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

The predictive model in a multi-object tracking environment with challenges like mutual occlusions, ambiguous backgrounds, appearance changes, and complex motions, is on the edge of drifting away from the actual position of objects. The trajectory’s errors would accumulate when tracks are followed by consecutive updates based on inaccurate locations (especially in online methods). On the other hand, even in the state-of-the-art methods, false positives and negatives are inseparable from detection outcome. In this work, we address both these problems. A simple method is proposed to predict the next probable location of each target based on the current location, direction, velocity, and also the mutual relationship between neighboring occluded persons. These predicted locations are refined using recursive filters and an affinity matrix is defined according to these next probable locations (existing tracks) and the current detections. Besides, a validation approach based on initial information of the scene and camera alignment is taken into account to decrease the number of false positives through a combined introduced detector. Moreover, retrieval of the missed detections, using a defined memory, improves the performance by putting false negatives into control. The proposed data association method in this paper can be applied in conjunction with any detection method through a tracking-by-detection approach. Experimental results on publicly available benchmark datasets indicate that our proposed method achieves higher geometric precision and tracking accuracy over many state-of-the-art approaches.

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

1. Introduction

The localization and tracking of pedestrians, has been always one of the widespread research topics; from the most basic computer vision applications to current hot topics of autonomous robots, self-driving cars, human-computer interaction systems and safety. Trajectories convey valuable information for the scenes analysis and thus generate vital input data for different applications (Klinger et al., 2017). Tracking purpose is the localization and making connections between goals positions over time, goal identifying and tracking each person’s path throughout the image sequences (Zhang et al., 2010).

While tracking targets maybe humans, vehicles, or animals, our focus in this paper is on pedestrians, which is the biggest challenge in crowded scenes, applying an online recursive Bayesian estimation approach.

There are two main difficulties; the predictive model in a multi-object tracking environment with challenges like mutual occlusions, ambiguous backgrounds, appearance changes, and complex motions, is on the verge of drifting away from the actual position of targets. The trajectory’s errors would accumulate when tracks are followed by consecutive updates based on approximate locations (especially in online methods). On the other hand, even in the state-of-the-art methods, false positives and negatives are inseparable from detection. In this work, we aim to settle both these problems.

The purpose of multiple object tracking (MOT) is to allocate a unique trajectory for each goal across the whole sequences. One sequence of multi object detection and tracking is presented in Fig. 1 as a block diagram.

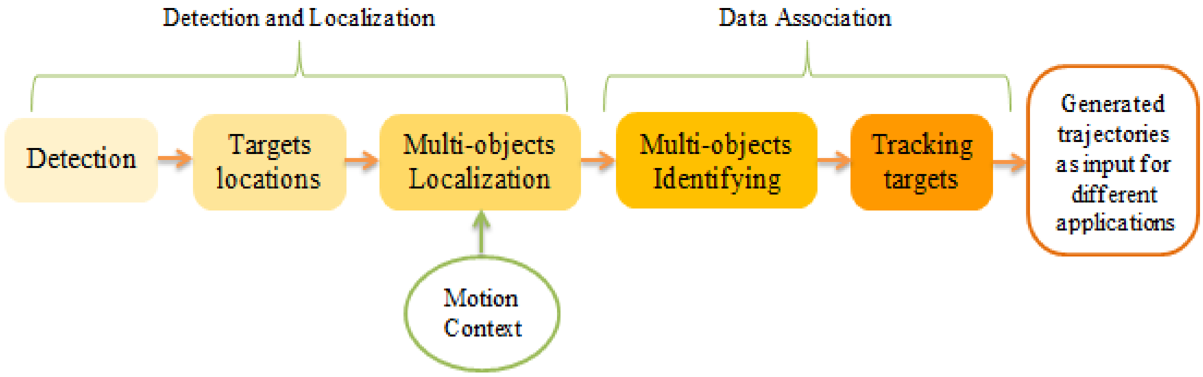
Detection and Localization
Data Association
Detection
Targets locations
Multi-objects Localization
Multi-objects Identifying
Tracking targets
Generated trajectories as input for different applications
Motion Context

Fig. 1. System diagram of each sequence in multiple object detection and tracking.

Detections in individual frames organize the observations in a recursive estimation framework. It is the initiation of many tracking systems. In situations with imprecise observations (resulted by the first block in Fig. 1), the generated trajectory is deviated easily (Klinger et al., 2017 and Ning et al., 2020). As mentioned before and seen in the diagram, in all tracking-by-detection methods (as the most frequent technique in MOT), tracking performance at the first stage relies on the detection accuracy.

To improve these first blocks efficiency in Fig. 1, additional information and helpful techniques like Bayesian models are applied; elimination of false positives by comparing them with previous observations is the general application of Bayesian models. So, at first step, a Dynamic Bayes Network is used to jointly inference the unknown parameters (the position of the goal tracked in the scene) for multi-objects localization block. In this study, Kalman filter combined with the observations obtained from the image, provide the evaluation possibility before subsequent processes. In such cases, discoveries are allowed to be corrected before using in the return filter.

in this paper, the unknown is estimated in a novel pedestrian detector by which the number of faults diminished significantly. However, complex scenes with long occlusions, unexpected motion changes and approximate detections are still challenging. To deal with these obstacles, we propose an Affinity matrix based on the existed trajectories and current detections.

Large number of false positives is an indispensable part of acquiring admissible detection rates that influences the reliability of automatic object detection (Stauffer and Grimson, 2000). Affinity matrix combined with our prior knowledge of the dataset, throughout a region specified by our novel scale factor, help us to lessen the risk of wrong detections. On the other hand, false negatives affect tracking performance. Often, failure to identify these cases is due to detector error. So, in the proposed method, by defining a temporary memory for each target, they can be almost completely controlled and recovered if the lost targets satisfy certain conditions.

Multi-object identification (next block in Fig. 1) is one of the most critical challenges in multi-object tracking. In common recursive works, applying classifiers in the target's categorization is usual. However, these strategies sharply diverge from the real object if the training data are stemmed from inappropriate examples. Tracking targets using the Hungarian algorithm and also global optimization methods are widespread. However, these methods not only can't handle occlusions in complex scenes but also have high computational complexity. In this study, tracking is performed separately for each individual in an additional step. Tracking-by-detection method in combination with the probabilistic approach is used in data association; moreover, linear programming is applied to find the optimal allocation of detections to trajectories in each frame.

As mentioned above, we also apply the previous information of targets, such as their locations, velocities, and appearances. Then, we define a disparity criterion based on this previous information. The appearance dissimilarity is calculated by matching color histograms.

The image sequences or video frames are in two-dimensional environments; while, the real-world coordinate is in three. As a result of the perspective features and image depth, the scale in front of the scene doesn't equal it at the end of the image. Considering the observed false positive or negative of detectors, it is clear that some wrongs happened aren't logically acceptable; for instance, a pedestrian on the sky or a very big or small detection compared to the dimensions of the image. To make the dimensions at all points of the scene analogous, we introduced a scale factor in this study. This factor can be calculated for each dataset experimentally or using our data prior knowledge (or even can be obtained during a separate learning algorithm). On the other hand, there is an additional challenge in datasets on moving platforms. In these conditions, for a steady enough camera movement (given a determined confidence coefficient) we will define the scale factor in a similar way.

The main scientific contributions of this study include:

- Proposing a new method of detection integrating high-efficiency methods and taking advantages of them.
- Defining an affinity matrix for each frame between the current detected goals and existed trajectories.
- Evaluation of detections using scene information and a novel defined scale factor.
- All in all, introducing a hierarchical data association method applicable in collaboration with all detection methods.

In the following, a review of our work's background is given in Section 2. In Section 3, the structure of our proposed method is explained. Experiments and results are represented in section 4. A conclusion on our work and the prospective future work are given in Section 5.

2. Related works:

As mentioned, detection and localization, data association, and motion context are three central building blocks of the proposed method. Frequently used tracking systems apply detection for each time-step, an association step to allocate detected objects to trajectories, and recursive filtering to discover linkages between image-based measurements and a motion model (Klinger et al., 2017). So, in the following, we focus on them, and for each block, we review the related works; also, in this section, to better clarify the place of our proposed method, we briefly mention our novelties compared with the related works.

2.1. Detection and localization

Apart from tracking techniques, different methods of detection are available to recognize goals (here pedestrians) in the scene. Currently, research topic of person detection has been allocated to features exploitation. Pedestrian features diversity and their large dimension are the most critical difficulties in pedestrian detection. Based on the theoretical analysis of six prevalent features of scale-invariant feature transform (SIFT), Haar, Histogram of gradients (HOG), speeded up robust features (SURF), local self-similarities (LSS) and local binary patterns (LBP), and also experimental results, Yao et al. (2015) in their work applied sparse representation and removed sparse feature subsets. Results demonstrated that sparse feature subsets are capable in the preserve of important components in these descriptors.

Combined approaches are prevalent. The combination usually happens for the detection procedures or their amalgamation with probabilistic strategies (Hernández-Aceituno et al., 2016). Hernández-Aceituno et al. introduced a Bayesian approach to the Viola-Jones algorithm that automatically detects pedestrians in image sequences. They presented a probabilistic interpretation of the first version of this algorithm and extended a method to estimate convolutions of statistical matrices. The resulted accuracy of this approach improved in a noticeable degree; however, there is still a high number of false positive. We apply both techniques in combined approaches. We aggregate different detection techniques with a focus on the number of FPs and FNs. We also take advantage of probabilistic strategies in combination with our proposed detector to avoid occlusions and missed detections by defining an affinity matrix and extra features.

One of the biggest obstacles in real-time human detection is the computational cost necessary in complex images. Ko et al. (2014) introduced a fast-human detector applying optimal levels of image scale and their adaptive region-of-interest (ROI). The proposed algorithm was effectively applied in real-world surveillance video sequences, with higher detection accuracy a better performance than those of other related methods. However, the reported number of FP in this study isn’t satisfying.

As a result of more reliable detector development, applying tracking-by-detection methods is getting commonplace. In this approach, detections are derived from detectors independently at each time-step and joined to generate trajectories. Yu et al. (2016), Ren et al. (2018), Dollar et al. (2014) and Stadler and Beyerer (2022) are examples of using this approach. Tracking-by-detection methods usually count on autonomous single frame detections. If these detections are inaccurate, trajectories are inclined to be updated towards wrong positions (Klinger et al., 2015). The proposed method by Yu et al. (2016) didn’t flourish to track humans far from the camera or people who walk closely as a result of the detector’s defeat dealing with these objects. This method follows an online tracking-by-detection approach, so the efficiency of the proposed tracking method, depends inevitably on the performance of the applied object detector. The method proposed in our paper is similar to this work in some aspects such as following the online tracking-by-detection approach and also appearance information, but we introduced a more efficient detector, and we don't rely only on it. We also applied the previous and feature knowledge of trajectories in each space.

As mentioned, a high number of false positives and acquiring acceptable detection rates are always accompanist. Additional clues like appearance, shape, and foreground information are required to alleviate the number of false positives (Klinger et al., 2017). Non-maximum suppression (NMS) is a technique to remove and control false positives related to one goal (detected more than once). Some authors, like Hoiem et al. (2008) and Klinger et al. (2017) worked on improving NMS approaches. They both shifted dimensions of the objects and the scene into 3D space and estimated the validations. For sequences in which training data is available, Klinger et al. (2017) also applied prior knowledge of the scene in the non-maximum suppression process. In this paper, we apply the idea of measurements transfer to 3D and generalize it by defining a scale factor without extra computation. Moreover, in addition to the generated mask resulted from the detector, we apply the primary information of the scene to eliminate the goals entrance or departure.

During the sequentially processing phase in image-based observations, errors in one step aren't capable to be checked and removed in the subsequent steps. To tackle this problem, Hoiem et al. (2008), Schindler et al. (2010), and Choi et al. (2013) incorporated different sources of information in a probabilistic graphical model framework, i.e., Bayesian networks. Felsberg and Larsson (2008) proposed a new approach in Bayesian tracking named channel-based tracking. Klinger et al. (2015) used a Dynamic Bayes Network with the tracked goals location as unknowns. Applying a dynamic model, the human detector and classifier, the unknowns are estimated in a probabilistic framework. We followed this approach to update locations.

Klinger et al. (2017) and Choi et al. (2013) utilized a joint probability model of observed and hidden variables. Observations include various sources of information. The main merit of this method is that errors that happened at one of the system elements can be corrected by the other observations and components. In the present paper, we extend their work with a different defined probabilistic graphical model. So, we introduce a binary model of the joint distribution. In the occluded scenes, making a decision is done through scores. On the other hand, different information sources in the approach by Klinger et al. (2017) not only put the burden on time and computational complexities, but also divert trajectories from their real paths. We apply a different combination of variables.

LiDAR-based pedestrian detection and tracking is one of the currents applying methods with applications such as security surveillance and human manner analysis. In spite of the LiDAR's high-resolution sensing capabilities, like any other approaches, there are difficulties attached to this method. Wang et al. (2022) addressed the LiDAR-based method's limitation in occlusion and complex movements applying three-dimensional measures.

These days, Deep Learning and neural networks are among the most powerful research topics improving rapidly. Detection isn't an exception and there is a vast amount of works done in this field. Although, our proposed association can be matched with any detection method, since we didn't apply this approach in this paper, we only limit ourselves to mentioning a couple of reviews. Ning et al. (2020) and Pervaiz et al. (2021) both, summarized the traditional and deep learning method's progress, results, research's issues and future outlooks of the problems. In addition, Pervaiz et al. (2021) introduced the pertaining dataset and evaluation criteria. As a result of extended researches, Stadler and Beyerer (2022) only focused on three deep learning's applications (video surveillance, human-machine interaction and analysis), comparison between two fields of applied technology (2D and 3D vision systems) and methods and strategies. Rahmaniar and Hernawan (2021) emphasized on embedded systems and their high processing powers using GPU and deep learning techniques in embedded platforms. Moreover, they represented a review of human detection models and their performances comparison.

2.2. Data association

In a multi-objective tracking scenario, an allocation problem must be solved to clarify the possible concords between detected goals and existed trajectories. Dedicating a weight into any possible conformity between detections and trajectories, relying on additional information and cues, a globally optimal solution can be discovered. Applying the Hungarian algorithm in polynomial time (O (n^3)) or probabilistic strategies like joint probabilistic data association (JPDA) introduced by Fortmann et al. (1983) are common approaches in this field. These methods are applicable for multi-object tracking, but constraints cannot be considered without using workarounds (Klinger et al., 2017). Yu et al. (2016) and Yang et al. (2021) proposed a hierarchically solution to the data association problem using the Hungarian algorithm with outputs of both independent trackers and detector.

Stadler and Beyerer (2022) also implemented a sequential data association, taking advantages of tracking-by-detection approach to cope with assignments problem in occluded environments. They also proposed a motion model to handle reidentification. There is an allowed delay in assignment in ambiguous situations, until the condition is clear again. Analogous to our proposed association method, this multi-hypothesis approach is applied on any tracking-by-detection framework and it applied its own introduced distance matrix.

Linear programming is a more recently proposed approach. In this method, an optimization problem provides the optimal dedication of detections to trajectories in every frame. Leal-Taixé et al., (2011) applied linear programming over the entire image sequence. They proposed a novel model to associate weights that are calculated depending on the predictive model and a classifier trained at runtime. We apply Linear Programming and define the linear equation and binary weights in a straightforward, efficient way. The weight in this method is earned using an introduced Affinity Matrix based on both the current and the previous knowledge of the scene. In addition to this matrix, we consider more predefined rules and thresholds according to the data.

In work presented by Yu et al. (2016), relying on different cues and information achieved from the attribution problem, detections and trajectories are divided into groups. These divided sets are assigned to each goal using the greedy algorithm. The proposed method operates through a hierarchical association framework. We extend this work and do all the different steps together and decide once. Besides, the performance of the method introduced by Yu et al. (2016) depends on the object detector efficiency. However, in our work, detection and data association are in combination. We also handle the problem of trajectory fragmentation over long occlusions.

2.3. State prediction and motion context

The motion model is a perception of a first-order Markov chain. Constant velocity and motion smoothness are usual assumptions in the motion model. Given the previous estimation of the state parameters, prior knowledge for the current frame is acquired. However, for prolonged occlusions, the first Markov chain is on the edge of deviation from the objective's correct position. On the other hand, people are required to show reactions to their environment. So, independent tracking of individuals is often not efficient enough. To handle this challenge, recent studies in this field work on contraction between pedestrians over the sequence (Klinger et al., 2017).

Many works have been done on modeling motion and formulating trajectory estimation as an energy minimization problem. In the study done by Milan et al. (2014), energy is defined as the sum of many penalty values such as aberration from a predicted manner like moving towards a specific goal or staying away from the collision. Pervaiz et al. (2021) and Yang et al. (2021) both defined their linear motion model based on the Kalman filter method and made prediction of the human 3D position.

Applying optical flow features is another method of motion context estimation. Leal-Taixé et al. (2014) applied these features in training a Random Forest classifier. The problem of optical flow is its usage limitation on online and also dynamic applications. Li et al. (2019) also proposed another new multi-object tracker following the popular tracking-by-detection scheme. They deal with the camera motion challenges with optical flow approaches and an extra tracking algorithm to end the missing detection problem. We extend the work by Leal-Taixé et al. (2014), and we apply background subtraction (BS) instead, to decrease the computational complexity caused by optical flow features extraction. The problem attached to these approaches is that they aren't applicable to the moving camera datasets.

What makes motion prediction challenging is that human's intention and many other factors like social relations, environment, and crowds affect their motion. In the study done by Rudenko et al. (2018), the authors formulated the task as a problem of the Markov decision process (MDP)[1] with a probabilistic approach. They introduced a weighted random walk algorithm in which each agent is impacted by its neighboring agents. They integrated local information of social groups, extracted the constraints imposed by the groups to their member's motion and applied this information in the motion model. This work didn't consider the people relation inside groups.

Applying the Gaussian Process approach is another method in motion context estimation. In most studies done based on Gaussian Process, the trajectory is treated as a regression problem. Ellis et al. (2009) followed this technique and estimated velocities based on the previous observations of the goals. Computational complexity (depending on the complexity of the scene) is the biggest weakness of this method. Kim et al. (2012), in their study, focused on the detection of the regions of interest using GP regression. Generated trajectories are in correlation depending on a specific distance from their current positions. This study got 2D locations as input and aimed at their velocities as target variables. Klinger et al. (2017) followed the related work and used a Gaussian Process Regression (GPR) model with velocities as target variables. In their work, a certain distance between the current object positions and their motion directions is defined. We generalize applying regression and consider only some previous time-steps. The number of frames is later evaluated as the problem parameter. The direction, velocity and crowed are all kept in mind.

[1] Markov decision process

3. Introduced multi-person tracking approach

In this section, a new hierarchical approach for multiple persons tracking is described. Observations are resulted from a pedestrian detector, prior information of the scene and the memory contents (section 3.4). The approach is based on tracking-by-detection techniques, each person's state at every time step (section 3.1) and vision-based observations. For each frame, an affinity matrix is defined between the current detected goals and existed trajectories (section 3.5). Considering neighboring pedestrians, detections extracted from the detector are associated either to an existing or to a new trajectory (section 3.6). The updated state then refines image positions and yields the basis for the prediction at the next time step (section 3.2 and 3.3). After data association, there may remain some unassigned detections or lost trajectories; it's the time for our affinity matrix to be completed considering a defined memory or the time for this memory to be filled respectively (section 3.7). The proposed multiple object tracking block diagram at each time-step is shown in Fig. 2.

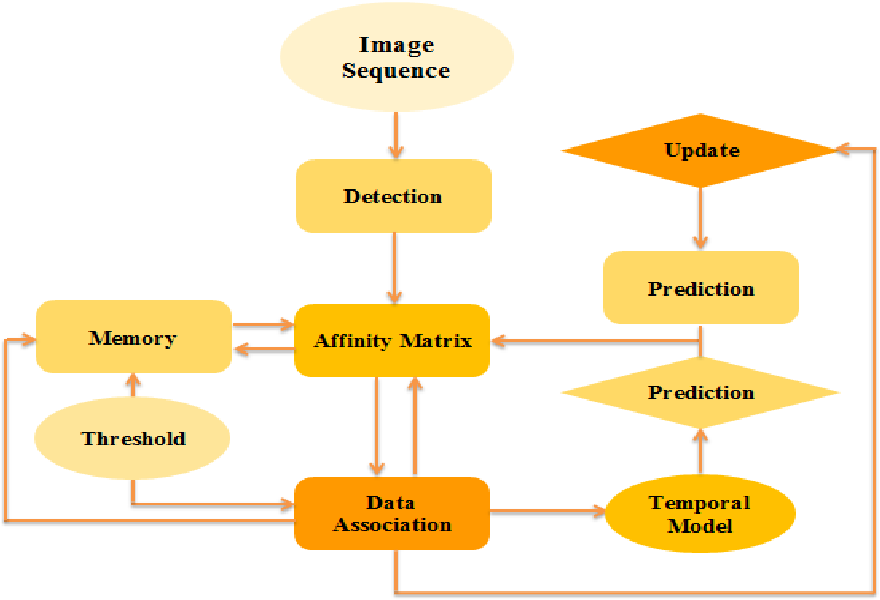

Fig. 2. System diagram. * Colors show the running steps of the algorithm (from light to dark). * Steps related to the Filters (like KF) are displayed by the lozenge, and steps related to the data and prior knowledge represented by oval.

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).

3.1. Proposed Dynamic Bayesian model

Many cases of the tracking systems are based on the detection in singular frames; in these methods, observations are organized in recursive estimation structures. Challenges on the scene including people's occlusion and complicated movements don't yield the prospective acceptable multi person tracking. For a tracking approach to be more reliable, we applied a Dynamic Bayesian Network (DBN). After extraction of measurement, features and observations at each frame, the network determines their relationship through frames. In this network, the location of the tracked person in the scene is unknown. In the proposed method, locations in the scene are related to the detections and hierarchical algorithm, considering the affinity matrix (Section 3.5). The joint probability density function of all variables can be factorized in accordance with the network structure as follow:

$$\mathrm{P}(\mathrm{x}_\mathrm{i}^\mathrm{t}, \mathrm{w}_\mathrm{i}^\mathrm{t}, \mathrm{w}_\mathrm{i}^{\mathrm{t}-1}, \mathrm{c}_\mathrm{det}^\mathrm{t}, \mathrm{IP}) \propto \mathrm{P}\big(\mathrm{w}_{\mathrm{i,t}}|\mathrm{w}_{\mathrm{j}=1,\dots,\mathrm{n},\mathrm{t}-1}\big)\mathrm{P}(\mathrm{x}_\mathrm{i}^\mathrm{t}|\mathrm{w}_\mathrm{i}^\mathrm{t}, \mathrm{IP})\mathrm{P}(\mathrm{c}_\mathrm{det}^\mathrm{t}|\mathrm{x}_\mathrm{i}^\mathrm{t})\mathrm{P}(\mathrm{x}_\mathrm{i}^\mathrm{t}|\mathrm{O}^\mathrm{t}). \quad (1)$$

Where, w_i^t is the system state and $x_i^t$the position of person *i* in the image at time step *t*. *IP* and *C*$_{det}$ are respectively the interesting places and detector confidence. In the following, we describe different terms of this formula.

For the computation of the posterior state $\mathrm{w}_\mathrm{i}^\mathrm{t}$ of our model a dynamic filter model is used. We apply Kalman and Unscented Kalman filters. The state vector $\mathrm{w}_\mathrm{i}^\mathrm{t} = [X, Y]^T$consists of the 2D position of each goal. There is an intermediate step of prediction through a temporal model that in combination with the filter's prediction is applied in computation of the image position variable $\mathrm{x_i}$. This magnitude is then used for the correction of the predicted state.

The goal's position in the image depends on its position in the previous frame, its speed and direction, detection occlusion, and the crowd on the scene. The challenges of occlusion and crowd is encountered using our affinity matrix. Based on the next predicted positions (Section 3.3) and the initial information of the scene, this probability $(\mathrm{P}(\mathrm{x}_\mathrm{i}^\mathrm{t}|\mathrm{w}_\mathrm{i}^\mathrm{t}, \mathrm{IP}))$ gets zero or one values.

Targets reference points are depicted by $\boldsymbol{x_i} = [x_i, y_i]^T$ for person *i*, where x_i and $\mathrm{y_i}$ are the column and row coordinates of goal's (this point is the bottom center point of the boundary rectangle). In our model the variable x_i is observed using the pedestrian detector. In situations, in which the detector misses the goals, the variable is only predicted by dynamic filters and considering the observations and previous information.

Interesting places

As mentioned above, *IP* addresses Interesting Places; it's designed to emphasize on regions where pedestrian's presentation occurs with higher frequency. A thresholding algorithm is set on detections according to pedestrians' distribution and camera location of applied dataset. Considering this threshold, the model credits discoveries. Figure 2(b) represents an example of the learned distribution related to one of the image sequences applied in our experiments. The situation in which the tilt angle of the camera doesn't remain approximately constant throughout the sequence (Fig. 3 left), the probabilities fluctuate in accordance with changes; However, as the camera slope remains almost constant during the sequences used in this data, the function at the bottom of the image remains almost unchanged (Klinger et al., 2014). Resulted probability $(\mathrm{P}(\mathrm{x}_\mathrm{i}^\mathrm{t}|\mathrm{IP}))$ for the whole test sequence is applicable and valid for any point and consecutive calculations aren't required. Unlike the related previous methods, in which a distribution of values between zero and one was used, in our proposed method, binary values are considered. To remove the unlikely areas, these probabilities can be calculated experimentally based on the initial information of the scene or learned through a separated algorithm as the work mentioned above. Then, the thresholds of possible areas for the individual’s presence are extracted.

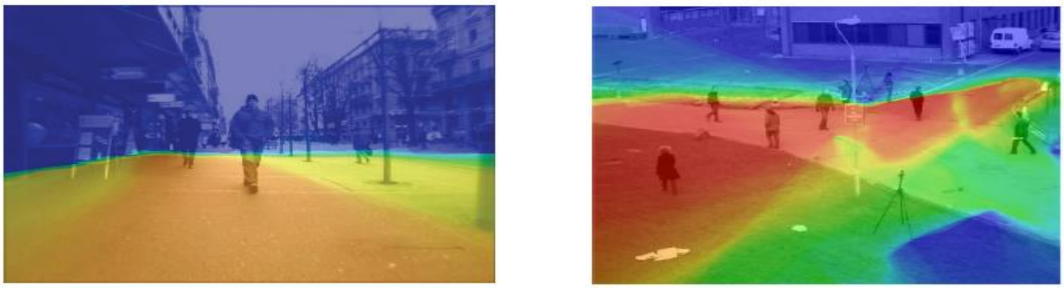

Fig. 3: Interesting places calculated for two different datasets (Klinger et al., 2017). More occurred areas are visualized in red and blue regions are referred to the area with no pedestrian through the image sequences of ETH-Bahnhof (left) and PETS09 (right) datasets.

Detector confidence

$P(\mathrm{c}_{\mathrm{det}}^{\mathrm{t}}|\mathrm{x}_{\mathrm{i}}^{\mathrm{t}})$ indicates the probability density function for any person’s presence in the scene. The magnitude is extracted from our introduced combined detector (Section 3.4). If at least one of the detectors recognize the person *i* at location x_i as a goal, the mask magnitude for the location will be one; otherwise, the magnitude equals zero. The resulted pdf is planned to focus on the areas in which pedestrians are mostly visited and is computed at each time step.

Occlusion model

Binary indicator O is used to model mutual occlusions of pedestrians. Conditional upon the person’s position, this index depicts whether a goal is supposed to be occluded or not (Klinger et al., 2014). Therefore:

$$P(x_i^t|O^t) = \begin{cases} 1, & \text{اگر}\; O(x_i^t) = 0 \\ 0, & \text{اگر}\; O(x_i^t) = 1 \end{cases} \tag{2}$$

In our proposed method $O(x_i^t)$ and $P(x_i^t|O^t)$ aren’t defined for each position; instead can be estimated for every pedestrian applying the Affinity Matrix (Section 3.5) and the distance between references points. Different conditions of occlusion are treated in data association.

3.2. Temporal Model and Prediction

People’s appearance features and their distances are useful cues in tracking; however, not adequate on their own (more specifically in cases with similar appearances and also in crowded environments). People in their trajectories react to the environment and to other people; so, motion model extraction in tracking-by-detection methods is very important.

In our introduced multi person tracking method, a dynamic model is proposed based on people’s position and the crowd around them (determined by the affinity matrix). In this model, the velocity and orientation for each person are interpolated depending on their velocity during n previous frames. This parameter is investigated in experiments to estimate the optimal magnitude. In crowded locations, assuming that people movement is more limited, the magnitude of n equals 1; so, in such situation the velocity is only dependent on the velocity in one previous step. In each frame, velocity is calculated in two directions through constant acceleration equations and the position in the next time step (frame) is predicted as follow:

$$X_t = X_{t-1} + v_{X,t}\,\Delta t \tag{3}$$

$$Y_t = Y_{t-1} + v_{y,t}\,\Delta t \tag{4}$$

These values in addition to filter's predictions are considered (section 3.3) as the next probable positions for each individual.

3.3. The scene model and observations

Large numbers of reported FP in detections, mostly happen outside the interesting regions or in areas with zero chances for goals presence. Considering images related to each dataset, location and orientation of the camera, perspective and depth of the images and the camera moving style, a significant number of FPs can be eliminated and enhance the tracking accuracy. In our work, a set of possibilities (or probabilities) for people to be presented at each point is applied. As mentioned above, there is only a binary probability for appearance or absence of a person at a point.

Taking the constant direction of the camera into consideration (in datasets with moving platform), we assume that the initial information of the scene (within a determined confidence) can be transmitted to all frames.

In addition, given the location and orientation of the camera and the perspective of the scene, a scale factor is defined; therewith dimensions and distances of the detected goals will be analogous. Using this factor, the height of the boundary rectangle of the discovered target is transferred into a metric scale proportional to the height of the discovered person. The proposed factor is used in two sectors. The first sector deals with acceptable heights for rectangular boundaries of detections, and the other is related to the probable areas defined at each step for each person's presence in the next frame (time step). Both applications are presented in Fig. 4. In this figure, probable areas and rectangular boundary dimension discrimination in the beginning and at the end of the scene, with different distances from the camera is obvious. All figures in this section are represented in the AVG-TownCentre sequence.

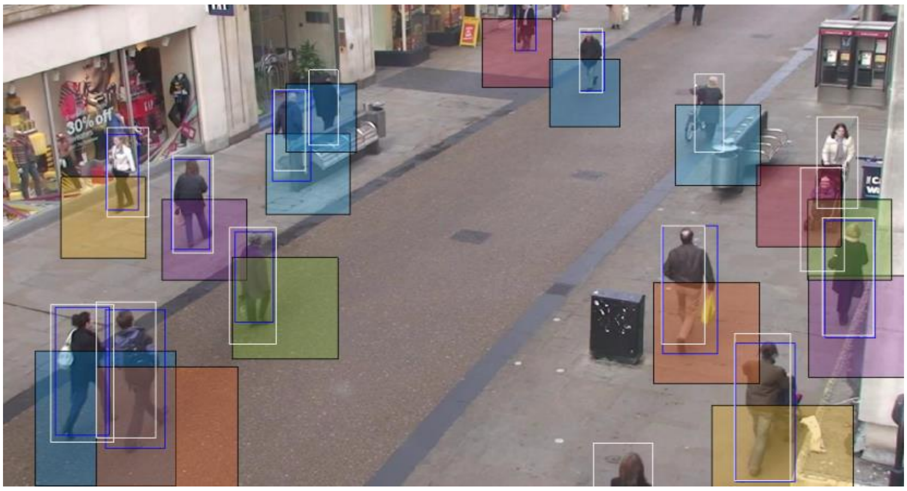

Fig 4: probable areas and rectangular boundary dimension difference in the beginning and at the end of the scene in the AVG-TownCentre sequence.

Similar to prior knowledge of the scene’s estimation, the scale factor is different for each data set and must be obtained in advance, by learning the distribution of the sequentially observed positions in a supervised approach or using trial and error methods.

3.4. Pedestrian Detector

In this study, a combination of three common detectors is used to detect individuals. The effects of Scale Factor and also presence/ absence in the IP is applied simultaneously.

In order to increase accuracy and reduce false positive discoveries, the background is first subtracted. The Background subtraction method is a common way to separate background elements from foregrounds, which is accessible by generating a background mask. This method is used to detect moving targets before a stationary camera. There are different methods to remove the background. It should be noted that in the dataset with a moving camera (like ETHZ Bahnhof) the background subtraction isn’t efficient. The result is represented in Fig. 5.

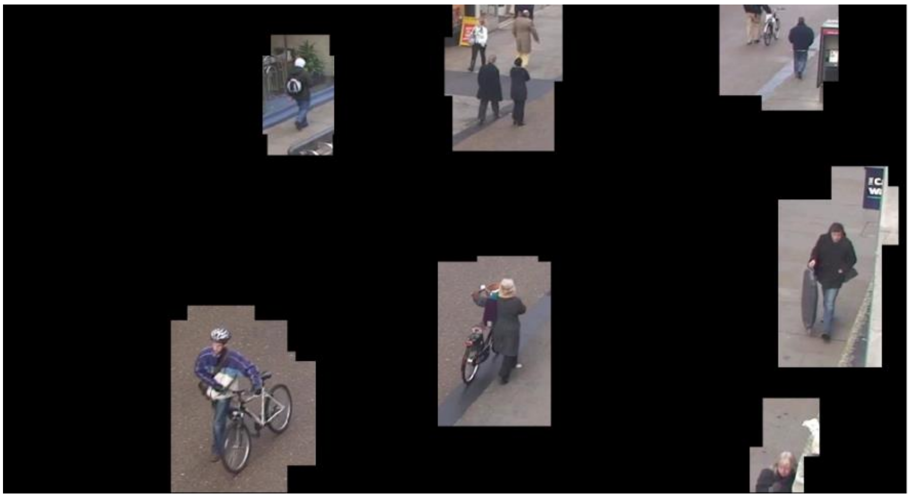

Fig 5: Detector's mask resulted by background subtraction in the AVG-TownCentre sequence.

We first apply the Gaussian Mixture Models for background subtraction by which the moving objects are segmented from the background. The method employs differences in pixel values and divides the motion pixels. Finding dissimilarities is operated based on pixels neighborhood values in consecutive frames by Chi- Square statistics. Then, a blob hole filling step is applied to fill the holes generated as a result of time difference. In the next level, a morphological erosion operation is carried out for noise elimination.

Areas in the resulted mask, without any farther processing, can be considered as goals. It is possible only when targets are objects like general moving vehicles; and also, in these cases, it is easy to recognize targets among identified groups that gather in the same sections. However, tracking targets such as pedestrians will result in many wrong detections. To solve this problem in this study, the remained area (pertaining the moving parts) is considered as a mask, on which two detection algorithms based on HOG/SVM and ACF[1] detectors are jointly applied. The combined approach increases the detection strength. Since two employed detectors follow different approaches, they have different detection capabilities. The detection results of ACF detector and HOG/SVM on the AVG-TownCentre sequence have been shown in the Fig. 6 and 7, respectively.

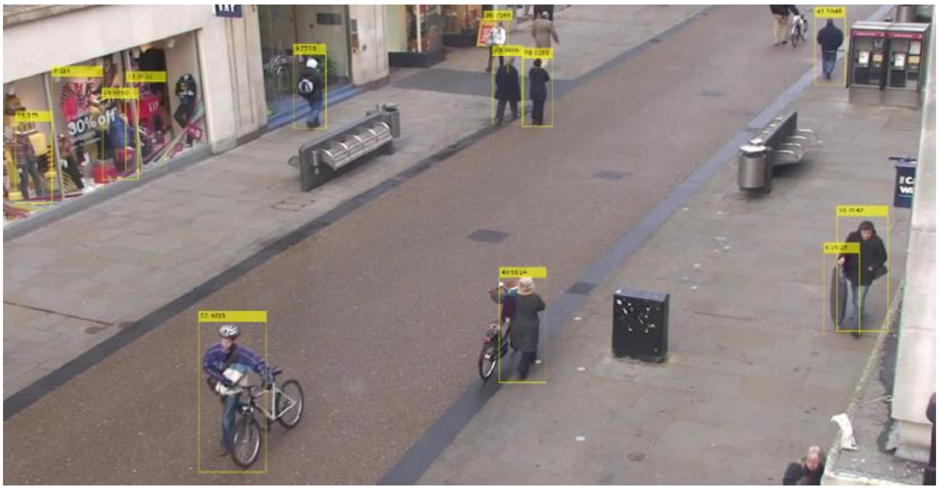

Fig 6: ACF Detector that extracts Aggregated Chanel Features and uses Adaboost classifier on the AVG-TownCentre sequence.

In each method of detection, it is possible to generate a group of boundary rectangles around any correct target. That happens as a result of the basis of the detection algorithms applied on the sliding window approach. In this method, depiction of many positive detections near the goal's true location is prevalent. Then, NMS nominates the detections mostly scored (Fig. 8 (Klinger et al., 2017)). As it can be seen in the Fig. 8, if there are still such cases, a threshold on overlap on the unity will be used; so that if the overlap is more than the threshold magnitude, the boundary rectangles will merge with an averaging method. The final results of the proposed hybrid detector can be seen in the Fig. 9.

[1] Aggregated Chanel Feature

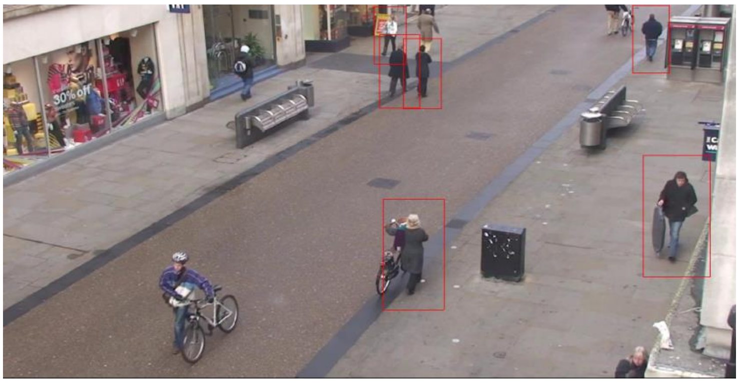

Fig 7: HOG/SVM Detector on the AVG-TownCentre sequence.

Detection k at time t is indicated as $d_t^k = [x_{d,t}^k, y_{d,t}^k, v_{d,t}^k, \dots \dots.]^T$ and a group of detections on frame t is in the form of $D_t(d_t^k \epsilon D_t)$, with indexes of $k \epsilon K_t \cong \{0, 1, \dots, K\}$; (zero is the representative of the missed trajectories).

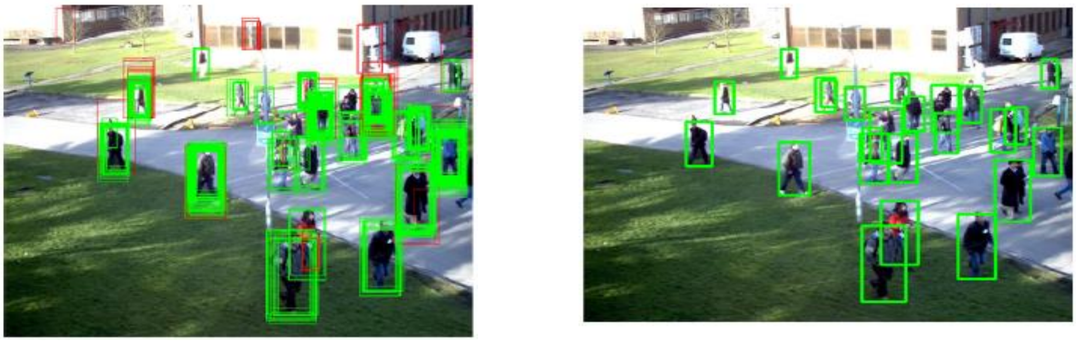

Fig 8: Raw detections (left), detections after NMS application (right) (Klinger et al., 2017).

Inria model is applied for static camera datasets and *caltech* model, with high diversity in detected pedestrian dimension, for dynamic one. Number of scales per octave is considered 8 and no maximum suppression is done according to the rectangular scores. The score demonstrates the confidence of the detector yielded by a cascade classifier and receives the magnitudes of minus infinity to plus infinity. The larger the magnitude, the more confidence is.

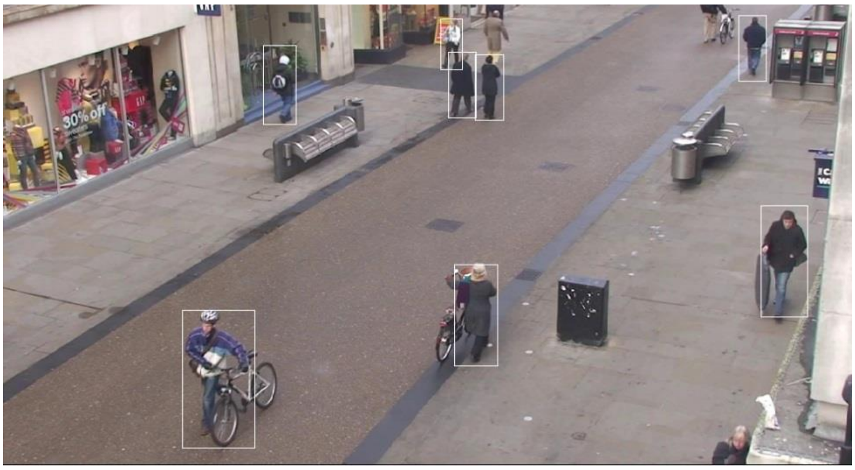

Fig 9: Combination of ACF and HOG/SVM Detectors in the AVG-TownCentre sequence.

3.5. Affinity Matrix

After each time step's detection, the proposed affinity matrix is generated between the current detected goals and the existed trajectories (in columns and rows respectively). For each goal (pedestrian) in the previous frame, a probable next area is defined in which the person might appear in the next time step (shown in Figure 4). In case any detected goal in the current frame appears in each probable area (each row), the related column would be given a figure.

This is a zero and one matrix, in which the rows represent the goals in the previous time step (existed trajectories in the current frame) and the columns show detections in the current frame. In an ideal circumstance, for each column there is one and only one column with a nonzero magnitude in rows. However, in real situation, some special states are possible to happen. Some of them are presented in Fig. 10. Over the time, not only states change but also the number of pedestrians presented in the current scene changes. So, at first the matrix dimension is proportional to the number of trajectories and detected goals. Then, in the following steps, it will be completed.

Referring to the affinity matrix, in cases where only one column of the matrix is related to only one row, considering a predetermined threshold of appearance similarity to the related goal in the previous time step, the label assigned to the referred row (in the previous frame) would assign to this column (corresponding trajectory).

The appearance similarity is investigated calculating the difference between color histogram of a goal (person) in the current frame analogous with the previous one. The comparison of appearance features is essential because, the goal appeared in this location may be related to a new or an occluded person or the considered goal may have already left the scene. So, the assignment can be done just for satisfied thresholds.

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

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

Fig 10: Possible states in Affinity Matrix.

Fig. 10 up left, with one total zero row, shows the situation in which no detected goal in the current frame appeared in the probable areas of the existed trajectories. The matrix represents the conditions that the related goal has left the scene or it is not detected by the detector (FN). If this situation happens in the middle of the scene, where according to the scene information (primary information of the dataset) person entrance is impossible, a predicted detection with this label is generated at the predicted next center for the *m* next frames. Tracking this generated person through prediction (without detection) continues up to *n* frames. After that, the label assigned to this person along with the appearance features, the last frame viewed and the reference point in this frame are stored in memory.

Fig. 10 up right with a column of zero indicates the conditions in which no trajectory is related to the current detections. This situation may happen in two cases; when the related goal is a pedestrian that has just entered the scene that must be assigned to a new trajectory or a situation when detected goal is a false positive resulted due to the detector's error (FP). To discriminate these conditions from each other, prior knowledge of the scene is applied. For example, if the detection is on the margins of the scene, it is possible to be allocated to a new path; otherwise, if the result is not comparable to a lost target in memory, this discovery is considered a FP and will be deleted.

In cases with more than one non-zero elements in a column, the current detection is compared with all non-zero rows in terms of reference distances and appearance characteristics and is assigned to the least different label. In the next two matrices (Fig. 10 down), it can be seen that the number of non-zero rows isn't equal to the number of nonzero columns. To better clarify situations that result in such affinity matrices, we have Fig, 11 in which squares represent probable areas and circles show the detections in the frame. The situations related to the second-row matrices of Fig 10 (left to right) are shown in Fig. 11 (left to right).

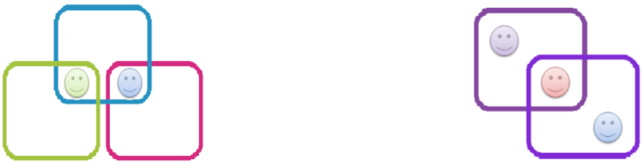

Fig 11: Possible situations of the scene due to occlusion. Missing a pedestrian, entrance or leave the scene, that all result to the detections wrong assignment to trajectories.

In such conditions, discriminations are calculated, scored, ordered and then assigned according to their scores. For the cases with a smaller number of columns (Fig. 10 down left), a column or in other words a missed detection must be added. This column will be assigned to the least score row. And in situation with a smaller number of non-zero rows (Fig. 10 down right), a row needs to be added. An example of Affinity Matrix in a time step and its corresponding frame are shown in Fig. 12. Considering the matrix, the relationship between trajectories number 9 and 10, 18 and 19 and also 14 and 16 are clear. This dependency can also be seen in the relevant frame.

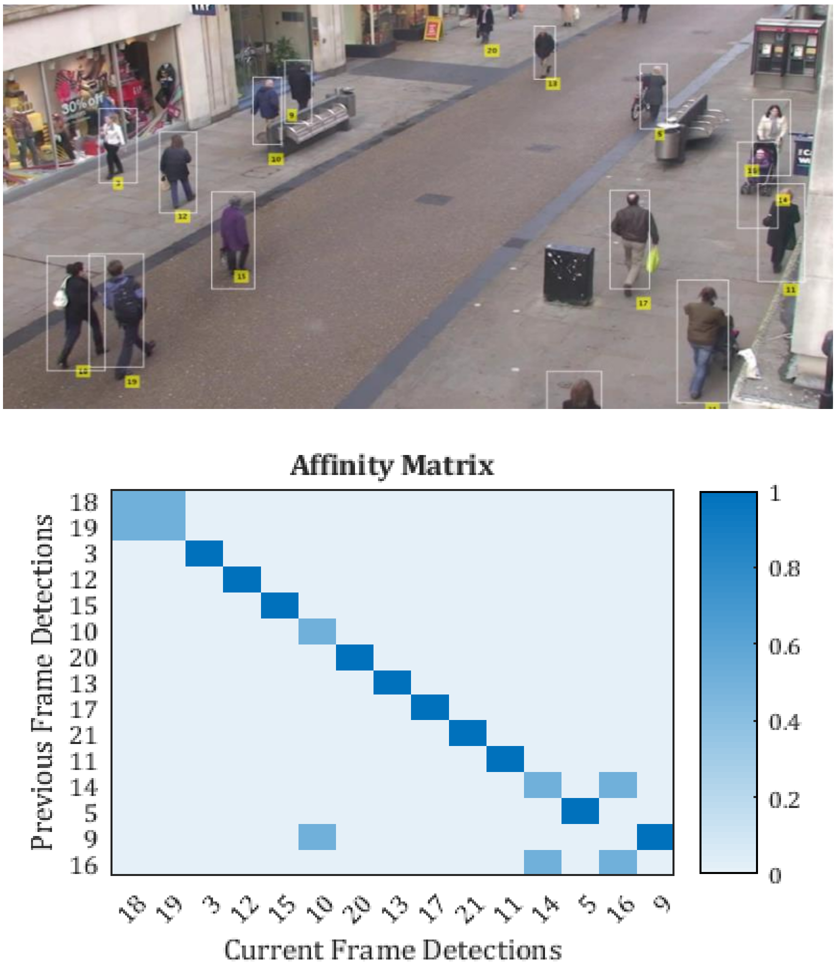

Fig 12: Affinity Matrix in a time step and its corresponding frame.

3.6. Data Association

Multi-person tracking algorithms based on tracking-by-detection methods are among the most common tracking approaches in recent years. Hybrid optimization and the dependency model are two main parts in multi-objective tracking to solve ambiguities in detection. Multi-object tracking receives the detector response and is formulated as an optimization problem. Short pathways or detector responses are learned by apparent characteristics, movement, position and size of the target along the video to long distances.

In our proposed approach, optimization is based on both past and future information sections in an online tracking. During online tracking, $T^i_{1:t} = \{x^i_k | 1 \leq \ t^i_s \leq k \leq \ t^i_e \leq$

$t\}$ shows trajectory of the person *i* up to time step *t*. These paths start at frame t_s^i and end at frame t_e^i. In addition, a set of trajectories until time *t* is indicated with $T_{1:t}$. Online multi-person tracking is considered to be the optimal combination of path $T_{1:t}$, with maximization of the posterior probability of $T_{1:t}$, given all detections until the frame *t*:

$$T_{1:t}^* = \ arg_{T_{1:t}} \max p(T_{1:t} | \ D_{1:t}) \tag{5}$$

Any pedestrian's path is provided by sequential states such as location $[x_i^t,\ y_i^t]$, velocity $[v_{xi}^t, v_{yi}^t]$, histogram in colors "$ColHist$", life span "age", rectangular boundary "$bbox$", next probable area "pre" and display possibility "vis". The target vector *i* in frame *t* is defined as:

$$s_t^i = \ [x_i^t,\ y_i^t, v_{xi}^t, v_{yi}^t, vis_i^t, ColHist_i^t, pre_i^t, age_i^t, bbox_i^t]^T; \tag{6}$$

Histogram of colors is applied in comparison of the detected goals and existed trajectories. Display possibility for each person at the beginning of its path equals *zero* and after *n* appearances in sequential frames turns into *one*. This parameter is defined to remove false positive detections and avoid wrong generation of new trajectories.

Following Klinger et al. (2017), to explore the optimal assignment, a joint probabilistic data association strategy is applied. At each frame, a set of *n* detections (D = {d_1, …, d_n} with d_k = {$x_{,k}$, $y_{,k}$}^{T} showing the reference point's coordinates) and *m* trajectories (T = {T_1, … , T_m, T^*} with likely new trajectories of T^*) are considered. With restrictions through which there may be only zero or one trajectory to be assigned with each detection, and for each trajectory there may exist only zero or one detection to be assigned with, the data association problem can be formulated as an integer linear program with binary variables:

maximize $\sum_{k\in[D]} w_i^k a_i^k$ *subject to the constraints*

$$\sum_{i\in[T]} a_i^k \ \leq 1\ \forall k \in [D], \tag{7}$$

$$\sum_{k\in[D]} a_i^k \ \leq 1\ \forall k \in [T \backslash \tau *],$$

where $a_i^k$is a binary indicator variable for the event that detection *k* is associated with trajectory *i*, and $w_i^k$is the weight of that association that's mainly assigned with *one* and *zero*. The magnitude is determined by Affinity Matrix at each time step and in occluded areas gets values between *one* and *zero*. [D] denotes the set of detection indices and [T] denotes the set of trajectory indices.

Association process between detections and trajectories

The process of association between detected goals and existed trajectories follows the equation:

$$A = \{assign^{i,k} |\ i \in T, k \in K\}. \tag{8}$$

where $assign^{i,k} = \{0, 1\}$; when detection k assign to the path i, the magnitude will be equal to *one* and otherwise will be *zero*.

3.7. Memory and information retrieval

As mentioned in section 3.5, in cases where a path lost in the middle of the scene, assuming the detector error as a false negative, the target is appended to our memory to be investigated later. The stored information includes the label assigned to this path, the appearance attributes, the last frame observed, and the reference point in this frame. This lost path remains in memory for m frames and if this element didn't come back to trajectories in the following *m* sequences, the memory would be discharged.

As described before, memory element usage occurs when a column with all zero elements appears in the affinity matrix and simultaneously, the related detection satisfies the conditions for allocation to one of the previous lost paths. Using this method, identified pathways corresponding to missing targets can be continuously recovered.

4. Experiments

In this section, the experimental results extracted from the proposed method for multi-object detection and tracking on different real-world datasets are represented. There is a concise explanation of applied datasets and evaluation criteria in Section 4.1. It is followed by the evaluation of the research parameters and their impact on precision and accuracy in Section 4.2. The parameters include: 1) the number of effective frames in prediction of next probable area in temporal model, 2) the number of frames stored in memory and 3) the number of frames in lost detection retrieval. Section 4.3 depicts the detector performance, the full model and all separated parts separately. Section 4.4 represents the results achieved applying scale factor and also different Dynamic Filters on MOT model and Section 4.5 is dedicated to experimentally compare the proposed approach with other recent MOT methods.

For single-frame pedestrian detection, we first apply the Gaussian Mixture Model for background subtraction by which the moving objects are segmented from the background. Afterwards, detectors trained with the *Inria* dataset are applied. In this dataset, the training images contain persons with a height of 96 pixels. Using a scale factor of *two* in HOG / SVM classifier, we succeed in detection of people appearing in the height of 48 pixels and larger. For the ACF detector, we applied *Inria* model static camera datasets and *caltech* model, with high diversity in detected pedestrian dimension, for dynamic one. Number of scales per octave is considered 8 and no maximum suppression is done according to the rectangular scores. The score demonstrates the confidence of the detector yielded by a cascade classifier and receives the magnitudes of minus infinity to plus infinity.

For the next probable areas, we define "search area" of 75 pixels (for dataset AVG-TownCentre) in two directions from the current center for each detected goal. To tune scale factor, four values of [1, 1.2, 1.5 and 2] are experimentally extracted. The scene is divided in four in longitudinal direction and each part's dimension is divided by values of scale factor respectively upwards. In other words, the highest section of the image (exposing the greatest distance from the camera), is divided by 2 to be analogous with the nearest section divided by 1. A distance threshold of 700 pixels and appearance threshold of 13000 units in association algorithm are defined. Visibility of a detection is *True* if it is detected for 3 frames uninterrupted. Image edges threshold is also experimentally determined according to the dataset.

4.1. Datasets and evaluation criteria

Our proposed method is evaluated on image sequences of both static camera set-up (the PETS 2009 campus image sequences and the AVG-TownCentre sequence), and moving platform (the ETHZ Bahnhof sequence). Following the work done by Klinger et al. (2017), the evaluation metrics used in the MOT benchmark are applied; (1) MOTA: Multi Object Tracking Accuracy, (2) MOTP: Multi Object Tracking Precision proposed by Bernardin and Stiefelhagen (2008), (3) FPPI: the number of false positives per image, (4) MT and (5) ML: the ratio of mostly tracked and mostly lost objects, respectively, (6) FP: the numbers of false positive and (7) FN: false negative detections, (8) IDs: the number of identity switches and (9) Frag.: the number of interruptions during the tracking of a person. MT happens for a pedestrian if it is tracked at least 80% of its presence and ML happens if the person is tracked at most 20% of its presence in sequentially images). The MOTA metric is a combined measure of tracking error (FPs, FNs and IDs.), in the range of minus infinity to one (no error). Tracking Precision metric demonstrates the geometric accuracy of a detected goal. The method used in all sequences is the same and the only difference is related to the detection and threshold stage. Basically, the application of background subtraction on datasets with moving cameras isn’t common; therefore, this step has been removed in ETHZ Bahnhof.

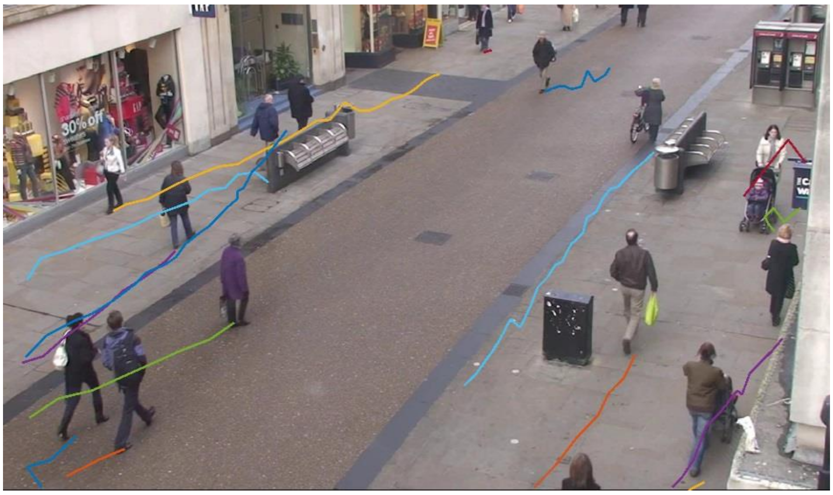

Fig 13: Tracking people in dataset AVG-Town Centre.

4.2. Research parameters investigation and impact

At this stage, the goal is to find the best values for the parameters defined in the algorithm. Experiments have been performed on the sequence of AVG-TownCentre images. This sequence of images has the widest distribution of pedestrians in the scene in terms of the number of people present in each frame simultaneously, occlusions, and the length of the route traveled by each person. In addition, these parameters do not depend on the position of the camera and even its movement; therefore, the resulting parameters can be used for all data sets studied.

Following the study conducted by Klinger et al. (2017), keeping the other parameters constant, the accuracy and precision results for each parameter were extracted and a decision was made based on the average of these two values.

$$\hat{p} = {}^{\arg max}{}_{p} S(p) \tag{9}$$

$$S(p) = (MOTA + MOTP)/2. \tag{10}$$

where p represents the algorithm parameters. These experiments have been reviewed for 50 frames. The results show situations where no recursive filter has been used. In each figure, the highest point of the mean indicates the optimal value for the parameter under consideration.

Memory

A lost detection in the middle of the scene (after it has got "Predicted" label for *m* frames), is stored in memory up to *n* frames. Then, the goal is removed or re-discovered on the scene. Memory helps to represent unidentified paths in the current frame. Values 1 to 11 for the frames were tested. Investigation and comparisons of both parameters of accuracy and precision (average) showed the value of 9 as the optimal value. Fig. 14 shows the values examined.

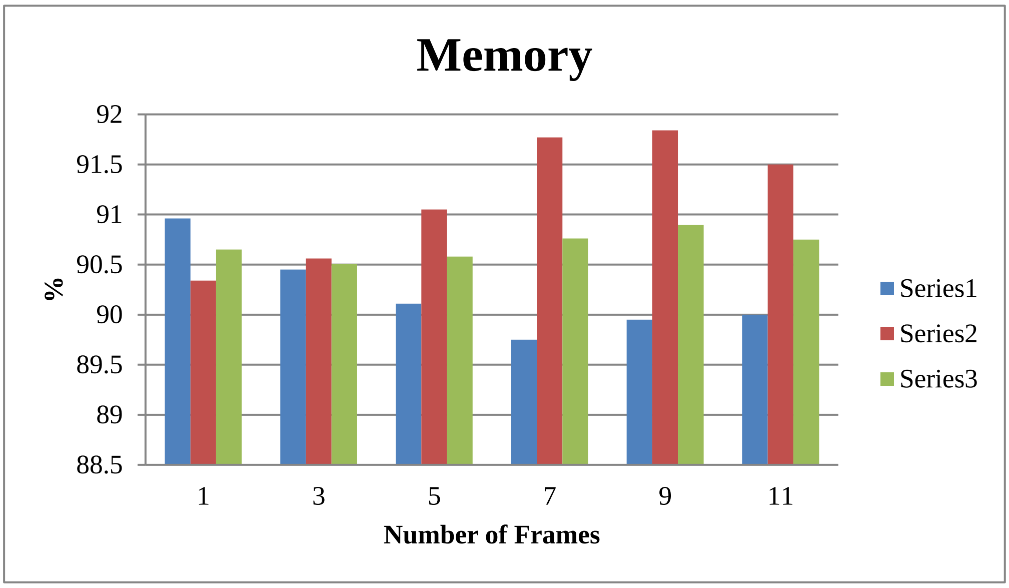

Fig 14: Parameter of Number of frames stored in memory and difference of the average of MOTA and MOTP metrics.

Number of effective frames in prediction of next probable area

Here, our goal is to investigate the number of effective frames for predicting next probable centers. Values 1 to 9 of the frames were examined. Investigations and comparisons of accuracy and precision (average) parameters showed the number of 7 frames as the optimal value. Fig. 15 shows the results of this study.

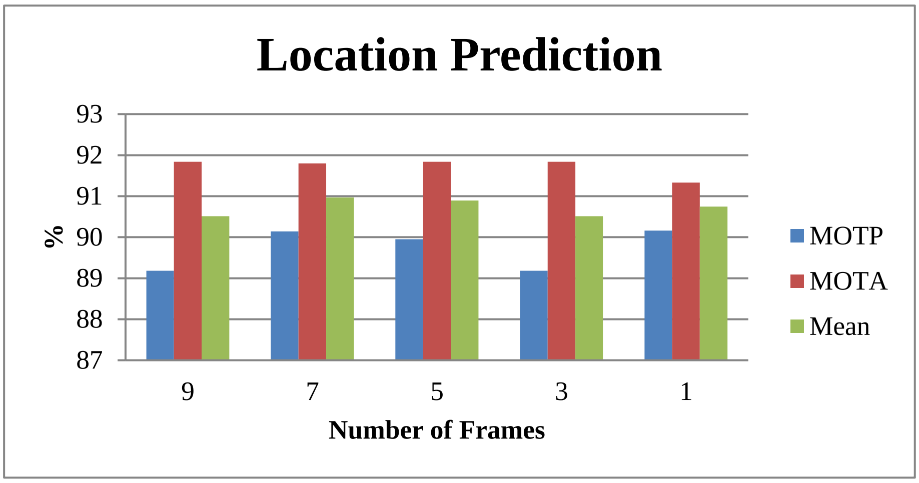

Fig 15: Parameter of Number of effective frames in prediction of next probable area in temporal model and difference of the average of MOTA and MOTP metrics.

Number of effective frames in prediction and retrieval of lost detection

A lost detection in the middle of the scene is generated and labeled as "Predicted" at the predicted next center and tracked up to *n* frames. Here, our goal is to investigate this parameter. To evaluate this parameter, values of 1 to 7 frames were examined. Investigations and comparisons of both accuracy and precision (average) parameters showed the number of 5 frames as the optimal value. Fig. 16 shows the results of this study.

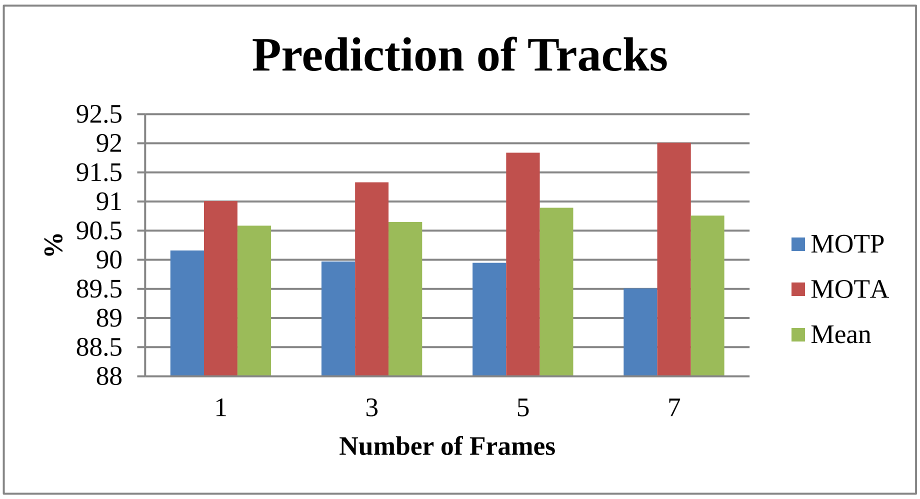

Fig 16: Parameter of Number of frames in lost detection retrieval and difference of the average of MOTA and MOTP metrics.

4.3. Detector performance

In order to evaluate the proposed detector, HOG and ACF methods, their combination with Background Subtraction, and full model (combination of these two detectors used after applying background subtraction) are separately investigated. Results are presented in table 1.

Table 1: Evaluation of detectors.

Method	FN	FP	FPPI	TP	Recall	MOTP
HOG	599	233	4.66	721	0.5462	62.5187
ACF	613	517	10.3400	722	0.5408	59.0244
HOG & BS	323	**4**	**0.08**	719	0.6900	**95.1411**
ACF & BS	169	14	0.2800	723	0.8105	91.2386
Proposed Method	**122**	18	0.3600	**732**	.**08571**	91.9424

As can be seen, false negatives mostly belong to the ACF method. Using HOG always reveals fewer goals than ACF. In ACF, however, many areas are misidentified as targets. Figures 5 and 6 also confirm the reported values. As it was expected, applying background subtraction technique diminishes the number of wrong detections significantly. On the other hand, following this method, the number of false negative also decreases. The highest amount of precision happens for the combination of HOG and BS. The result is dependent on the boundary rectangles. As it is clear in extracted figures, boundaries drawn out of different methods are localized differently in different sizes. The proposed method doesn’t yield the least number of wrong detections. However, it provides the least false negatives, the most correct detections and therefore the most recall.

4.4. Evaluation of the proposed model and investigation of different types of recursive filters used

Specifying parameters by which the optimal results for the training dataset are delivered, now, we work toward demonstrating experimentally the efficiency of Filters application. Experiments are executed on the AVG-TownCentre sequence. The results of applying Kalman, and Unscented Kalman filters in the proposed algorithm and also non-filtered mode on the AVG-Town Center data collection are given in Tables 2. Evaluation criteria are described in section 4.1.

The best results for accuracy and precision belong to the Unscented filter, which selects the minimum set of sample points (known as sigma points) around the mean. Sigma points are then distributed along nonlinear functions, from which a new mean and covariance are generated. In fact, this filter detects the small effects of non-linear pathways and has good results, especially in the case of incorrect positives, which has affected the accuracy of the results.

Table 2: The results of the proposed method in combination with Recursive filters.

Method	MOTA	MOTP	FPPI	MT (%)	ML (%)	FP	FN	IDs	Frg.
Non- Filtered	91.98	**89.32**	0.34	97.67	0	154	**160**	**140**	90
Kalman Filter	92.15	88.87	0.36	95.34	0	164	190	**140**	**74**
Unscented Kalman Filter	**92.60**	**89.32**	**0.32**	95	0	**144**	192	174	84

Comparison of the results shows that the use of filters does not have much effect on the accuracy rate (less than 1.7% difference between the maximum and minimum values). We think the reason for this is the closeness of the values estimated by the time model with the values predicted by the filters. Also, filters do not play a role in the dimensions of boundary rectangles, which have an important effect on measuring accuracy.

There is an investigation on effects of scale factor in Table 3. The comparison is made on the non-filtered model. As it was expected, the Scale Factor has an important influence on multi-object tracking accuracy. Removing the scale factor results a noticeable increase in the number of false positives and negatives. Scale factor affected the algorithm in two scopes; in the acceptable size of pedestrians and also in definition of the next probable area of existence.

Table 3: The results of the proposed method in combination with Recursive filters.

Method	MOTA	MOTP	FPPI	MT (%)	ML (%)	FP	FN	IDs	Frg.
Without Scale Factor	85.70	89.07	0.51	**98**	0	232	413	164	45
With Scale Factor	**91.98**	**89.32**	**0.34**	97.67	0	**154**	**160**	**140**	90

Wrong detections smaller than reasonable dimensions near the camera or larger than reasonable dimensions far from the camera are common. So, considering acceptable size of pedestrians and applying scale factor to control these sizes, handles the number of FPs. On the other hand, there is a logical area for goals appearance around their current positions; however, this area isn't the same through the scene. Therefore, applying scale factor in definition of the next probable area of existence has an important role in reducing the number of false negatives.

4.5. Full Model Evaluation

Considering the parameters that yield optimal results for the training dataset, now, our aim is now to investigate the performance of the full proposed model empirically. Also, we compare our method with some state-of-the-art methods. Experiments are conducted on the AVG-TownCentre, PETS09_S2L1 and ETHZ Bahnhof sequences. Table 4 shows the results of the comparisons. The results aren't available on all data sets for all compared methods; so, only the available results for each method are reported. In addition, due to the different depths of images in the data, differences in the number of pedestrian's diversity per data, and different overlap between datasets, the results aren't analogous. Keeping inaccessibility of time-tested resources in mind, and hardware differences that do not provide comparable results, the comparison in running-time of methods is ignored.

Table 4: The results of the proposed method in comparison with previous studies. Best values are printed 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**

We report the results of our proposed method on AVG-TownCentre dataset in comparison with studies conducted by Klinger et al. (2017), Liao et al. (2018), Leal-Taixé et al. (2011) and Pellegrini et al. (2009). We further report these results for the image sequence of PETS 2009 comparing studies by Ren et al. (2018), Yang et al. (2019), Yang et al. (2018) and Liao et al. (2018). For ETHZ Bahnhof data with a moving camera, the proposed method was compared with the work of Klinger et al. (2017) and Yoon et al. (2015). The results shown in Tables 4 reflect the benefits of using the full model.

Leal-Taixé et al. (2011) proposed a method of global optimization on the whole sequence of images. Pellegrini et al. (2009) worked on measuring the predicted speed using minimizing energy function. Study done by Ren et al. (2018) is based on a social force model in which each person's position in the next prediction frame and the application of the Hongarin algorithm on the weighted interval matrix between individuals. Yang et al. (2019) worked based on merging two-stage information with an improved sparse dependency model based on appearance features and a rank-based motion-based dependency model. Yang et al. (2018) studied on the generalization of a single-track detector to a multi-target, using the visual features and decision process of Markov. Other mentioned methods in comparison are explained earlier.

As it can be seen, the number of false positives for AVG-TownCentre is significantly less than its same value in other methods (155 vs. 640 and the other values greater than 1000). This value is in the second place for PETS09_S2L1. In motion camera data, this value is again lower than the other available false positives with a considerable difference (155 vs. 923). We think this false positive reduction is the results of using a combined detector, the background subtraction algorithm, as well as the application of thresholds based on scene information.

The number of false negatives in our approach for both stationary camera datasets is the lowest among the other methods studied (660 vs. 2820 and other values greater than 4000 and 14 vs. 183); for data on moving platform, the number of incorrectly detected goals is slightly less than number of false negatives among the reported magnitudes. In the proposed approach, the number of false negatives can be controlled by tuning the defined memory and also the prediction of lost detections (generated by the detector).

The numbers of identity switches and misidentification of individuals in our proposed method don't acquire the lowest value for any of the investigated data, but it is always close to the lowest value. The exception is seen in PETS09_S2L1, in which the occlusion and prolonged pauses among targets with similar appearances is clearly visible simultaneously.

The number of interruptions in our proposed method is the lowest for both data with fixed camera (7 in 19 and 90 compared to 184); although this magnitude isn't the lowest value for ETHZ Bahnhof, it doesn't differ much from the minimum value. The reason for this reduction in comparison with previous research methods is again the memory application and the prediction of lost discoveries (resulted by the detector). Both of which help to achieve these goals and prevent interruptions in individuals tracking; In such a way that, if these cases occur, they are predicted on the stage until the targets are rediscovered.

Accuracy of AVG-TownCentre and PETS09_S2L1, with 87.74% and 94.85%, respectively, was the highest compared to other research methods. The obtained values, especially in AVG-TownCentre, with 75.75% as the highest result gained by Liao and his colleagues, are significantly good. This is reported to be 66.86% for data with moving camera, which is 17% lower than the highest value in the Yoon's study. The results in this image sequence indicate the weakness of the adjusted threshold values in this set.

In all cases investigated for all image sequences, geometric accuracy or precision in the proposed approach allocated the highest value to itself, with an average difference of

15% to 30% for all three datasets. Explaining this issue, we can mention the temporal model presented in combination with recursive filters in two steps of prediction and updating; even in cases where the goals have been recovered, the best forecast of the next possible situation has been made.

5. Conclusion

We have presented an online probabilistic approach for localization and tracking of multi-person in image sequences. The main basis of the proposed model is the affinity matrix defined between the discoveries in the current frame and the available trajectories. The matrix is generated at each frame after detection and is completed during the next steps. The detector used is a combination of three conventional detection methods that results in a significant reduction in the number of false and missed detections. A simple temporal model is defined based on the speed and location of current frames. Short-term memory is used to recover lost data from the detector. The threshold values for each data are determined by the position of the camera and the three-dimensional perspective properties of images. Adjustable algorithm parameters were checked, and optimal values were determined. The introduced method is analogous to the previous state of the art studies, and perfect results have been obtained in the field of accuracy and precision. The results showed that the proposed approach could be used in both data series with fixed and moving cameras and applying an affinity matrix with defined memory reduced error due to identity changes and interruptions to a large extent. However, there are still some challenges in these fields.

False negatives still occur; this happens due to the detector weakness (more specifically in AVG-TownCentre, targets revealing on the edge of the scene is the most difficult). For future studies, improving detection using an approach similar to that used in HOG / SVM, but by features in optical form is proposed. In future work, the tracking errors of identity switches will be addressed by more precise bounding rectangles. It must be profitable to add boundary sizes to the unknowns of dynamic filters to handle the results of detectors and solve the data association problem more efficiently. Also, the definition of their central point instead of the point at the bottom may improve the precision. The distance and relative movement of pedestrians is more information that can help to this end. Moreover, using 3D features and additional information-based thresholds, especially for the mobile data set, performance improvement may also be achievable.

References

Attique Khan, M.; Mittal, M.; Mohan Goyal, L.; Roy, S.; “A deep survey on supervised learning based human detection and activity classification methods”, Multimedia Tools and Applications, 80:27867–27923, 2021.

Bernardin, K.; Stiefelhagen, R.; “Evaluating Multiple Object Tracking Performance: The clear MOT Metrics,” EURASIP Journal on Image and Video Processing. Vol. 2008, Article Id 246309, pp. 1–10, 2008.

Brunetti, A.; Buongiorno, D.; Trotta, G. F.; Bevilacqu, V.; “Computer vision and deep learning techniques for pedestrian detection and tracking: A survey”, Neurocomputing, 300:17–33, 2018.

Choi, W.; Pantofaru, C.; Savarese, S.; “A general framework for tracking multiple people from a moving camera,” IEEE Trans. Pattern Anal. Mach. Intell., 35 (7), pp. 1577–1591, 2013.

Dollar, P.; R. Appel, S. Belongie, et al, “Fast feature pyramids for object detection”, IEEE Trans. Pattern Anal. Mach. Intell., 36, (8), pp. 1532–1545, 2014.

Ellis, D.; Sommerlade, E.; Reid, I.; “Modelling pedestrian trajectory patterns with gaussian processes,” IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1229–1234, 2009.

Felsberg, J.; Larsson, F.; “Learning Bayesian Tracking for Motion Estimation,” The 1st International Workshop on Machine Learning for Vision-based Motion Analysis, 2008.

Fortmann, T.E.;Bar-Shalom, Y.; Scheffe, M.; “Sonar tracking of multiple targets using joint probabilistic data association,” IEEE J. Oceanic Eng. 8 (3), pp. 173–184, 1983.

Hernández-Aceituno, J.; Acosta, L.; D. Piñeiro, J.; “Pedestrian Detection in Crowded Environments through Bayesian Prediction of Sequential Probability Matrices,” Journal of Sensors, 4697260, 2016.

Hoiem, D.; Efros, A.A.; Hebert, M.; “Putting objects in perspective,” Int. J. Comput. Vision 80 (1), pp. 3–15. 2008.

Kim, K.; Lee, D.; Essa, I.; “Detecting regions of interest in dynamic scenes with camera motions,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1258–1265, 2012.

Klinger, T.; Rottensteiner, F.; Heipke, C.; “A dynamic Bayes Network for visual pedestrian tracking”. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives 40 (2014), Nr. 3, pp. 145–150, 2014.

Klinger, T.; Rottensteiner, F.; Heipke, C.; “Probabilistic multi-person tracking using dynamic bayes networks,” ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. II-3/W5, pp. 435–442, 2015.

Klinger, T.; Rottensteiner, F.; Heipke, C.; “Probabilistic multi-person localisation and tracking in image sequences,” ISPRS Journal of Photogrammetry and Remote Sensing, no. 127, pp. 73–88, 2017.

Ko, B. C.; Jeong, M.; Nam, J.; “Fast Human Detection for Intelligent Monitoring Using Surveillance Visible Sensors,” Journal of Sensors, no. 14, 2014.

Leal-Taixé, L.; Fenzi, M.; Kuznetsova, A.; Rosenhahn, B.; Savarese, S.; “Learning an image-based motion context for multiple people tracking,”IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3542–3549, 2014.

Leal-Taixé, L.; Pons-Moll, G.; Rosenhahn, B.; “Everybody needs somebody: modeling social and grouping behavior on a linear programming multiple people tracker”, IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 120–127, 2011.

Li, W.; Mu, J.; Liu, G.; “Multiple Object Tracking with Motion and Appearance Cues,” IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019.

Liao, M.; Xiao, G.; “Online Multi person tracking based on sparse representation”, IEEE International Conference on Progress in Informatics and Computing (PIC), 2018.

Milan, S.; Roth, K.; Schindler, K.; “Continuous energy minimization for multitarget tracking,” IEEE Trans. Pattern Anal. Mach. Intell. 36 (1), pp. 58–72, 2014.

Ning, C.; Menglu, L.; Hao, Y.; Xueping, S.; Yunhong, L.: “Survey of pedestrian detection with occlusion; Complex & Intelligent Systems pp. 1/11, 2020.

Pellegrini, S.; Ess, A.; Schindler, K.; Van Gool, L.; “You’ll never walk alone: Modeling social behavior for multi-target tracking”, IEEE International Conference on Computer Vision (ICCV), pp. 261–268, 2009.

Pervaiz, M.; Jalal, A.; and Kim, K.; "Hybrid Algorithm for Multi People Counting and Tracking for Smart Surveillance," 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST), pp. 530-535, 2021.

Rahmaniar, W.; Hernawan, A.; “Real-Time Human Detection Using Deep Learning on Embedded Platforms: A Review”, Journal of Robot. Control, JRC 2021, 2, 462–468, 2021.

Ren, H.; Xu, F.; Zou, F.; Jia, K.; Di, P.; Kang, J.; “Multi-pedestrian Tracking Based on Social Forces”, IEEE International Conference on Intelligence and Safety for Robotics, Shenyang, China, August 24–27, 2018.

Rudenko, A.; L. Palmieri, A. J. Lilienthal and K. O. Arras, "Human Motion Prediction Under Social Grouping Constraints," 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, 2018, pp. 3358-3364, 2018.

Schindler, K.; Ess, A.; Leibe, B.; Van Gool, L.; “Automatic detection and tracking of pedestrians from a moving stereo rig,” ISPRS J. Photogramm. Remote Sensing 65 (6), pp. 523–537. 2010.

Stadler, D.; Beyerer, J.; “Modelling Ambiguous Assignments for Multi-Person Tracking in Crowds,” 2022 IEEE/CVF Winter Conference on Application of Computer Vision Workshop (WACVW), pp. 133-142, 2022.

Stauffer, C.; Grimson, W.; “Learning patterns of activity using real-time tracking,” IEEE Trans. Pattern Anal. Mach. Intell. 22 (8), pp. 747–757. 2000.

Wang, W.; Chang, X.; Yang, J.; Xu, G.; “LiDAR-Based Dense Pedestrian Detection and Tracking”, Applied Sciences, vol. 12, no. 4, 2022.

Yang, A. L.; Ren, H. Y.; Fei, M. R.; Naeem, W.; “Multi-person vision tracking approach based on human body localization features,” Advances in Manufacturing, 9(4), 496-508, 2021.

Yang, H.; Li, J.; Liu, J.; Zhang, Y.; Wu, Z.; Pei, Z.; “Multi-Pedestrian Tracking Based on Improved Two Step Data Association”, IEEE, Volume: 7, pp. 100780–100794, 2019.

Yang, T.; Cappelle, C.; Ruichek, Y.; Bagdouri, M.E.; “Online Multi-object Tracking Combining Optical Flow and Compressive Tracking in Markov Decision Process”, 2018.

Yao, S.; Wang, T.; Shen, W.; Pan, S.; Chong, Y.; Ding, F.; “Feature Selection and Pedestrian Detection Based on Sparse Representation,” Journal of Pone, 2015.

Yoon, J.H.; Yang, M.H.; Lim, J.; Yoon, K.J.; “Bayesian multi-object tracking using motion context from multiple objects,” IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 33–40, 2015.

Yu, J.; Kim, D.; Ku, B.; Ko, H.; “Online Multi-Object Tracking based on Hierarchical Association Framework,” ISPRS Journal of Photogrammetry and Remote Sensing, IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2016.

Zhang, D.; Xia, F.; Yang, Z.; Yao, L.; Zhao, W.; “Localization Technologies for Indoor Human Tracking,” Future Information Technology (FutureTech), 5th International, 2010.