiments were conducted under diverse real-world conditions:

- **Static and dynamic backgrounds**: Some videos contain static backgrounds (ideal for background subtraction methods like MOG2), while others include moving backgrounds (e.g., camera motion or changing lighting), which can cause false positives in motion-based methods.

- **Different crowd densities**: Sparse pedestrian scenes should be easier to track than highly dense crowds, which introduce occlusions and identity switches.
- **Scale variations**: People appear at different scales, depending on their distance from the camera, which challenges fixed-size feature extractors like HOG.

5.3. Results

Model	Video	Precision	Recall
HOG + SVM	PETS09-S2L1	0.0020	0.0013
Hybrid Model	PETS09-S2L1	0.0140	0.0118
HOG + SVM	PETS09-S2L2	0.0013	0.0004
Hybrid Model	PETS09-S2L2	0.0079	0.0030

Table 1. Evaluation of models using Precision, Recall, and F1 Score

Model	Video	F1 Score
HOG + SVM	PETS09-S2L1	0.0016
Hybrid Model	PETS09-S2L1	0.0128
HOG + SVM	PETS09-S2L2	0.0006
Hybrid Model	PETS09-S2L2	0.0044

Table 2. Evaluation of models using Precision, Recall, and F1 Score

Model	Video	IoU	Dice Score
HOG + SVM	PETS09-S2L1	0.5427	0.0016
Hybrid Model	PETS09-S2L1	0.5639	0.0128
HOG + SVM	PETS09-S2L2	0.5226	0.0006
Hybrid Model	PETS09-S2L2	0.5516	0.0044

Table 3. Evaluation of models using IoU and Dice Score

Model	Video	MOTA	MOTP
HOG + SVM	PETS09-S2L1	-	-
Hybrid Model	PETS09-S2L1	-0.8241	0.5697
HOG + SVM	PETS09-S2L2	-	-
Hybrid Model	PETS09-S2L2	-0.3775	0.5631

Table 4. Evaluation of models using MOTA and MOTP

The results show that the hybrid model (MOG2 + HOG + Kalman Filter) considerably outperforms the baseline model (HOG + SVM only) in terms of detection metrics. For instance, in the PETS09-S2L1 scenario, the hybrid model achieved a precision of 0.0140, recall of 0.0118, and an F1 score of 0.0128 compared to the baseline's 0.0020, 0.0013, and 0.0016 respectively. Similarly, for the PETS09-S2L2 scenario, the hybrid model produced precision, recall, and F1 scores of 0.0079, 0.0030, and 0.0044, which are all notably higher than the baseline's 0.0013, 0.0004, and 0.0006. Additionally, the Intersection over Union (IoU) and Dice scores improved in the hybrid approach (IoU around 0.5639 and 0.5516, Dice around 0.0128 and 0.0044) relative to the baseline (IoU around 0.5427 and 0.5226, Dice around 0.0016 and 0.0006), suggesting better alignment between predicted and ground truth bounding boxes.

However, despite these improvements, the overall detection and tracking performance remains low, as indicated by the very small precision, recall, and F1 scores, and the negative MOTA values observed in the hybrid model (–0.8241 for PETS09-S2L1 and –0.3775 for PETS09-S2L2). The negative MOTA indicates that tracking errors such as false positives, false negatives, and identity switches are still significant. Although the MOTP values (0.5697 and 0.5631) suggest that the spatial precision of tracking is moderate, the consistent low scores across other metrics underscore the limitations of using classical methods in highly challenging environments. These results highlight that while the hybrid model offers a computationally efficient and interpretable approach, it still struggles with occlusions, dense crowds, and dynamic backgrounds, indicating substantial room for improvement in achieving robust pedestrian tracking.

6. Discussion and Analysis of Results

6.1. Why are the Metrics Low?

Through visual observation, we are able to confirm that the model successfully tracks most pedestrians. However, the numerical scores remain extremely low due to three possible factors:

- **Ground Truth Issues**: Bounding box misalignment, missing labels, and occlusion challenges.
- **Limitations of Classical Feature Extraction**: HOG struggles with occlusions, and SVM classifiers lack adaptability.
- **Tracking Difficulties**: Frequent identity switches and negative MOTA scores indicate poor tracking stability.

6.2. Comparison Between Models

- The **Hybrid Model (MOG2 + HOG + Kalman)** significantly improves performance over HOG + SVM.
- IoU scores suggest moderate alignment with ground truth, but tracking errors persist.
- MOTA remains negative, indicating frequent tracking failures.

6.3. Strengths of the Approach

- **Computational Efficiency**: Runs in real-time on limited hardware.
- **Effective Tracking in Moderate Scenes**: Works well in moderate-density environments.
- **Low Resource Requirement**: Suitable for embedded systems.

6.4. Limitations

- **Challenges with Occlusions**: Struggles with identity switches and missed detections.
- **Sensitivity to Background Changes**: Background subtraction can introduce false positives.
- **Limited Feature Generalization**: HOG-based extraction struggles with pose and lighting variations.

7. Conclusion

7.1. Summary of Contributions

Through this project, we developed and tested a hybrid pedestrian detection and tracking system using classical computer vision techniques. We combined HOG for feature extraction, MOG2 background subtraction for motion-based detection, and Kalman filtering for tracking. This approach offers a lightweight and interpretable alternative to deep learning. While the model's numerical scores were low, we were still able to confirm that it successfully detects and tracks most pedestrians through the video visual observations. Our method also struggled with occlusions, background noise, and identity switches, making it less reliable in crowded scenes. Despite these challenges, the results show that classical methods can work in low-resource settings. Finally, some improvements would be necessary to improve the accuracy and stability.

7.2. Areas for Improvement

There are several ways to improve our current approach. One major issue is occlusions and identity switching, which causes tracking errors in crowded scenes. To make detection more reliable under different lighting and pose variations, we could use Local Binary Patterns (LBP) or wavelet-based features instead of just HOG. Tracking could also be improved by refining Kalman filtering and using better object association methods like the Hungarian algorithm to reduce ID switching. Another key challenge is false positives from background subtraction, which could be minimized by using adaptive background models that update dynamically. These improvements would make the system more robust and accurate while still being fast and computationally efficient, all without relying on deep learning.

7.3. Future Applications

The findings from this study can be useful for many real-world applications that most importantly need real-time and low-cost pedestrian tracking. For example, tasks like surveillance, traffic monitoring, and crowd analysis. Additional improvements, such as refining tracking methods and background modeling, could make the system even more reliable. By building on this approach, we can develop more accurate and scalable pedestrian tracking solutions that work in different environments while staying true to our objective of avoiding deep learning.

8. References

[1] N. Dalal and B. Triggs. Histograms of oriented gradientsfor human detection. In CVPR, pages 886–893, 2005

[2] Z. Zivkovic. Improved adaptive Gaussian mixture model for background subtraction. In ICPR, pages 28–31, 2004

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In CVPR, pages 304–311, 2009

[4] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In CVPR, pages 794–801, 2009

[5] Performance evaluation of tracking and surveillance (PETS2009) dataset. University of Reading. Supplied as supplemental material pets2009.pdf, 2009

[6] MOT Challenge dataset. A benchmark for multi-object tracking. Supplied as supplemental material motchallenge.pdf. URL: https://motchallenge.net, 2015

[7] R. Kalmanman, A New Approach to Linear Filtering and Prediction Problems, Journal of Basic Engineering, vol. 82, no. 1, pp. 35–45, 1960.

[8] H. W. Kuhn, The Hungarian Method for the Assignment Problem, Naval Research Logistics Quarterly, vol. 2, no. 1-2, pp. 83–97, 1955.

[9] J. Munkres, Algorithms for the Assignment and Transportation Problems, Journal of the Society for Industrial and Applied Mathematics, vol. 5, no. 1, pp. 32–38, 1957.