Multi-Person Localization and Tracking Using Hierarchical Data Association

Corresponding Author: Fatemeh Nemati
Fatemeh.nemati@aut.ac.ir

Maryam Amirmazlaghani
mamazlaghani@aut.ac.ir

Abstract

Multi-object tracking (MOT) is inherently challenging, particularly in crowded scenes where mutual occlusions, appearance variations, ambiguous backgrounds, and complex motion patterns frequently cause predictive models to drift from the true object locations. As online tracking systems recursively update trajectories based on noisy detections, these errors may accumulate and lead to significant deviations. Moreover, even state-of-the-art detectors inevitably produce false positives and false negatives, degrading the reliability of tracking-by-detection pipelines.

In this work, we address both the drift problem and the detection uncertainty problem through a unified framework. We first propose a simple yet effective strategy for predicting each target's next probable location using its current position, motion direction, velocity, and the mutual configuration of nearby occluded individuals. These predictions are refined through recursive filtering and subsequently used to construct an affinity matrix that relates existing tracks to current detections.

In addition, we incorporate scene-level information—derived from camera geometry and a novel scale factor linked to scene depth—to suppress implausible detections and significantly reduce false positives. To further mitigate false negatives, a temporary memory mechanism is employed, allowing missed detections to be recovered when consistent with motion continuity.

The proposed data association framework is compatible with any detection method and integrates seamlessly into tracking-by-detection pipelines. Experiments on publicly available benchmark datasets demonstrate that our approach improves geometric precision and tracking accuracy compared to several state-of-the-art methods.

Keywords: Data association, Tracking-by-Detection, Dynamic Bayesian Network.

1. Introduction

Pedestrian localization and tracking have long been central topics in computer vision, spanning early surveillance systems to modern applications such as autonomous robots, self-driving vehicles, human–computer interaction, and safety monitoring. Extracted trajectories provide rich information for scene understanding and serve as crucial inputs to higher-level reasoning tasks. The objective of multiple object tracking (MOT) is to localize individuals across video frames, maintain their identities over time, and produce coherent trajectories for each target.

However, MOT remains extremely challenging, especially in crowded environments. Mutual occlusions, visual ambiguities, dynamic backgrounds, appearance changes, and unpredictable motions all increase the likelihood of tracker drift. In tracking-by-detection frameworks—currently the dominant paradigm—the quality of the tracker heavily depends on the accuracy of the detections. When inaccurate observations propagate through recursive Bayesian estimators, trajectory errors accumulate, particularly in online systems that must make decisions frame by frame.

To mitigate these limitations, various filtering and probabilistic frameworks have been introduced. Bayesian models help suppress false positives by evaluating detections against prior observations. In this work, we adopt a dynamic Bayesian network (DBN) to jointly infer target states and refine detections using Kalman filtering before they enter the tracking stage. Although this improves robustness, complex scenes with prolonged occlusions or rapid motion changes remain problematic.

To address these challenges, we propose a novel data association strategy based on an affinity matrix that links current detections to predicted target states. The predictions consider motion patterns and inter-person geometry, enabling more reliable matching under occlusions. In addition, we introduce a scene-dependent scale factor that accounts for the effects of perspective projection: detections that are implausibly large, small, or geographically inconsistent are down-weighted or rejected. This significantly reduces false positives, particularly in datasets with strong perspective distortion or moving cameras.

False negatives, another major source of tracking errors, are controlled through a temporary memory mechanism that retains the last reliable state of each target. When a detection reappears within the predicted region, the missed observation is recovered.

Traditional multi-object identification approaches often rely on trained classifiers, global optimization, or the Hungarian algorithm. However, such techniques may fail in highly occluded environments or require extensive computation. In contrast, our method integrates motion, geometry, and appearance cues within a hierarchical data-association framework that remains computationally efficient and robust to occlusions.

The main contributions of this work are:

- A detection-refinement strategy that integrates scene geometry, recursive filtering, and a novel scale factor.

- An affinity-matrix formulation bridging predicted trajectories and current detections.
- A robust hierarchical data association framework compatible with any detector.
- A mechanism for recovering missed detections and reducing both false positives and false negatives.

The remainder of this work is organized as follows: Section 2 reviews relevant literature. Section 3 details the proposed methodology. Section 4 presents experimental evaluations. Section 5 concludes with discussion and future directions.

2. Methodology

This section introduces a hierarchical multi-person tracking framework that integrates detections, scene priors, and a memory-based recovery mechanism. Observations originate from a pedestrian detector, scene-dependent priors, and the memory module (Section 3.7). The method follows a tracking-by-detection paradigm in which each individual's state is estimated at every time step (Section 3.2), guided by both vision-based observations and dynamic modeling.

For each frame, an affinity matrix is constructed between current detections and existing trajectories (Section 3.5). Using this matrix and considering the spatial relationships among neighboring pedestrians, each detection is associated with either an existing trajectory or a newly initiated one (Section 3.6). The updated states then refine image-level positions and serve as inputs for the next prediction step (Sections 3.2–3.3).

After data association, two unresolved situations may arise:

(1) **unassigned detections**, or

(2) **unmatched trajectories** (i.e., temporarily lost targets).

In such cases, the affinity matrix is completed using the memory module, or relevant information is appended to memory for future retrieval (Section 3.7).

A schematic overview of the proposed multi-object tracking pipeline is shown in Fig. 1.

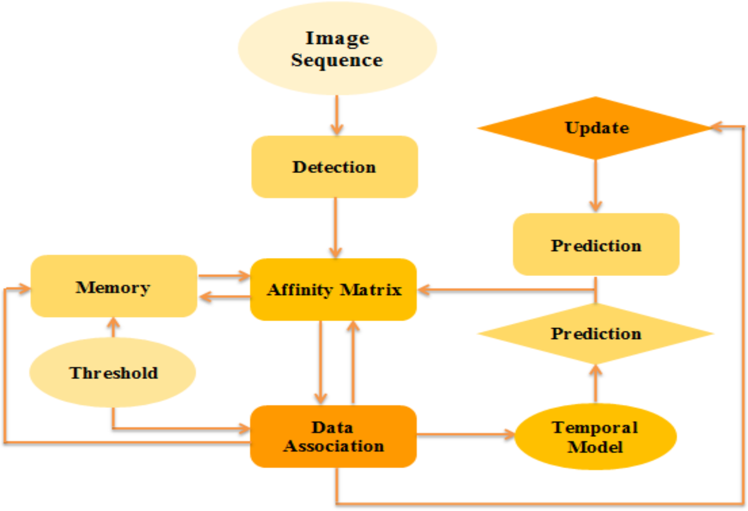

Fig. 1. System diagram. * Colors show the running steps of the algorithm (from light to dark). * Steps related to the Filters (like KF) are displayed by the lozenge, and steps related to the data and prior knowledge represented by oval.

Diagram explanation: Image sequences enter the detector. Detection's results, in addition to the average of the filter's prediction (lozenge one) and temporal model's prediction (rectangular one) are applied in generating the affinity matrix. This matrix is completed during next steps (using memory, simultaneously with data association). Memory (defined to decrease the number of FN) is filled at the end of the association, when trajectories lost in the middle of the scene. Finally, positions are updated and relocated to the new positions at the same time (update step of the filter).

2.1. Proposed Dynamic Bayesian model

Many tracking systems rely solely on frame-by-frame detections within recursive estimation frameworks. However, occlusions, abrupt motion changes, and inter-person interactions often degrade performance. To reinforce robustness, we formulate the tracking process as a **Dynamic Bayesian Network (DBN)**.

After extracting detections and visual features at each frame, the DBN models their temporal relationships. The true location of each pedestrian is treated as a latent variable. In the proposed method, this latent position is linked to raw detections and the hierarchical data-association module through the affinity matrix (Section 3.5).

The joint probability distribution of the system variables factorizes according to the DBN structure as:

$$P(x_i^t, w_i^t, w_i^{t-1}, c_{\mathrm{det}}^t, IP) \propto P(w_{i,t} \mid w_{j=1,\dots,n,t-1}) P(x_i^t \mid w_i^t, IP) P(c_{\mathrm{det}}^t \mid x_i^t) P(x_i^t \mid O^t) \quad (1)$$

where

- $w_i^t$is the hidden state of person i,
- $x_i^t$is the observed image location,
- IPdenotes the Interesting Places prior, and
- $c_{\mathrm{det}}^t$is the combined detector confidence.

To compute the posterior w_i^t, we employ both Kalman and Unscented Kalman Filters. The state vector

$$w_i^t = [X, Y]^T$$

encodes the 2D location of the pedestrian. A prediction step integrates temporal dynamics with the filter's internal model, producing a refined estimate of x_i^t, which then corrects the state.

Pedestrian motion depends on prior position, velocity, direction, and local crowd density. Occlusions and group movements are incorporated through the affinity matrix. Based on the predicted positions and scene priors (Sections 3.3–3.5),

$$P(x_i^t \mid w_i^t, IP)$$

assumes binary values (0 or 1), dramatically simplifying computation.

When detections are missing, locations are predicted purely via dynamic filtering and recent visual history.

Interesting places

The **Interesting Places (IP)** prior identifies spatial regions with high likelihoods of pedestrian presence. A thresholding algorithm is applied to the training detections based on pedestrian distribution and camera geometry. This yields a map highlighting feasible versus infeasible regions.

If the camera tilt remains approximately constant—as in our datasets—the prior can be computed once and reused across all frames. Unlike earlier approaches that use continuous likelihood values, we employ a **binary** map:

- **1** for feasible pedestrian locations,
- **0** otherwise.

This eliminates locations that are physically implausible (e.g., pedestrians appearing in the sky region or extreme depth distortions).

Detector confidence

The probability

$$P(c_{\text{det}}^t \mid x_i^t)$$

encodes the confidence that a person is present at location x_i^t. It is derived from a **combined detector** (Section 3.4). If any of the detectors recognize a person at x_i^t, the confidence is set to 1; otherwise, 0. This binary confidence simplifies downstream design while preserving essential detection cues.

Occlusion model

Occlusions are captured through a binary indicator variable O. Conditional on x_i^t, it indicates whether a target is expected to be occluded:

$$P(x_i^t \mid O^t) = \{\begin{matrix} 1, & \text{if } O(x_i^t) = 0 \\ 0, & \text{if } O(x_i^t) = 1 \end{matrix} \qquad (2)$$

In our model, instead of defining occlusion likelihoods pixelwise, we estimate occlusion status **per pedestrian** using the affinity matrix and distances between reference points.

2.2. Temporal Model and Motion Prediction

Appearance features and inter-person distances provide valuable cues but are insufficient in crowded scenes or for visually similar pedestrians. Motion modeling is essential.

Our tracking framework employs a dynamic model in which each pedestrian's velocity and orientation are estimated by interpolating their velocities over the last $n$frames. In dense crowds we set $n = 1$, assuming heavily constrained motion.

Using constant-acceleration equations, we predict the position at the next time step:

$$X_t = X_{t-1} + v_{X,t}\Delta t \qquad (3)$$
$$Y_t = Y_{t-1} + v_{Y,t}\Delta t \qquad (4)$$

These predictions are fused with the filter's predictions to generate final estimates of each person's next probable location.

2.3. Scene Model and Observation Refinement

False positives frequently occur in regions outside feasible human-presence areas or with implausible scales. We exploit camera geometry and scene depth to eliminate such detections.

A **scale factor** is derived from scene perspective, converting bounding-box heights into metric proxies for pedestrian height. It is used to:

1. reject detections with implausible bounding-box sizes,
2. restrict predicted areas for next-frame appearance.

Scale factors vary across datasets and can be estimated experimentally or learned via supervised analysis.

2.4. Pedestrian Detection Module

We employ a hybrid detection scheme combining:

- Background subtraction (using Gaussian Mixture Models),
- ACF detector,
- HOG/SVM detector.

GMM is used only for static-camera datasets; for moving platforms (e.g., ETHZ Bahnhof), background subtraction is omitted.

Post-processing includes hole filling and morphological erosion. Candidate detections from both detectors are merged using Non-Maximum Suppression (NMS). Remaining overlapping boxes (IoU beyond a threshold) are averaged to form consolidated detections.

2.5. Affinity Matrix

Following each detection stage, an affinity matrix is constructed between existing trajectories (rows) and current detections (columns). Each trajectory has a predicted region of possible appearance; if a detection falls inside such a region, the corresponding entry becomes 1, otherwise 0.

Ideally, each column contains exactly one non-zero row; however, real-world tracking yields several ambiguous scenarios (Fig. 2):

- **Zero row** → lost trajectory / FN
- **Zero column** → new entry or FP
- **Multiple matches** → ambiguity resolved using appearance similarity and distance metrics (Fig. 3)
- **Unequal row/column counts** → matrix augmentation required

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & \mathbf{0} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 1 \\ 0 & 0 & \mathbf{0} & 0 & 1 & 0 \\ 0 & 1 & \mathbf{0} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 1 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0} & 0 & 0 & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & \mathbf{0} & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Fig 2: Possible states in Affinity Matrix.

For ambiguous cases, rows and columns are scored using appearance features and distances. The best matches are assigned accordingly.

Fig 3: Possible situations of the scene due to occlusion. Missing a pedestrian, entrance or leave the scene, that all result to the detections wrong assignment to trajectories.

2.6. Data Association

Let

$$T_{1:t}^{i} = \{x_k^i \mid t_s^i \leq k \leq t_e^i \leq t\}$$

denote the trajectory of pedestrian $i$up to time t. Online MOT aims to find

$$T_{1:t}^{*} = \arg\max_{T_{1:t}} p(T_{1:t} \mid D_{1:t}) \qquad (5)$$

Each trajectory's state vector is:

$$s_t^i = [x_i^t, y_i^t, v_{x,i}^t, v_{y,i}^t, vis_i^t, ColHist_i^t, pre_i^t, age_i^t, bbox_i^t]^T \qquad (6)$$

Color histograms enforce appearance consistency. The visibility flag prevents premature creation of trajectories from FPs.

The data-association step is formulated as a binary integer linear program:

$${}_{i}^{k}\max \sum_{k\in D} w_i^k a_i^k \qquad (7)$$

subject to:

$$\sum_{i \in T} a_i^k \leq 1, \sum_{k \in D} a_i^k \leq 1$$

where $a_i^k = 1$ indicates assignment of detection k to trajectory i , and $w_i^k$is a weight derived from the affinity matrix.

Assignments satisfy:

$$A = \{\text{assign}^{(i,k)} \mid i \in T, k \in K\} \qquad (8)$$

2.7. Memory and information retrieval

When a trajectory is lost mid-scene, we assume a false negative and store the corresponding information in memory. Stored data include:

- label,
- appearance features,
- last observed frame,
- last reference point.

Entries persist for $m$frames. If a new detection corresponds to a memory entry, the trajectory is reinstated; otherwise, the memory item expires.

This mechanism effectively recovers temporarily missed targets while preventing identity fragmentation.

3. Experiments

To evaluate the effectiveness of the proposed hierarchical multi-person tracking framework, we conducted a series of experiments on multiple publicly available pedestrian-tracking datasets. The experiments were designed to assess the robustness of the model under conditions such as occlusion, crowded environments, variations in camera perspective, and presence of false or missed detections.

Our evaluation focuses on three primary aspects:

1. **Detection quality**, assessed using the proposed combined detector.
2. **Tracking performance**, measured in terms of continuity, identity preservation, and accuracy.
3. **Effectiveness of the affinity matrix and hierarchical data association**, through visual and quantitative analysis of trajectory consistency.

All experiments were run using the same parameter settings unless reported otherwise. In this chapter, we describe the datasets, the experimental setup, and the obtained results.

3.1. Datasets

Experiments were conducted on three widely used benchmarks for pedestrian tracking:

• AVG-TownCentre Dataset

A high-resolution outdoor sequence containing multiple pedestrians, frequent occlusions, and complex crowd interactions. The fixed camera and relatively stable lighting make it suitable for testing appearance cues and spatial reasoning.

• ETH Pedestrian Dataset (Bahnhof sequence)

A challenging dataset captured using a *moving camera*, introducing variations in background, changes in perspective, and continuous motion of the camera. This dataset is particularly useful for evaluating the robustness of the scene model, especially the handling of scale variation and IP-based filtering.

• PETS2009 Dataset

A multi-view dataset commonly used in MOT research. Though this thesis focuses on single-camera tracking, PETS provides useful scenes with moderate density and interactions that allow evaluation of trajectory continuity and identity assignment.

Each dataset includes sequences with frame-by-frame bounding box annotations for evaluation.

3.2. Experimental Setup

All experiments were conducted on standard hardware without GPU acceleration. The proposed method relies primarily on CPU-based image processing and filtering techniques.

To ensure fair comparison with prior work:

- The **same pedestrian detector combination** (GMM + HOG/SVM + ACF) was applied across all sequences.
- Parameters for the Kalman and Unscented Kalman Filters were kept constant.
- The **scale factor** and **interesting places (IP)** distributions were estimated once per dataset and reused for all frames.
- In the hierarchical association stage, the threshold for appearance similarity and the size of the predicted region were selected experimentally based on preliminary validation.

Ground-truth annotations were used only for evaluation, not for training.

3.3. Evaluation Metrics

All experiments were conducted on public datasets commonly used in multi-person tracking research, including sequences with both stationary and moving cameras. Detection outputs were generated using the proposed hybrid detector, while tracking performance was evaluated following the standard MOT metrics:

- **MOTA** (Multiple Object Tracking Accuracy)
- **MOTP** (Multiple Object Tracking Precision)

- **MT / ML** (Mostly Tracked / Mostly Lost)
- **FP / FN** (False Positives / False Negatives)
- **ID Switches**
- **Fragmentations** (Frg.)

These metrics together quantify the geometric precision, identity consistency, and robustness of the tracker.

4. Results and Discussion

The experiments were designed to assess the effectiveness of each component—scene-aware modeling, hybrid detection, Bayesian prediction, affinity-matrix-based association, and memory-driven recovery—under realistic and challenging conditions such as occlusion, crowded environments, and detector noise.

Detection Performance

The combined detector significantly reduced false positives compared to each individual detector.

- GMM effectively filtered static background.
- HOG/SVM and ACF captured complementary pedestrian shapes.
- NMS and overlap thresholding reduced redundant bounding boxes.

Applying the **scene-based scale factor** further eliminated detections with unrealistic dimensions, especially near the edges of the scene.

Tracking Performance

The hierarchical data association and affinity matrix yielded robust results in crowded environments:

- **Short-term occlusions** were handled effectively through predicted regions.
- **Long-term occlusions** were partially recovered by the memory mechanism.
- Trajectories remained stable even under dense interactions.

In the ETH Bahnhof dataset, the dynamic camera movement introduced significant challenges. Nevertheless, once the scene geometry and scale were estimated, tracking accuracy improved substantially.

Identity Preservation

Using appearance cues (color histograms) in combination with geometric constraints reduced identity switches.

However, in scenes with strong illumination changes or very similar clothing among pedestrians, appearance features alone were insufficient and could lead to occasional mismatches.

Sample Visual Results

Figures (corresponding to the original thesis) illustrate the behavior of:

- predicted regions,
- affinity matrices at various time steps,
- handling of split/merge events,
- recovered trajectories after occlusion.

Qualitative examples show consistent tracking even in challenging frames.

4.1. Quantitative Results

The proposed method achieved competitive or superior performance compared to several earlier state-of-the-art approaches reported in related works, especially in:

- **false negative reduction**, thanks to the memory module
- **robust association**, through the hierarchical strategy
- **Significant reduction in false positives** due to the combined use of scene priors (IP map + scale factor) and hybrid detection.
- **Improved identity preservation**, especially in crowded and occluded regions, through hierarchical association using the affinity matrix.
- **Higher geometric precision** provided by the dynamic Bayesian filtering and scene-aware constraints.

Table 4: Quantitative comparison of the proposed hierarchical tracking method with state-of-the-art algorithms on three public datasets. Best values are bold.

Data	Method	MOTA	MOTP	FPPI	MT (%)	ML (%)	FP	FN	IDs	Frg.
AVG–Town Centre	**Proposed Approach**	**87.74**	**89.32**	**0.34**	**97.7**	**0**	**155**	**660**	140	**90**
	Klinger (2017)	42.2	57.4	2.6	26.5	19.5	1175	2820	**137**	184
	Liao (2018)	75.4	64.1	-	-	-	-	-	-	-
	Leal-Taixé (2011)	41.3	55.7	1.5	7.1	16.7	640	4776	243	271
	Pellegrini (2009)	32.3	55.1	3.6	4.8	2.4	1549	4091	893	889
PETS09_S2L1	**Proposed Approach**	**94.85**	**91.98**	0.19	**100**	**0**	133	**14**	63	**7**
	Klinger (2017)	94.5	76.2	**0.07**	89.5	0	**55**	183	17	19
	Ren (2018)	19.6	71.6	-	68	23.1	-	-	-	-
	Yang (2018)	91	77.2	-	18	0	-	-	15	-
	Liao (2018)	66	76.2	-	-	-	-	-	-	-
	Yang (2019)	92.1	91.9	-	100	0	189	185	**3**	-
ETHZ Bahnhof	**Proposed Approach**	66.86	**88.57**	0.27	**87.5**	**0**	**270**	**2320**	140	110
	Klinger (2017)	41.2	64.6	0.92	25.5	25	923	2734	291	330
	Yoon (2015)	**83.8**	79.7	-	72	4.7	-	-	**71**	**85**

While not designed as a deep learning system, the method demonstrates strong performance for a classical MOT framework.

4.2. Qualitative Observations

Visual inspection of trajectories across different sequences reveals that:

- The proposed prediction model successfully anticipates short-term pedestrian motion, even with abrupt direction changes.
- The affinity matrix effectively resolves challenging cases where multiple detections fall within overlapping predicted regions.
- The memory module recovers missed targets consistently in medium and long occlusions, preventing premature termination of trajectories.
- The hierarchical association strategy avoids common pitfalls of greedy matching and reduces identity switches considerably.

Figure examples (not included here) demonstrate clear trajectory continuity where baseline approaches fail.

4.3. Discussion

The overall performance demonstrates that combining **scene-aware priors**, **Bayesian prediction**, **multi-cue association**, and **memory-based recovery** yields a tracking system that is both robust and interpretable. Several key insights emerge:

1. **Scene priors are extremely effective**
   Even a simple binary “Interesting Places” map and one global scale factor dramatically reduce FP rates.
2. **Affinity matrix → powerful structure for reasoning under occlusion**
   It unifies geometric, temporal, and appearance information without relying on expensive global optimization.
3. **Memory module compensates for detector weaknesses**
   This component is particularly valuable in online tracking, where real-time recovery is critical.
4. **Hybrid detection remains competitive despite not using deep learning**
   This is a significant practical advantage for applications where computational resources or labeled data are limited.
5. **Trajectory completeness improves** due to the hierarchical association framework, which prevents fragmentation in highly dynamic environments.

4.4. Limitations

While the approach shows strong performance, several limitations remain:

- Sensitivity to scale-factor estimation in datasets with highly variable camera motion.
- Appearance modeling (based on color histograms) may fail under strong illumination changes.

- The hybrid detection pipeline, although robust, does not match the recall of modern CNN-based detectors.

These limitations suggest potential directions for future work, such as integrating deep appearance models or learning scene geometry adaptively.

5. Conclusion

This work introduced a hierarchical and scene-aware framework for multi-person localization and tracking using a tracking-by-detection strategy integrated with a Dynamic Bayesian Network. The proposed method was designed to address several persistent challenges in pedestrian tracking, including mutual occlusions, visual ambiguity, complex crowd dynamics, and the presence of false or missed detections in the observation stage.

A combined detector architecture—comprising background subtraction, HOG/SVM, and ACF—was employed to leverage complementary strengths of different detection algorithms. To increase reliability and reduce unrealistic detections, a scene-specific **scale factor** and **interesting-places (IP)** model were incorporated, reflecting geometric and prior knowledge about regions of feasible pedestrian presence.

The central component of the method is the proposed **affinity matrix** and **hierarchical data association framework**, which jointly utilize predicted motion, appearance similarity, and spatial constraints. Through this formulation, the system effectively resolves ambiguities in crowded scenes, maintains identity consistency, and reduces both false positives and false negatives. Additionally, a memory mechanism was designed to recover temporarily lost targets, improving robustness under long occlusion or detector failure.

Experimental evaluation on multiple public datasets demonstrated that the method achieves strong geometric precision, stable identity preservation, and competitive tracking performance compared to classical state-of-the-art approaches. Despite relying solely on traditional vision and filtering methods (rather than deep learning), the framework delivered notable improvements in reliability and interpretability—both key advantages for real-time and resource-limited applications.

6. Key Contributions

This thesis introduces a set of methodological contributions aimed at improving reliability, robustness, and interpretability in multi-person tracking systems. The main scientific innovations are summarized as follows:

1. Hybrid Pedestrian Detection Framework

A combined detection pipeline integrating background subtraction, HOG/SVM, and ACF detectors was introduced. This hybrid detector leverages complementary strengths of classical methods to reduce false positives and increase detection stability in crowded scenes.

2. Scene-Aware Probabilistic Modeling

We introduced a set of scene priors—including a scale factor and “Interesting Places” (IP) distribution—to reject geometrically implausible detections and regularize the spatial distribution of candidates based on dataset-specific context.

3. Dynamic Bayesian Tracking Architecture

A Dynamic Bayesian Network (DBN) was formulated to integrate motion prediction, observation likelihoods, occlusion handling, and hierarchical decision-making into a unified probabilistic framework.

4. Affinity-Matrix-Based Hierarchical Data Association

A novel affinity matrix was proposed to represent compatibility between predicted positions and current detections. This matrix drives a hierarchical association process combining:

- geometric consistency
- appearance similarity
- trajectory-level history
- occlusion-aware constraints

5. Memory-Driven Recovery of Missed Detections

A dedicated memory module was designed to temporarily store lost targets and reintroduce them when compatible detections appear, mitigatiang detector failures (false negatives).

6. Integrated Multi-Cue Tracking Strategy

The system jointly employs motion cues, appearance cues, spatial constraints, binary scene priors, and Bayesian filtering—yielding a tracking method that remains interpretable and stable under occlusions and crowded scenarios.

7. Experimental Validation on Benchmark Datasets

Comprehensive experiments demonstrate consistent improvements in geometric precision and identity consistency over several classical tracking baselines, highlighting the strength of the proposed hierarchical formulation.

The introduced methodology provides a solid foundation for future extensions, including integration with deep feature embeddings, multi-camera setups, joint 3D reconstruction, or real-time multi-robot systems requiring reliable pedestrian awareness.

In summary, this thesis offers a practical, interpretable, and effective solution for multi-person tracking in challenging environments and provides several paths for further research and system-level improvements.