Robust Online Multi-object Tracking by Maximum a Posteriori Estimation with Sequential Trajectory Prior

Min Yang(✉), Mingtao Pei, Jiajun Shen, and Yunde Jia

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, People's Republic of China
{yangminbit,peimt,shenjiajun,jiayunde}@bit.edu.cn

Abstract. This paper address the problem of online multi-object tracking by using the Maximum a Posteriori (MAP) framework. Given the observations up to the current frame, we estimate the optimal object trajectories by solving two MAP estimation problems: object detection and trajectory-detection association. By introducing the sequential trajectory prior, *i.e.*, the prior information from previous frames about "good" trajectories, into MAP estimation, the output of the pre-trained object detector is refined and the correctness of the association between trajectories and detections is enhanced. In addition, the sequential trajectory prior allows the two MAP stages interact with each other in a sequential manner, which facilitates online multi-object tracking. Our experiments on publicly available challenging datasets demonstrate that the proposed algorithm provides superior performance in various complex scenes.

Keywords: Online multi-object tracking · Data association · Maximum a posteriori estimation · Sequential trajectory prior

1 Introduction

Multi-object tracking is a very challenging problem, especially in complex scenes, due to frequent occlusions and interactions among similar-looking objects. Driven by the recent development of object detectors [1–3], tracking-by-detection has become a popular technique for multi-object tracking. With the detection responses provided by detectors, tracking-by-detection approaches associate these detections across frames to form the trajectories of objects.

Many tracking methods [4–6] address the association problem in a large temporal window, which seek for the optimum detection assignments by considering a batch of frames at a time. Due to the utilization of future information, they can handle detection errors and tracking failures caused by occlusions. However, it is difficult to apply the batch methods to time-critical applications, since they provide tracking results with a significant temporal delay.

Our work focuses on online multi-object tracking which only considers observations up to the current frame and sequentially builds trajectories via frame-by-frame association with online provided detections. Compared with the batch

S. Arik et al. (Eds.): ICONIP 2015, Part I, LNCS 9489, pp. 623–633, 2015.
DOI: 10.1007/978-3-319-26532-2_69

methods, online tracking systems [7–10] can be applied to real-time applications, but suffer from performance degradation in complex scenes. We aim to overcome the limitations for online multi-object tracking and to achieve high quality tracking results in complex scenes.

In this paper, we formulate the online multi-object tracking problem under a Beyesian framework, and treat detection and association as two collaborative maximum a posteriori (MAP) estimation problems by introducing the *sequential trajectory prior*. The basic idea is that the observations from previous frames contain useful prior information to assist the estimation of object trajectories in the current frame. Intuitively, it is better to allow the high-confidence trajectories to guide the current estimation of hard-to-see detections. And, for trajectory-detection association, more reliable detections are likely linked to high-confidence trajectories. We thus model such cues as the sequential trajectory prior, and use MAP estimation to simultaneously refine the detector output and enhance the trajectory-detection association correctness. We show that the two MAP stages interact with each other via the sequential trajectory prior: high-confidence trajectories from previous frame provide reliable prior information to refine the detections in the detection stage, and accurate detections facilitate the association stage to generate more confident trajectories. Our experiments demonstrate that the resulting algorithm provides superior tracking performance in various complex scenes.

Previous methods [10–12] exploit the prior information from previous frames for online multi-object tracking. Luo *et al.* [11] introduced a spatio-temporal consistency constraint to their online detector learning stage. Bae and Yoon [10] used trajectory confidence to assist their local and global association approach. Their work is extended in [12] by introducing a track existence probability into data association. However, these methods utilize the prior information only in the detection or association task. In contrast, we explicitly introduce the sequential trajectory prior into both the detection and association stages by using a unified MAP framework. As a result, the online multi-object tracking performance is significantly improved especially in complex scenes.

2 Our Approach

2.1 Problem Formulation

Let $\mathbb{X}_{1:t}$, $\mathbb{Y}_{1:t}$ and $\mathbb{Z}_{1:t}$ be the trajectories, detections and observed images up to frame t, respectively. We adopt a Bayesian approach to formulate the online multi-object tracking problem, where trajectories $\mathbb{X}_{1:t}$ and detections $\mathbb{Y}_{1:t}$ are random variables and the goal is to maximize the joint posterior distribution over $\mathbb{X}_{1:t}$ and $\mathbb{Y}_{1:t}$ given observed images $\mathbb{Z}_{1:t}$. Formally,

$$\begin{aligned}(\mathbb{X}^*_{1:t}, \mathbb{Y}^*_{1:t}) &= \underset{\mathbb{X}_{1:t}, \mathbb{Y}_{1:t}}{\arg\max}\, P\left(\mathbb{X}_{1:t}, \mathbb{Y}_{1:t} | \mathbb{Z}_{1:t}\right) \\ &= \underset{\mathbb{X}_{1:t}, \mathbb{Y}_{1:t}}{\arg\max}\, P\left(\mathbb{X}_{1:t} | \mathbb{Y}_{1:t}, \mathbb{Z}_{1:t}\right) P\left(\mathbb{Y}_{1:t} | \mathbb{Z}_{1:t}\right), \end{aligned} \tag{1}$$

where the second equation used the definition of conditional probability. Since it is impossible to globally optimize Eq. (1) using brute force search, we expand the original formulation by sequentially estimating the current trajectories $\mathbb{X}_t$ and detections $\mathbb{Y}_t$ conditional on the previous results using the tracking-by-detection strategy. The problem is then decomposed into two MAP estimation stages:

$$\text{(detection)} \quad \mathbb{Y}_t^* = \arg\max_{\mathbb{Y}_t} P\left(\mathbb{Y}_t | \mathbb{Z}_{1:t}\right), \tag{2}$$

$$\text{(association)} \quad \mathbb{X}_t^* = \arg\max_{\mathbb{X}_t} P\left(\mathbb{X}_t | \mathbb{Y}_t^*, \mathbb{X}_{t-1}\right). \tag{3}$$

Specifically, in the detection stage, we obtain a MAP estimation of the detections $\mathbb{Y}_t^*$ by considering the observed images up to the current frame $\mathbb{Z}_{1:t}$. The trajectory estimation problem is then reformulated as a MAP estimation of pairwise associations between $\mathbb{X}_{t-1}$ and $\mathbb{Y}_t^*$ in the association stage.

2.2 Detection Refinement with MAP Estimation

Based on the Bayesian rule, the MAP estimation of the detections $\mathbb{Y}_t^*$ defined in Eq. (2) can be represented as

$$\mathbb{Y}_t^* = \arg\max_{\mathbb{Y}_t} \frac{P\left(\mathbb{Z}_t | \mathbb{Y}_t, \mathbb{Z}_{1:t-1}\right) P\left(\mathbb{Y}_t | \mathbb{Z}_{1:t-1}\right)}{P\left(\mathbb{Z}_t | \mathbb{Z}_{1:t-1}\right)}, \tag{4}$$

where $P\left(\mathbb{Z}_t | \mathbb{Y}_t, \mathbb{Z}_{1:t-1}\right)$ models the observation likelihood function which measures how well the hypothetical detections explain the observed image, and $P\left(\mathbb{Y}_t | \mathbb{Z}_{1:t-1}\right)$ is a prior detection probability which represents the prior information collected from the previous observations.

Prior Detection Probability. We approximately compute the prior detection probability based on the spatio-temporal consistency assumption during tracking. That is, the object states in two subsequent frames should not change drastically. Intuitively, the detections in frame t are much likely to appear around the trajectories from frame $(t-1)$. To utilize such prior, we predict the object states of high-confidence trajectories through Kalman filters, and use the predicted states to produce a density map to represent the prior detection probability. The trajectory confidence is defined by Eq. (8) in Sect. 2.3. Formally, we compute a density map D_t^k for a specific confident object k at frame t as

$$D_t^k(\mathbf{p}) = \exp\left(-\frac{\|\mathbf{p} - \mathbf{p}_k\|^2}{2\sigma_k^2}\right), \tag{5}$$

where $\mathbf{p}$ is the image position, $\mathbf{p}_k$ is the predicted position of object k, and σ_k is the scale parameter which is proportional to the scale of object k (set to 5 times the object scale in our implementation). Suppose that we have c confident objects from high-confidence trajectories in frame $(t-1)$, the density map D_t corresponding to $P\left(\mathbb{Y}_t | \mathbb{Z}_{1:t-1}\right)$ is generated by combining the density maps of all confident objects, expressed as $D_t = \max(D_t^0, D_t^1, \ldots, D_t^c)$. Note that D_t^0 is a const density map where the prior detection probability for each position is equal to 0.5, which is used to prevent the suppression of newly appeared objects.

Observation Likelihood Function. We revisit the detection confidence map produced by the pre-trained object detector to represent the observation likelihood function $P\left(\mathbb{Z}_t|\mathbb{Y}_t, \mathbb{Z}_{1:t-1}\right)$. Following the general object detection strategy, we generate the hypothetical detections $\mathbb{Y}_t$ in multiple scales. Hence, $P\left(\mathbb{Z}_t|\mathbb{Y}_t, \mathbb{Z}_{1:t-1}\right)$ is expressed as multiple confidence maps by applying the object detector to the observed image $\mathbb{Z}_t$ in multiple scales.

Posterior Detection Probability. Combining the observation likelihood function and the prior detection probability mentioned above, we can estimate the posterior detection probability as indicated in Eq. (4). Since the normalized term $P\left(\mathbb{Z}_t|\mathbb{Z}_{1:t-1}\right)$ is constant, we simply use the density map D_t to refine the multiple confidence maps produced by the detector. Then the optimal detections $\mathbb{Y}_t^*$ is obtained by applying non-maximum suppression to the refined confidence maps. Most existing methods use the observation likelihood $P\left(\mathbb{Z}_t|\mathbb{Y}_t, \mathbb{Z}_{1:t-1}\right)$ to approximate the posterior $P\left(\mathbb{Y}_t|\mathbb{Z}_{1:t}\right)$, which actually ignores the useful prior information. In this paper, we employ the prior information from previous frames to model a prior detection probability $P\left(\mathbb{Y}_t|\mathbb{Z}_{1:t-1}\right)$ which actually refines the detector output in a principle manner.

2.3 Data Association with MAP Estimation

Since the number of all possible enumerations of $\mathbb{X}_t$ given the existing trajectories $\mathbb{X}_{t-1}$ and the refined detections $\mathbb{Y}_t^*$ is huge, directly solving Eq. (3) is intractable. We turn to solve a data association problem and then obtain the optimal trajectories $\mathbb{X}_t^*$ by updating $\mathbb{X}_{t-1}$ with the associated detections.

Suppose that we have m trajectories $\mathbb{X}_{t-1} = \{X^i\}_{i=1}^m$ at frame $t-1$ and n refined detections $\mathbb{Y}_t^* = \{\mathbf{y}^j\}_{j=1}^n$ at frame t, where X^i is the trajectory of the i-th object and $\mathbf{y}^j$ is the j-th refined detection. Note that we drop the time index for simplicity since the association is exactly between $\mathbb{X}_{t-1}$ and $\mathbb{Y}_t^*$. We define an event $\Psi_{i,j}$ to represent that the j-th refined detection is associated with the i-th trajectory. Then, the pairwise association problem between $\mathbb{X}_{t-1}$ and $\mathbb{Y}_t^*$ can be expressed as a MAP estimation formulation,

$$\Psi_{i,j}^* = \underset{\Psi_{i,j}}{\arg\max}\, P\left(\Psi_{i,j}|\mathbb{Y}_t^*, \mathbb{X}_{t-1}\right), \tag{6}$$

where $P\left(\Psi_{i,j}|\mathbb{Y}_t^*, \mathbb{X}_{t-1}\right)$ is the the posterior association probability. It can be computed by applying the Bayesian rule,

$$P\left(\Psi_{i,j}|\mathbb{Y}_t^*, \mathbb{X}_{t-1}\right) = \frac{P\left(\mathbb{Y}_t^*|\Psi_{i,j}, \mathbb{X}_{t-1}\right) P\left(\Psi_{i,j}|\mathbb{X}_{t-1}\right)}{P\left(\mathbb{Y}_t^*|\mathbb{X}_{t-1}\right)}, \tag{7}$$

where $P\left(\mathbb{Y}_t^*|\Psi_{i,j}, \mathbb{X}_{t-1}\right)$ is the likelihood that indicates the possibility of observing the detections $\mathbb{Y}_t^*$ given the existing trajectories $\mathbb{X}_{t-1}$ and the association

$\Psi_{i,j}$, and $P\left(\Psi_{i,j}|\mathbb{X}_{t-1}\right)$ is the prior association probability that measures the possibility of the association $\Psi_{i,j}$ before data association.

Prior Association Probability. To compute the prior association probability $P\left(\Psi_{i,j}|\mathbb{X}_{t-1}\right)$, we exploit two kinds of prior information before performing data association: the trajectory confidence and the detection reliability.

Similar to [10], we use a trajectory confidence score function $\Delta(X^i)$ to measure the reliability of an existing trajectory X^i,

$$\Delta(X^i) = \exp\left(-\beta \cdot \frac{M}{L}\right) \times \left(\frac{1}{L}\sum_{k\in\Omega^i} \Phi_k^i\right), \tag{8}$$

where L is the number of frames in which the trajectory has associated detections, $M = |X^i| - L$ is the number of frames in which the object is missing, Ω^i indicates the set of frames in which the trajectory X^i has associated detections, Φ_k^i is the posterior association probability between X^i and the associated detection at frame k, and β is a control parameter depending on the detection performance. Since the trajectory confidence lies in $[0, 1]$, we consider a trajectory as a high-confidence when $\Delta(X^i) > 0.5$.

The reliability of a detection $\mathbf{y}^j$ can be directly represented as the posterior defined in Sect. 2.2, simply denoted as $\delta(\mathbf{y}^j)$. Then the prior $P\left(\Psi_{i,j}|\mathbb{X}_{t-1}\right)$ can be intuitively approximated as

$$P\left(\Psi_{i,j}|\mathbb{X}_{t-1}\right) \approx \frac{\delta(\mathbf{y}^j)}{\sum_{v=1}^{n}\delta(\mathbf{y}^v)} \cdot \Delta(X^i), \tag{9}$$

where we impose the constraint that the association events for a trajectory X^i are mutually exclusive.

Observation Likelihood Function. We assume that the detections in $\mathbb{Y}_t^*$ are conditionally independent given the existing trajectories $\mathbb{X}_{t-1}$ and the association $\Psi_{i,j}$. Then the likelihood $P\left(\mathbb{Y}_t^*|\Psi_{i,j}, \mathbb{X}_{t-1}\right)$ can be computed as

$$P\left(\mathbb{Y}_t^*|\Psi_{i,j}, \mathbb{X}_{t-1}\right) = \prod_{v=1}^{n} P\left(\mathbf{y}^v|\Psi_{i,j}, \mathbb{X}_{t-1}\right). \tag{10}$$

Note that $P\left(\mathbf{y}^j|\Psi_{i,j}, \mathbb{X}_{t-1}\right) = P\left(\mathbf{y}^j|X^i\right)$ is the association likelihood between $\mathbf{y}^j$ and X^i. We compute the the association likelihood by using the appearance, shape, and motion cues, similar to [7]. The remaining task is to estimate the likelihood $P\left(\mathbf{y}^v|\Psi_{i,j}, \mathbb{X}_{t-1}\right)$ with $v \neq j$ which can be explained as the probability that the detection $\mathbf{y}^v$ is not originated from the trajectory X^i.

We consider two situations where the detection $\mathbf{y}^v$ can be observed: $\mathbf{y}^v$ is originated from other trajectories except X^i, or $\mathbf{y}^v$ is a false positive detection. Using the definition of marginal probability, the likelihood $P\left(\mathbf{y}^v|\Psi_{i,j}, \mathbb{X}_{t-1}\right)$ with $v \neq j$ can be computed by

$$\begin{aligned} P\left(\mathbf{y}^v|\mathit{\Psi}_{i,j},\mathbb{X}_{t-1}\right) &= P\left(\mathbf{y}^v,\mathit{\Psi}_{0,v}|\mathit{\Psi}_{i,j},\mathbb{X}_{t-1}\right)+\sum_{u\neq i}P\left(\mathbf{y}^v,\mathit{\Psi}_{u,v}|\mathit{\Psi}_{i,j},\mathbb{X}_{t-1}\right)\\ &= P\left(\mathbf{y}^v,\mathit{\Psi}_{0,v}|\mathbb{X}_{t-1}\right)+\sum_{u\neq i}P\left(\mathbf{y}^v,\mathit{\Psi}_{u,v}|\mathbb{X}_{t-1}\right), \end{aligned} \tag{11}$$

where $\mathit{\Psi}_{0,v}$ means that the detection $\mathbf{y}^v$ is not associated with any trajectory. Denote $P_{u,v} = P\left(\mathit{\Psi}_{u,v}|\mathbb{X}_{t-1}\right)$ as the prior association probability defined in Eq. (9), and $\rho = P\left(\mathbf{y}^v|\mathit{\Psi}_{0,v},\mathbb{X}_{t-1}\right)$ as the const probability that a detection becomes false positive, we have

$$P\left(\mathbf{y}^v,\mathit{\Psi}_{0,v}|\mathbb{X}_{t-1}\right) = P\left(\mathbf{y}^v|\mathit{\Psi}_{0,v},\mathbb{X}_{t-1}\right)P\left(\mathit{\Psi}_{0,v}|\mathbb{X}_{t-1}\right) = \rho\cdot\prod_{u=1}^{m}\left(1-P_{u,v}\right), \tag{12}$$

$$P\left(\mathbf{y}^v,\mathit{\Psi}_{u,v}|\mathbb{X}_{t-1}\right) = P\left(\mathbf{y}^v|\mathit{\Psi}_{u,v},\mathbb{X}_{t-1}\right)P\left(\mathit{\Psi}_{u,v}|\mathbb{X}_{t-1}\right) = P\left(\mathbf{y}^v|X^u\right)P_{u,v}, \tag{13}$$

and thus

$$P\left(\mathbf{y}^v|\mathit{\Psi}_{i,j},\mathbb{X}_{t-1}\right) = \rho\cdot\prod_{u=1}^{m}\left(1-P_{u,v}\right)+\sum_{u\neq i}P\left(\mathbf{y}^v|X^u\right)P_{u,v}. \tag{14}$$

Then the observation likelihood function $P\left(\mathbb{Y}_t^*|\mathit{\Psi}_{i,j},\mathbb{X}_{t-1}\right)$ can be obtained by substituting Eq. (14) into Eq. (10),

$$P\left(\mathbb{Y}_t^*|\mathit{\Psi}_{i,j},\mathbb{X}_{t-1}\right) = P\left(\mathbf{y}^j|X^i\right)\prod_{v\neq j}\Theta_{i,j}^v, \tag{15}$$

where we denote $\Theta_{i,j}^v = P\left(\mathbf{y}^v|\mathit{\Psi}_{i,j},\mathbb{X}_{t-1}\right)$ with $v\neq j$ for simplicity.

Posterior Association Probability. Denote the normalization term in Eq. (7) as $\gamma = P\left(\mathbb{Y}_t^*|\mathbb{X}_{t-1}\right)$, we can derive the posterior as

$$P\left(\mathit{\Psi}_{i,j}|\mathbb{Y}_t^*,\mathbb{X}_{t-1}\right) = \gamma^{-1}P_{i,j}P\left(\mathbf{y}^j|X^i\right)\prod_{v\neq j}\Theta_{i,j}^v. \tag{16}$$

In a similar manner, the posterior association probability for the non association event $\mathit{\Psi}_{i,0}$ of the trajectory X^i can be acquired by

$$\begin{aligned} P(\mathit{\Psi}_{i,0}|\mathbb{Y}_t^*,\mathbb{X}_{t-1}) &= \frac{P\left(\mathbb{Y}_t^*|\mathit{\Psi}_{i,0},\mathbb{X}_{t-1}\right)P\left(\mathit{\Psi}_{i,0}|\mathbb{X}_{t-1}\right)}{P\left(\mathbb{Y}_t^*|\mathbb{X}_{t-1}\right)}\\ &= \gamma^{-1}\left(1-\sum_{j=1}^{n}P_{i,j}\right)\prod_{v=1}^{n}\Theta_{i,0}^v. \end{aligned} \tag{17}$$

Using the fact that $\sum_{j=1}^{n}P\left(\mathit{\Psi}_{i,j}|\mathbb{Y}_t^*,\mathbb{X}_{t-1}\right)+P\left(\mathit{\Psi}_{i,0}|\mathbb{Y}_t^*,\mathbb{X}_{t-1}\right)=1$, the normalization term γ can be computed as

$$\begin{aligned} \gamma &= \sum_{j=1}^{n}\left(P_{i,j}P\left(\mathbf{y}^j|X^i\right)\prod_{v\neq j}\Theta_{i,j}^v\right)+\left(1-\sum_{j=1}^{n}P_{i,j}\right)\prod_{v=1}^{n}\Theta_{i,0}^v\\ &= \left(1-\sum_{j=1}^{n}P_{i,j}+\sum_{j=1}^{n}Q_{i,j}\right)\prod_{v=1}^{n}\Theta_{i,0}^v, \end{aligned} \tag{18}$$

where $Q_{i,j} = P_{i,j} P\left(\mathbf{y}^j | X^i\right) / \Theta_{i,0}^j$, and the second equation uses the fact $\Theta_{i,j}^v = \Theta_{i,0}^v$ when $v \neq j$.

Data Association. With the posterior probabilities given by Eqs. (16) and (17), the data association problem of Eq. (6) can be solved by the Hungarian algorithm [13]. Specifically, a association cost matrix $S = [s_{ij}]_{m \times n}$ is constructed with each entry $s_{ij} = -\log(P\left(\Psi_{i,j} | \mathbb{Y}_t^*, \mathbb{X}_{t-1}\right))$ to indicate the cost when j-th refined detection is associated with the i-th trajectory. Then the optimal trajectory-detection pairs are determined by minimizing the total cost in $S_{m \times n}$. When the association cost of a trajectory-detection pair is less than the cost of non association $-\log(P\left(\Psi_{i,0} | \mathbb{Y}_t^*, \mathbb{X}_{t-1}\right))$, the detection $\mathbf{y}^j$ is associated with X^i. A Kalman filter is used to refine the object states for a trajectory, with the associated detections as the measurement data. Then the confidence $\Delta(X^i)$ is updated using Eq. (8). The detections that are not associated with any existing trajectories are used to initialize a new potential trajectory. Once the length of a potential trajectory grows over a threshold (set to 5 frames in our implementation), it gets formally initialized.

3 Experiments

In this section, we give a detailed analysis of our approach compared to the state-of-the-art in multi-object tracking. The state-of-the-art trackers include DP [4], TBD [6], CEM [5] and CMOT [10], in which the CMOT tracker is online algorithms while the other trackers perform multi-object tracking in a batch mode. We report the results by using the source codes publicly provided by the authors with the same object detector and their default parameters.

3.1 Implementation Details

Our online multi-object tracking algorithm is implemented in MATLAB, and operates entirely in the image coordinate without camera or ground plane calibration. Without code optimization and parallel programming, our algorithm runs at about 10 fps on an Intel Core i7 3.5 GHz PC with 16 GB memory. The system parameters that need to be set beforehand include the control factor β in Eq. (8), and the const probability ρ in Eq. (14). In our implementation, we empirically set $\beta = 2$ and $\rho = 0.1$ for all experiments.

3.2 Datasets and Object Detector

We use the following datasets for performance evaluation: *PETS2009* dataset [14], *TUD* dataset [15], and *ETH Mobile Scene* (*ETHMS*) [16]. The *PETS2009* dataset shows an out door survivance scene where large amount of pedestrians enter and exit the filed-of-view. We adopt the widely used $S2L1$ and $S2L2$ sequences for evaluation. In the *TUD* dataset, the sequences *Campus*, *Crossing* and *Stadtmitte* are used, where the challenges include severe occlusions between

Table 1. Quantitative comparison results. Batch methods are marked with an asterisk. **Bold** scores highlight the best results.

Method	MOTA↑	MOTP↑	FP↓	FN↓	MT↑	ML↓	IDS↓	FG↓
⋆DP [4]	31.1%	**71.6%**	3,695	11,890	19.6%	33.2%	3,177	1,277
⋆CEM [5]	39.7%	70.7%	4,656	11,411	24.5%	34.0%	349	**640**
⋆TBD [6]	35.4%	71.4%	6,267	9,995	27.5%	**31.3%**	1,329	1,025
CMOT [10]	21.7%	69.9%	7,912	11,354	20.1%	33.4%	1,998	1,139
Ours (w/o all)	27.3%	70.5%	4,855	13,293	21.5%	41.3%	679	990
Ours (w/o MAP assoc.)	43.5%	71.2%	3,982	10,764	24.7%	37.0%	634	931
Ours (w/o MAP det.)	40.8%	70.9%	4,292	11,521	28.1%	34.0%	312	790
Ours (with all)	**49.0%**	71.2%	**3,603**	**9,942**	**31.0%**	32.8%	**235**	754

objects and low viewpoint. In the *ETHMS* dataset, we evaluate our algorithm on the sequences *Bahnhof*, *Jelmoli* and *SunnyDay*, which are taken by a moving camera in crowded street scenes. In total, the test datasets contain over 3500 frames and 368 annotated trajectories (27240 bounding boxes). For fair comparison, we use the ground truth publicly provided by Milan *et al.* [5].

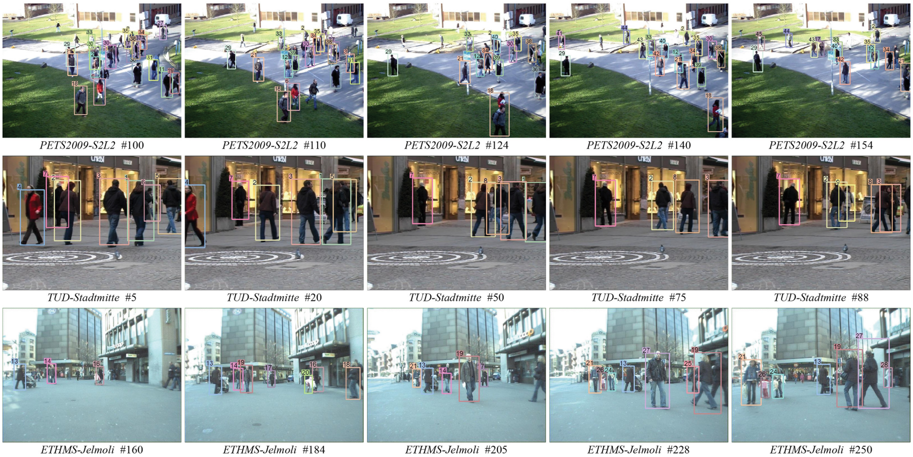
To efficiently acquire online detections, we use the aggregate channel features object detector [3] which can be operated in almost real time. The detector is trained on the INRIA dataset [1] with default parameters.

3.3 Evaluation Metrics

We use the widely accepted CLEAR performance metrics [17] for quantitative evaluation: the multiple object tracking precision (MOTP↑) that evaluates average overlap rate between true positive tracking results and the ground truth, and the multiple object tracking accuracy (MOTA↑) which indicates the accuracy composed of false positives (FP↓), false negatives (FN↓) and identity switches (IDS↓). Additionally, we report measures defined by Li *et al.* [18], including the percentage of mostly tracked (MT↑) and mostly lost (ML↓) ground truth trajectories, as well as the number of times that a ground truth trajectory is interrupted (FG↓). Here, ↑ means that higher scores indicate better results, and ↓ represents that lower is better.

3.4 Results and Discussion

Quantitative results of our algorithm compared with the state-of-the-art tracking methods on the datasets are listed in Table 1, and sample results are shown in Fig. 1. Overall, our algorithm outperforms the competing online tracker CMOT, and achieves competitive results compared to the state-of-the-art batch methods (*i.e.*, DP, TBD and CEM). It owes to the proposed two collaborative MAP estimation stages which simultaneously incorporate the sequential trajectory prior into both the detection and association procedures during tracking. As can be observed from the quantitative evaluation results, our algorithm achieves far



PETS2009-S2L2 #100 *PETS2009-S2L2* #110 *PETS2009-S2L2* #124 *PETS2009-S2L2* #140 *PETS2009-S2L2* #154

TUD-Stadtmitte #5 *TUD-Stadtmitte* #20 *TUD-Stadtmitte* #50 *TUD-Stadtmitte* #75 *TUD-Stadtmitte* #88

ETHMS-Jelmoli #160 *ETHMS-Jelmoli* #184 *ETHMS-Jelmoli* #205 *ETHMS-Jelmoli* #228 *ETHMS-Jelmoli* #250

Fig. 1. Sample tracking results of our method on three representative test video sequences (*PETS2009-S2L2*, *TUD-Stadtmitte* and *ETHMS-Jelmoli*). At each frame, objects with different IDs are indicated by bounding boxes with different colors.

superior performance in terms of MOTA, FP and FN, which indicates that the detection refinement stage integrating with the sequential trajectory prior significantly facilitates the tracking process. In addition, we achieve excellent results in terms of MT, ML, IDS and FG, demonstrating that the combination of association likelihood and sequential prior benefits the correct association between trajectories and detections. As shown in the qualitative examples of tracking results in Fig. 1, our method is able to accurately track the target persons under various challenging conditions.

To demonstrate the effectiveness of the proposed two MAP estimation stages with the sequential trajectory prior, we build three baseline algorithms to do validation and analyze various aspects of our approach. The comparison results between our approach and three baseline algorithms are also listed in Table 1, where removal of the MAP estimation stage means removal the prior and only using the likelihood as most tracking methods do. As can be seen from the comparison results, the baseline algorithm without both of the two MAP estimation stages shows severe performance degradation. Using sequential trajectory prior to refine the detections results in significant improvement on MOTA and FN, which validates that the sequential trajectory prior indeed assists the detector to recall more accuracy detections. In addition, incorporating sequential trajectory prior to trajectory-detection association apparently improves the accuracy in terms of MT, ML, IDS and FG, which demonstrated that the association correctness is improved by using the MAP estimation of the posterior association probability. The proposed algorithm considers the sequential trajectory prior in both the detection and association stages, and thus shows the best performance.

4 Conclusion

We have proposed an online multi-object tracking-by-detection algorithm by using the Maximum a Posteriori (MAP) framework. To account for noisy detections and improve trajectory-detection association correctness, we exploit the prior information contained in previous frames, such as the positions of objects that most likely to appear, the adaptive confidences of trajectories and the detection reliability, to guide the detection and association stages in the current frame. By using these sequential trajectory priors in MAP, the tracker is able to recall more reliable detections and alleviate the ambiguity of trajectory-detection association, and achieves great improvement on tracking performance.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
2. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE TPAMI **32**(9), 1627–1645 (2010)
3. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE TPAMI **36**(8), 1532–1545 (2014)
4. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR, pp. 1201–1208 (2011)
5. Milan, A., Roth, S., Schindler, K.: Continuous energy minimization for multitarget tracking. IEEE TPAMI **36**(1), 58–72 (2014)
6. Geiger, A., Lauer, M., Wojek, C., Stiller, C., Urtasun, R.: 3D traffic scene understanding from movable platforms. IEEE TPAMI **36**(5), 1012–1025 (2014)
7. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. IJCV **75**(2), 247–266 (2007)
8. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multiperson tracking-by-detection from a single, uncalibrated camera. IEEE TPAMI **33**(9), 1820–1833 (2011)
9. Shu, G., Dehghan, A., Oreifej, O., Hand, E., Shah, M.: Part-based multiple-person tracking with partial occlusion handling. In: CVPR, pp. 1815–1821 (2012)
10. Bae, S.H., Yoon, K.J.: Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In: CVPR, pp. 1218–1225 (2014)
11. Luo, W., Kim, T.K., Stenger, B., Zhao, X., Cipolla, R.: Bi-label propagation for generic multiple object tracking. In: CVPR, pp. 1290–1297 (2014)
12. Bae, S.H., Yoon, K.J.: Robust online multi-object tracking with data association and track management. IEEE TIP **23**(7), 2820–2833 (2014)
13. Kuhn, H.W.: The Hungarian method for the assignment problem. Nav. Res. Logistics Q. **2**(1–2), 83–97 (1995)
14. Ellis, A., Shahrokni, A., Ferryman, J.M.: PETS2009 and Winter-PETS 2009 results: a combined evaluation. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter), pp. 1–8 (2009)
15. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR, pp. 1–8 (2008)

16. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR, pp. 1–8 (2008)
17. Keni, B., Rainer, S.: Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP J. Image Video Process. (2008)
18. Li, Y., Huang, C., Nevatia, R.: Learning to associate: hybridboosted multi-target tracker for crowded scene. In: CVPR, pp. 2953–2960 (2009)