Unified Hierarchical Multi-Object Tracking using Global Data Association

Martin Hofmann, Michael Haag, Gerhard Rigoll
Institute for Human-Machine Communication
Technische Universität München
martin.hofmann@tum.de, michael.haag@mytum.de, rigoll@tum.de

Abstract

This paper presents a unified hierarchical multi-object tracking scheme. The problem of simultaneously tracking multiple objects is cast as a global MAP problem which aims at maximizing the probability of trajectories given the observations in each frame. Directly solving this problem is infeasible, due to computational considerations and the difficulty of reliably estimate necessary transition probabilities. Without breaking the MAP formulation, we propose a three stage hierarchical tracking framework which makes solving the MAP feasible. In addition, using a hierarchical framework allows for modeling inter-object occlusions. Occlusion handling thus smoothly and implicitly integrates into the proposed framework without any explicit occlusion reasoning. Finally, we evaluate the proposed method on the publicly available PETS 2009 tracking data and show improvements over the current state of the art for most sequences.

1. Introduction

Simultaneously tracking multiple objects in video data is still a difficult and only partially solved problem in computer vision. Tracking single objects is conceptionally simpler. Here, the object can be detected in each frame and the trajectory arises, simply by connecting the detections over time. However, in the case of multiple objects, a major problems becomes the association of detections to trajectories. To overcome this problem, on the one hand the identity of the object can be established (i.e. by color, shape features) and on the other hand constraints on smooth motion and continuity can be applied.

A second major issue in multi object tracking are mutual occlusions. While object motion may in general be considered independent, the visual effect of multiple objects is by far not independent. Thus, objects occluding each other strongly influence their mutual visual appearance up to the case that one object is completely occluded by the other.

In our approach we simultaneously address both issues in a unified hierarchical framework. First, we formulate our tracking algorithm as a maximum a posteriori problem, which is commonly done in global tracking approaches. In order to make the MAP approach tractable, a multitude of independence assumptions have to be made. While some of these assumptions are reasonable, others are not valid in practice. Most notable, using a Markov chain in the definition of the transition probabilities would require all information of an object (appearance, motion, etc.) to be present in a single frame. However, both motion and appearance can hardly be captured in just a single frame.

To overcome these limitations, we still keep the MAP formulation and process the data in three hierarchical stages. This allows multiple frames to be aggregated at lower hierarchical levels. Thus, reliable motion and appearance information can be captured, which improves association at the higher levels.

2. Related Work

Target tracking has been studied extensively. Local tracking approaches using for example the Kalman Filter [16] have high precision and localization accuracy, but fail in multi object scenarios where association of detections and trajectories becomes a major issue. In order to cope with multiple objects, trajectories can be optimized one by one, e.g. using Dynamic programming [4, 9, 10]. This, however, largely ignores mutual influence of the trajectories. Multi-Hypothesis Tracking (MHT) [14] and Joint Probabilistic Data Association Filters (JPDAF)[7] overcome this problem by jointly optimizing trajectories, but these methods suffer from the combinatorial hypotheses space. Another class of recent and very successful approaches define the tracking as a global optimization over the complete sequence [17][13][8]. Here, a global posterior probability is formulated and maximized. While these methods are conceptionally solid and fast algorithms exist (e.g Hungarian Algorithm [11]), in order to make the method tractable many independence assumptions (as outlined in the introduction) have to be made, which is limiting in practice.

(a) Detection Input (b) Tracking Result

Figure 1: (a) detection responses in the input frame with one false positive and two misses due to occlusion, (b) final tracking result.

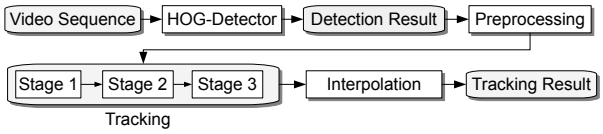

Figure 2: Schematic of the hierarchical tracking framework

3. Unified Tracking Framework

The outline of our tracking approach is depicted in Figure 2. We follow the successful tracking-by-detection paradigm. Thus, in a first step, objects of interest are detected with an off-the-shelf object detector (we use [5]). In a preprocessing step, detection responses are pruned according to geometric constraints and prior knowledge. Using the known camera calibration, many false positives (i.e. trees in the background) can be removed.

For the actual tracking algorithm we use a consistent three stage architecture which builds on a unified MAP formulation, but uses different parameter settings at each stage:

First, at the low level stage, track fragments are generated by conservatively linking detection responses that are very likely to belong to the same track. Second, at the middle level stage, the obtained track fragments are associated into longer tracks by formulating the tracking task as a MAP problem which is solved by the Hungarian algorithm. Finally, at the high level stage, the resulting tracks from the previous stage are refined and grouped into long-range tracks by simply adapting the MAP parameters. After tracking, the obtained data is interpolated leading to the final tracking result.

3.1. Maximum A Posteriori Formulation

First, we will describe the basic underlying maximum a posteriori (MAP) problem regarding the task of tracking multiple objects in a video sequence.

As presented in [17], we assume a set of object observations being detection responses from the detector. This set $\mathcal{X}$ contains various information for each detection i, so that $\mathcal{X} = \{\mathbf{x}_i\}$ with $\mathbf{x}_i = (x_i, s_i, t_i)$, where x_i is the position, s_i is the size and t_i is the time index of the object in the video sequence.

The main aim is to find appropriate trajectories based on these object observations. Hereby, a trajectory hypothesis T_k can be described as a list of observations so that $T_k = \{\mathbf{x}_{k_1}, \mathbf{x}_{k_2}, \ldots, \mathbf{x}_{k_{l_k}}\}$ with $\mathbf{x}_{k_i} \in \mathcal{X}$, whereas an association hypothesis $\mathcal{T}$ is defined as a set of all trajectory hypotheses leading to $\mathcal{T} = \{T_k\}$.

The overall tracking goal leads to the objective to maximize the posteriori probability of $\mathcal{T}$ given the set $\mathcal{X}$ of object observations, so we can formulate:

$$\mathcal{T}^* = \underset{\mathcal{T}}{\textbf{argmax}}\, P(\mathcal{T}|\mathcal{X}) \tag{1}$$

$$= \underset{\mathcal{T}}{\textbf{argmax}}\, P(\mathcal{X}|\mathcal{T})P(\mathcal{T}) \tag{2}$$

$$= \underset{\mathcal{T}}{\textbf{argmax}} \prod_i P(\mathbf{x}_i|\mathcal{T})P(\mathcal{T}) \tag{3}$$

For the last conversion of the equation, we assume that the likelihood probabilities are conditionally independent given the hypothesis $\mathcal{T}$.

Assuming that trajectories do not overlap ($\mathcal{T}_k \cap \mathcal{T}_l = \emptyset, \forall k \neq l$) and that the motion of trajectories are independent of each other, the prior $P(\mathcal{T})$ can be further factorized:

$$\mathcal{T}^* = \underset{\mathcal{T}}{\textbf{argmax}} \prod_i P(\mathbf{x}_i|\mathcal{T}) \prod_{\mathcal{T}_k \in \mathcal{T}} P(\mathcal{T}_k) \tag{4}$$

The prior of each trajectory $P(\mathcal{T}_k)$ in (4) is modeled with a Markov chain:

$$\begin{aligned} P(\mathcal{T}_k) =& P(\{\mathbf{x}_{k_0}, \mathbf{x}_{k_1}, \ldots, \mathbf{x}_{k_{l_k}}\}) \\ =& P_{entr}(\mathbf{x}_{k_0}) P_{link}(\mathbf{x}_{k_1}|\mathbf{x}_{k_0}) P_{link}(\mathbf{x}_{k_2}|\mathbf{x}_{k_1}) \\ & \ldots P_{link}(\mathbf{x}_{k_{l_k}}|\mathbf{x}_{k_{l_k-1}}) P_{exit}(\mathbf{x}_{k_{l_k}}) \end{aligned} \tag{5}$$

Using a Markov chain implies that all information is captured at a given point in time and following nodes only depend on the previous and not all other nodes. While this is a frequently made assumption, we feel that it is fundamentally flawed, because in practice, insufficient information can be captured in a single frame. The resulting probabilities P_{entr}, P_{exit} and $P_{link}(\mathbf{x}_{k_{i+1}}|\mathbf{x}_{k_i})$ are essential for the tracking algorithm and are specified in the Section 3.2.

The second factor in (4), $P(\mathbf{x}_i|\mathcal{T})$, is the likelihood function of observation $\mathbf{x}_i$:

$$P(\mathbf{x}_i|\mathcal{T}) = \begin{cases} P_{tp} & \text{if } \exists\, \mathcal{T}_k \in \mathcal{T}, x_i \in \mathcal{T}_k \\ P_{fp} & \text{otherwise} \end{cases} \tag{6}$$

with P_{tp} being the true positive and P_{fp} being false positive probability. The probabilities $P_{tp} = 1 - P_{fp}$, and $P_{tn} = 1 - P_{fn}$ (used in Equation (14)) depend on the score output s of the object detector (i.e. a high detector score leads to to high P_{tp}). The relations $P_{tp} \leftarrow s$ and $P_{tn} \leftarrow s$ are learned on annotated ground truth.

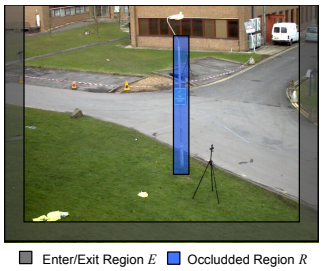

Figure 3: Illustration of the entrance/exit area E and the occlusion region R.

3.2. MAP Parameters

As shown above, there are several parameters in the above framework, which need to have meaningful values. Most important are enter and exit probabilities (P_{entr} and P_{exit}), as well as the transition probability P_{link}.

Enter and Exit Probabilities We assume that tracks can only start and end at the edge region E (as depicted in Figure 3). Except for the first (last) frame, where tracks can also start (end) independently of the position. We set the enter and exit probabilities near the edge higher than in the center.

In order to estimate these enter and exit probabilities, we assume them to be generally the same ($P_{entr} = P_{exit}$). We can now show that these parameters are related to the minimal required track length l.

Consider the probability P_a as the likelihood that a group of l linked detections is surely a real track as well as the probability P_b as the likelihood that this group only consists of false alarms. Then we get, with Equation (5) and (6):

$$P_a = P_{entr} \prod_k (P_{link} P_{tp}) P_{exit} \tag{7}$$

$$P_b = \prod_k P_{fp} = P_{fp}^l \tag{8}$$

With $P_{link} = 1$ we can write:

$$P_a = P_{entr}^2 P_{tp}^l \tag{9}$$

The MAP formulation in (4) decides for a trajectory, if $P_a > P_b$, that is if the observation probability is higher than assuming pure false positives. Therefore, when we set the two equal, we get:

$$\begin{aligned} P_a &= P_b \\ P_{entr}^2 P_{tp}^l &= P_{fp}^l \\ P_{exit} = P_{entr} &= \left(\frac{P_{fp}}{P_{tp}}\right)^{l/2} \end{aligned} \tag{10}$$

The last equation shows that the enter and exit probabilities will affect the minimal length of the obtained tracks. Vice versa, by choosing a certain minimal track length l, we can calculated P_{entr} and P_{exit}.

Transition Probability The transition probabilities P_{link} are described by the similarity of the linked detection responses. We formulate three independent aspects of similarity: Appearance, frame skip and motion similarity. Thus, we formulate the transition probability as follows:

$$P_{link}(\mathbf{x}_j|\mathbf{x}_i) = P(a_j|a_i)P(\Delta t)P(v_j|v_i) \tag{11}$$

Appearance Model The appearance term $P(a_j|a_i)$ is based on RGB histograms a_j and a_i. The similarity in appearance is defined based on the Bhattacharyya distance $A_{ij} = \sum_{i=1}^{n} \sqrt{(\sum a_i \cdot \sum a_j)}$:

$$P(a_j|a_i) = \frac{\mathcal{N}(A_{ij}; A_s, \sigma_s^2)}{\mathcal{N}(A_{ij}; A_s, \sigma_s^2) + \mathcal{N}(A_{ij}; A_d, \sigma_d^2)} \tag{12}$$

with $\mathcal{N}(x; A_s, \sigma_s^2)$ and $\mathcal{N}(x; A_d, \sigma_d^2)$ being the normal distributions of A_{ij} between the same object and different objects respectively. The parameters $(A_s, \sigma_s, A_d, \sigma_d)$ are learned from annotated ground truth.

Frame Skip The term $P(\Delta t)$ models the frame skip. This allows us to handle miss detections. In addition, this term is used to model occlusions caused by known stationary objects (e.g. the light pole in PETS2009) as well as by dynamically moving objects.

Since the tracklets can also be linked over non-consecutive frames which leads to a certain frame gap Δt, with F being the set of skipped frames, we model this as a time gap component. It is defined by an exponential model as follows:

$$P(\Delta t) = \prod_{t \in F} P(t) \tag{13}$$

$$\text{and } P(t) = \begin{cases} P_{tn} & \text{if } x_i' \in R \\ P_{fn} & \text{otherwise} \end{cases} \tag{14}$$

Here, R is the occlusion region which is defined according to static and dynamic objects in the scene.

Motion Model The third term of the transition probability considers the motion similarity of two tracklets. Let v_j and v_i be the normalized movement vectors calculated as the mean of the frame wise movement of detection responses within the tracklets. Then we get:

$$P(v_j|v_i) = 1 - \frac{1}{2}\|v_j - v_i\|_1 \tag{15}$$

In consequence, this term measures the similarity of the track movement assuming a person is not likely to change its direction abruptly.

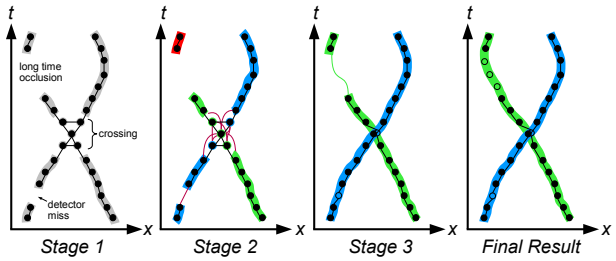

Figure 4: Exemplary illustration of the hierarchical tracking strategy. Stage 1: build small tracklets and direct links based on spatial overlap; Stage 2: use frame skip (here set to 1) to handle small occlusions and misses; Stage 3: handle longtime occlusions

4. Hierarchical Tracking Strategy

For our multi-target tracking algorithm we pursue a hierarchical bottom-up strategy. During the whole tracking progress the underlying MAP framework stays consistent, solely the parameter settings are adjusted stepwise.

We divide the tracking task into different stages resulting into iteratively growing tracks that will finally lead to a set of target trajectories. Hereby, each tracking estimate achieved by a stage is used as starting point of the following stage. An exemplary illustration of our stage wise tracking progress is depicted in Figure 4.

4.1. Stage 1: Tracklet Generation

In the first stage, we conservatively group a set of detections out of consecutive frames that fulfill the following three conditions: (1) there is exactly one detection per frame in a set, (2) every detection is connected with exactly one adjacent window, (3) the distance in appearance passes a certain threshold.

These conditions can be cast into the MAP formulation, i.e. $P_{entr} = P_{exit} = 1$ and $P_{link} = 1$, if detections have significant overlap to exactly one predecessor and exactly one successor, otherwise $P_{link} = 0$. To optimize this MAP formulation, simple heuristics can be used, rather than using complex optimization techniques.

4.2. Stage 2: Mid-Range

The input to the second stage are the tracklets generated in the first stage. Therefore, each observation $\mathbf{x}_k$ now consists of multiple detections from multiple adjacent frames. Thus, multiple detections (which would be treated independently without hierarchy) can be jointly analyzed.

Entrance and Exit Probability For the minimal length of a track, we choose a weak prior and set $l = 12$. This will fully describe the probabilities P_{enter} and P_{exit} (using (10)). By choosing this minimal length, single (false) detections will be ignored and the tracklet size will be in the mid-range. For the region of entrance and exit, $\mathbf{x}_k \in E$ (see Figure 3), we set $l = 2$ (also for stage 3) in order to allow shorter tracks at these regions. If this distinction of cases is not made, tracks will not end or start in the entrance area leading to errors (like identity switches between persons).

Transition Probability Regarding the transition probability P_{link}, we have to adapt the equation, since the length of the input tracklets are not long enough to reliably estimate a movement vector. Thus, we ignore the movement term ($P(v_j|v_i) = 1$) and use the given information, namely appearance and frame skip, leading to:

$$P_{link,stage_2}(\mathbf{x}_j|\mathbf{x}_i) = P(a_j|a_i)P(\Delta t) \tag{16}$$

with $P(a_j|a_i)$ describing the histogram distances and $P(\Delta t)$ being defined by a fixed occlusion region R as described in 3.2.

Frame Skip To be able to cope with missed detections and occlusions, we now allow all tracklets to be linked to other tracklets from non-consecutive frames. Unlike in the first stage, the transitions can now skip a certain number of frames $\Delta t \leq \Delta t_{max}$ by up to Δt_{max} frames.

4.3. Stage 3: Long-Range

Subsequently, the obtained tracking result is used to create a new hierarchical stage. Here the tracklets are built based on the ID that the tracking process in the previous stage has calculated.

Entrance and Exit Probability Since we now refine the tracklets and search for longer trajectories, we use a stronger prior regarding the probabilities P_{enter} and P_{exit}. Hence, we set the minimal length for a track much higher. By choosing an increased length, we force the tracker to build up tracklets that will be in the long-range.

Transition Probability As the input tracklets lie now in the mid-range length, we can use the information of the movements in addition. Hence, we use the full Equation (11):

$$P_{link,stage_3}(\mathbf{x}_j|\mathbf{x}_i) = P(a_j|a_i)P(\Delta t)P(v_j|v_i) \tag{17}$$

Occluded Region In the third stage, we treat all obtained tracklets (from the second stage) as 'static' objects and therefore as possible occluders. Thus, we add the space which a tracklet requires to the occlusion area R. This has an influence on the frame skip term $P(\Delta t)$ in Equation (14). Now, a frame skip between two tracklets becomes more likely, if other (already found) objects are the reason for the occlusion.

Frame Skip In order to handle long time occlusions, we allow much higher frame skips and set Δt_{max} to about 70. By doing so, occlusions that are present over various frames can be solved.

4.4. Optimization and Post-Processing

In all three stages we defined a MAP problem. In the first stage, where the structure is very simple, the solution can be achieved with heuristics. In stage two and three, we use the Hungarian Algorithm [11] to solve the MAP problem.

After all three stages, the obtained tracking result is interpolated in two ways: On the one hand, all gaps within a track are linearly interpolated (these gaps can occur if frame skips are allowed), on the other hand the tracks are extended at the beginning and ending point in case some detections were lost at the start and end of a trajectory. Finally, the output is smoothed using a moving average filter of span 15 pixels for position and span 20 pixels for height/width.

5. Results and Evaluation

We apply our hierarchical association framework to the multiple pedestrian tracking problem. In this section, we will present our results using various datasets and compare our algorithm with other state-of-the-art systems.

5.1. Performance Metrics for Multiple Object Tracking

We use the widespread measures introduced in [15] called Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP), that became the recent de facto standard. Additionally, we apply further metrics that are presented in [12]. These are Identity Switches (IDS), Track Fragments (FM), Mostly Tracked (MT), Partly Tracked (PT) and Mostly Lost (ML), Precision and Recall. A summary and short description of the used measures is given in Table 1.

The conception of these tracking metrics allows the judgment of precision, the capacity for tracking an object consistently over time, as well as various configuration errors made by the tracker, e.g. false positives, misses or mismatches.

5.2. Datasets

The evaluation of the proposed algorithm is carried out on various video sequences from publicly available datasets.

We use two sequences from the PETS 2009 benchmark dataset [6] for our experiments. For both sequences, the dataset provides views from multiple camera angles, however, we only use one camera (View 1). The first sequence (S2L1, 795 frames total) shows an outdoor scene with numerous pedestrians which occlude each other various times.

Table 1: Evaluation metrics (according to [12] and [15])

Name	Definition
MOTA	Multiple Object Tracking Accuracy: It combines all error types and is normalized with the total No. of targets. The higher the better.
MOTP	Multiple Object Tracking Precision: The normalized distance between the objects and tracker hypotheses. The higher the better.
IDS	ID Switches: Number of times that a tracked trajectory switches its matched ground truth identity. The smaller the better.
FM	Fragments: Number of times that a ground truth trajectory is interrupted. The smaller the better.
MT	Mostly tracked: Percentage of ground truth trajectories that are tracked for more than 80% in length. The higher the better.
ML	Mostly lost: Percentage of ground truth trajectories that are tracked for less than 20% in length. The smaller the better.
PT	Partially tracked: 100% - MT - ML.
Recall	correctly matched objects / total ground truth objects
Precision	correctly matched objects / total output objects

The second sequence (S2L2, 436 frames total) shows a denser crowd, making tracking significantly more challenging.

In addition, we use the dataset presented in [1], called 'TUD-Stadtmitte'. It consists of 170 frames and depicts a busy pedestrian zone in a city center. Here, the camera has a very low view point. Generally, the occlusions in this scene are more difficult to handle since a person standing in the foreground can occlude multiple individuals standing behind it for a longer period in time.

5.3. Operational Settings

The most critical stage of our multi stage tracking system is the second one. Here, the two main components are the control of the outcoming track length and the creation of links that are non-consecutive allowing frame skips. For each component, there is a parameter that influences the tracking efficiency. Both of these parameters will be evaluated in the following.

Frame Skip Regarding the non-consecutive linking, the maximum number of allowed frame skips Δt_{max} (in stage 2) is a crucial factor for performance. In order to demonstrate the effect of the parameter Δt_{max} on the tracking result, we show in Figure 5 the system performance in dependence of the frame skip Δt_{max} (on PETS 2009 S2L1).

As can be seen, there is a very high increase in accuracy at the beginning, since here the tracking system can overcome missing detection either from misses owing to errors of the detector in use or from short inter-object occlusions.

As the frame skip is increased further, the accuracy increases as well (the long time occlusion can now be solved properly) until it reaches a certain saturation at roughly

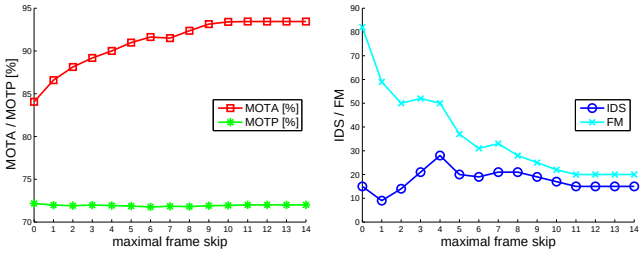

Figure 5: Influence of maximally allowed frame skip Δt_{max} on MOTA, MOTP, IDS and FM. (PETS 2009 S2L1)

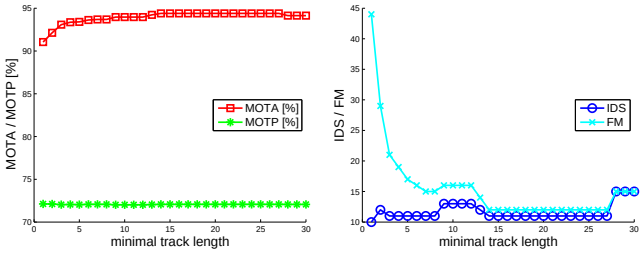

Figure 6: Influence of the minimal track length l on MOTA, MOTP, IDS and FM. (PETS 2009 S2L1).

$\Delta t_{max} = 10$. Hence, as a default and for our other experiments, we set $\Delta t_{max} = 12$.

Minimal Track Length The second main component of the second stage is the minimal tracklet length l.

To show the influence of this parameter, we plot (in Figure 6) MOTA, MOTP, the numbers of ID switches and fragmentations in relation to the chosen minimal track length (on PETS 2009 S2L1).

When increasing the minimal track length, fragmentations are significantly decreased and MOTA is significantly increased, while MOTP and ID switches stay almost constant. This is, because for small minimal track length l, only those detections will be connected, which are very likely to belong together, leading to high number of fragmentation. With increased l, longer tracks are encouraged, so bad detections, misses and occlusions can be overcome. With l being in the range of 15 to 25, the number of fragmentations stays nearly the same and an optimum is found.

If l is increased further, fragmentation and misses increase. If l is set too high, fragmentation will increase again, since then the tracker seeks very long tracks and thus, tracklets that are not likely to belong together are combined.

5.4. Overall Results

Table 2 presents quantitative results of our approach on all datasets. We show results of stage 2 (a) and stage 3 (b) including interpolation, but without smoothing. The final output (c) also includes smoothing. As expected, we see a distinguishable performance gain of the third stage over the second stage. Since the track length increases, consequently, the number of fragments decreases. In most cases, the number of ID switches stays the same, because in the third stage, tracklets are only merged (which cannot solve false IDs from the previous stage). Smoothing the final trajectories gives additional performance gain: Tracking Precision (MOTP) is increased, while fragments and ID switches are decreased. This is due to the fact that ground truth trajectories have also been smoothed. Thus, smoothing makes our output more similar to the ground truth.

Figure 7: Tracking results on PETS 2009 S2L1

On scenarios with a medium dense crowd (like PETS-S2L1 and S3MF1), our approach yields tracking results with a significantly high accuracy and a very small rate of fragmentations.

In comparison to the work in [2][3] our tracking system achieves slightly better results in sequences with sparse or medium density. Contrastingly, our approach leads to moderately smaller accuracy in scenes showing high density crowds.

6. Conclusion and Outlook

We presented a structured method for multi object tracking. The problem was first defined as a typical maximum a posteriori problem, which was made tractable by several independence assumptions. The (somewhat flawed) Markov assumption was also used in our approach, however, using a hierarchical processing, the downside of this assumption (i.e. track fragments and id switches) could be overcome. We showed the relation of enter and exit probabilities to the expected trajectory length. This new parameter, together with the maximal allowable frame skip were the major parameters in our hierarchical formulation. We evaluate both parameters over a substantial range and reason for the optimal operating point. In the evaluation, it can be seen that our approach achieves state of the art results. In many scenarios the hierarchical approach leads to a performance increase. Especially the number of track fragments and the number of ID switches could be decreased significantly.

Sequence	Method	MOTA [%]	MOTP [%]	MT [%]	PT [%]	ML [%]	FM	IDS	Rec. [%]	Prec. [%]
PETS S2L1	[2]	88.3	75.7	86.96	4.35	8.70	-	-	-	-
	[3]	81.4	**76.1**	82.61	17.39	**0**	21	15	-	-
	(a)	95.02	72.33	95.65	4.35	**0**	15	10	97.35	98.64
	(b)	96.54	72.05	**100**	0	**0**	13	10	98.47	98.66
	(c)	**97.83**	75.30	**100**	0	**0**	**8**	**8**	**99.00**	**99.14**
PETS S3MF1	[2]	96.3	**84.1**	**100**	0	**0**	-	-	-	-
	(a)	98.62	72.42	**100**	0	**0**	**0**	**0**	**100**	98.19
	(b)	98.62	72.42	**100**	0	**0**	**0**	**0**	**100**	98.19
	(c)	**99.21**	77.65	**100**	0	**0**	**0**	**0**	**100**	**99.14**
TUD Stadtmitte	[2]	68.6	64.0	55.56	0	44.44	-	-	-	-
	[3]	60.5	65.8	66.7	33.3	**0**	**4**	**7**	-	-
	(a)	61.58	64.59	60	40	**0**	29	19	81.15	83.11
	(b)	65.76	64.59	80	20	**0**	26	19	88.99	81.62
	(c)	**72.37**	**72.02**	**90**	0	10	**10**	**8**	**92.76**	**83.52**
PETS S2L2	[2]	**60.2**	**60.5**	33.33	56	**10.67**	-	-	-	-
	(a)	51.09	58.01	31.58	50	18.42	128	159	60.73	89.94
	(b)	51.98	57.50	34.21	47.37	18.42	135	164	62.00	89.76
	(c)	57.14	56.36	**39.47**	42.11	18.42	**59**	**67**	**63.83**	**92.14**

Table 2: Quantitative results of our tracking system for each dataset used. For each sequence we report results of our approach after the second stage (a), results after third stage (b) and results after final smoothing (c). Results are compared to "Global Occlusion Reasoning" [2] and "Continuous Energy Minimization"[3].

References

[1] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, 06/2010 2010. 5

[2] A. Andriyenko, S. Roth, and K. Schindler. An analytical formulation of global occlusion reasoning for multi-target tracking. In *ICCV Workshops*, pages 1839–1846. IEEE, 2011. 6, 7

[3] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. In *CVPR*, Proc. IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, 2011. 6, 7

[4] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 744 – 750, june 2006. 1

[5] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2008*, 2008. 2

[6] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. In *Winter-PETS*, pages 1–6, 2009. 5

[7] T. Fortmann, Y. Bar-Shalom, and M. Scheffe. Sonar tracking of multiple targets using joint probabilistic data association. *Oceanic Engineering, IEEE Journal of*, 8(3):173 – 184, jul 1983. 1

[8] J. F. Henriques, R. Caseiro, and J. Batista. Globally optimal solution to multi-object tracking with merged measurements. In *ICCV*, pages 2470–2477. IEEE, 2011. 1

[9] M. Hofmann, M. Kaiser, H. Aliakbarpour, and G. Rigoll. Fusion of multi-modal sensors in a voxel occupancy grid for tracking and behaviour analysis. *12th International Workshop on Image Analysis for Multimedia Interactive Services, Delft, The Netherlands*, 2011. 1

[10] M. Hofmann, M. Kaiser, N. Lehment, and G. Rigoll. Event detection in a smart home environment using viterbi filtering and graph cuts in a 3d voxel occupancy grid. *International Conference on Computer Vision Theory and Applications, Algarve, Portugal*, 2011. 1

[11] H. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 1, 5

[12] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *CVPR*, pages 2953–2960. IEEE, 2009. 5

[13] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, pages 1201 –1208, june 2011. 1

[14] D. Reid. An algorithm for tracking multiple targets. *Automatic Control, IEEE Transactions on*, 24(6):843 – 854, dec 1979. 1

[15] R. Stiefelhagen and J. S. Garofolo, editors. *Multimodal Technologies for Perception of Humans, First International Evaluation Workshop on Classification of Events, Activities and Relationships*, volume 4122 of *Lecture Notes in Computer Science*. Springer, 2007. 5

[16] G. Welch and G. Bishop. An introduction to the kalman filter. Technical report, Chapel Hill, NC, USA, 1995. 1

[17] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*. IEEE Computer Society, 2008. 1, 2