Detection- and Trajectory-Level Exclusion in Multiple Object Tracking

Anton Milan[1] Konrad Schindler[2] Stefan Roth[1]

[1]Department of Computer Science, TU Darmstadt [2]Photogrammetry and Remote Sensing Group, ETH Zürich

Abstract

When tracking multiple targets in crowded scenarios, modeling mutual exclusion between distinct targets becomes important at two levels: (1) in data association, each target observation should support at most one trajectory and each trajectory should be assigned at most one observation per frame; (2) in trajectory estimation, two trajectories should remain spatially separated at all times to avoid collisions. Yet, existing trackers often sidestep these important constraints. We address this using a mixed discrete-continuous conditional random field (CRF) that explicitly models both types of constraints: Exclusion between conflicting observations with supermodular pairwise terms, and exclusion between trajectories by generalizing global label costs to suppress the co-occurrence of incompatible labels (trajectories). We develop an expansion move-based MAP estimation scheme that handles both non-submodular constraints and pairwise global label costs. Furthermore, we perform a statistical analysis of ground-truth trajectories to derive appropriate CRF potentials for modeling data fidelity, target dynamics, and inter-target occlusion.

1. Introduction

The task of visual multi-target tracking is to recover the spatio-temporal trajectories of a (usually unknown) number of targets from a video sequence. Tracking multiple targets – often people or vehicles – has a wide range of applications ranging from robotics to video surveillance. Even though the field has made tremendous progress since the early works [*e.g.*, 10], modern systems still have clear limitations, especially as the observed scenes get more crowded. This is not entirely surprising, since the solution space grows rapidly as the number of visible targets and the length of their trajectories increases. Moreover, physical limits mandate a growing number of constraints (such as mutual exclusion) as more targets are in close proximity to each other.

Tracking in realistic sequences is further complicated by background clutter, poor contrast, and partial or full occlusions, such as from other targets. *Tracking-by-detection* approaches that rely on powerful object (class) detectors are thus becoming increasingly popular [5, 13, 21, 23, 25, 26]. In this case, targets are detected independently in each frame with an offline-trained object detector. This not only addresses adverse imaging conditions, but also reduces drift and allows to bridge severe occlusions and other temporary loss of evidence. We follow this approach here.

(a) (b)

Figure 1. Typical failure cases *(top)* are addressed with the proposed discrete-continuous CRF *(bottom)*: Detections are forced to take on different labels *(a)* and physically overlapping trajectories are suppressed even if they do not share detections *(b)*.

A large class of trackers aims to further improve robustness by processing entire frame batches, rather than inferring the current state solely from the track history [*e.g.*, 3, 5, 16, 26]. While this may lead to potential contradictions in frames that occur in different batches and to a mild time lag, the crucial advantage is greater reliability, since longer time windows afford both more data and stronger models. Batch-type multi-target trackers typically formulate a joint energy function for all targets in all frames.

We can distinguish two categories of batch approaches: The vast majority focuses on purely discrete optimization for solving either data association [23, 26, *etc.*] or trajectory estimation [*e.g.*, 5]. This allows one to encode complex constraints, including inter-object exclusion, in a natural way. The disadvantage is that the trajectories need to be discretized themselves, hence the necessarily finite spatial resolution can limit tracking performance and lead to visible

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.472

artifacts. Moreover, exclusion is only handled either at the data-association level [23, 26], or at the trajectory level [5]. The second group of methods uses alternative state spaces, such as purely continuous [3] or mixed discrete-continuous [4] formulations. While these allow estimating target trajectories in continuous space, they have other drawbacks. [3] needed to employ an optimization scheme with ad-hoc jump moves; [4] used the label cost framework of [12], which is not designed for imposing exclusion constraints.

In this paper, we propose a mixed discrete-continuous conditional random field (CRF) for multi-target tracking that aims to combine the advantages of continuous-space trajectory estimation with the advantages of discrete methods for enforcing exclusion constraints. We specifically address mutual exclusion both at the data-association and at the trajectory level (*cf*. Fig. 1). We thus go beyond previous discrete-continuous trackers [4] that do not perform explicit exclusion reasoning, and beyond previous discrete approaches that model exclusion only at the trajectory [5] or at the data-association level [26].

We make the following contributions: *(i)* We extend the global label subset costs of [12] to pairwise label costs that allow penalizing the co-occurrence of competing labels in the solution; *(ii)* we show how pairwise co-occurrence label costs can be used to model trajectory exclusion for multi-target tracking; *(iii)* we enforce physically plausible data association with non-submodular pairwise constraints; *(iv)* we propose an iterative MAP estimation scheme based on expansion moves for the resulting non-submodular multi-label CRF energy with pairwise label costs; and *(v)* we analyze the statistical properties of real trajectories and observations on ground truth data to derive CRF potentials for various model components (data fidelity, dynamic model, occlusion duration). Together, these advances yield a more faithful and more accurate model of multi-target tracking, which nevertheless remains tractable and delivers improved tracking results. To the best of our knowledge our approach is the first to combine both unique data association of individual observations and physical collision-avoidance at the trajectory level in a common model.

2. Related Work

Visual tracking has been an ongoing research topic for decades, and a full literature review is beyond our scope. In the presence of a single target, tracking can be performed by estimating the target location and motion in every frame and building the trajectory through interpolation [17]. In the presence of multiple targets an additional challenge arises, often referred to as *data association*: each detection must be assigned a target identifier or discarded as a false alarm. Filtering approaches, such as JPDA [15] or particle filters [8, 22], estimate the state and association online, *i.e*. by only considering the past states and present observations.

Although online processing is desirable for time critical applications, batch approaches have become increasingly popular due to their superior robustness [3, 5, 16, 21, 26]. To keep the optimization tractable, the objective can be discretized and (near-)optimal solutions can be obtained by linear programming relaxations [5, 16] or computing the maximal flow in a network graph [21, 23, 26]. Constraints on data association can also be mapped to other graph problems, for which efficient (approximate) algorithms exist [9, 25]. To overcome the drawback of discretization, [3] formulates the objective entirely in the continuous domain and combines gradient descent and greedy jump moves for optimization. The mixed discrete-continuous model of [4] preserves the benefits of a continuous trajectory space, while allowing for discrete data association. Trajectory level constraints are formulated as global label costs, and optimized with a graph cut-based discrete-continuous scheme [12].

Mutual exclusion between targets, *i.e*. a term that penalizes or entirely disallows solutions where two or more targets collide, is a crucial property of multi-target tracking. Approaches that focus on data association [*e.g*., 21, 23, 26] usually represent the state space of target trajectories through the underlying detections. While this allows a one-to-one mapping between each detection and each trajectory, situations where the data is missing are not captured properly, hence two trajectories may in fact intersect. Grid-based methods [*e.g*., 5] explicitly model mutual exclusion between target locations at the trajectory level by imposing linear constraints. However, such a discrete grid is a somewhat crude approximation of the continuous trajectory space. A continuous mutual exclusion term [3] leads to a non-convex objective that is optimized with ad-hoc jumps.

In the present work we follow the basic idea of a mixed discrete-continuous representation [4]. Aside from representing trajectories in continuous space, this allows us to impose mutual exclusion simultaneously at the data association and at the trajectory level using the discrete part of the model. While these constraints go beyond the capabilities of the label cost optimization framework of [12], we show how it can be extended appropriately.

3. CRF Model

Following the increasingly popular *tracking-by-detection* framework [2–5, 8, 21, 23–26], we rely on an independently obtained set of target hypotheses $\mathbf{D}$. To make different tracking systems comparable, we use publicly available detector responses [3, 19, 24] throughout, which are generated by popular object detectors [*e.g*., 11]. The individual detections then serve as input data for the reconstruction of the final trajectories.

The ultimate goal of such *tracking-by-detection* approaches is twofold: *(i)* Every detection needs to be explained correctly, *i.e*. either assigned to a target or identified

as clutter. Here it is important to not only enforce a unique assignment for each detection, but also to constrain that two simultaneous observations must not be assigned to the same target. *(ii)* The resulting trajectories have to explain the observations in a physically plausible way, *i.e*. all velocities must remain within physical limits and trajectories must not overlap, because two objects cannot occupy the same physical space at the same time. Especially this latter constraint poses a challenge and has often been neglected in order to keep the optimization simple [4, 21, 26].

In this work we address both challenges, at two different levels: The first one is handled by introducing pairwise terms between competing detections to avoid an unnatural interpretation of the data. That constraint on its own is insufficient, however, since it could lead to phantom trajectories that mirror existing ones, but are physically not plausible. The second challenge is thus approached at the trajectory level. We introduce a novel pairwise co-occurrence label cost that is applied only if both labels are present in a solution. Although these model components introduce pairwise terms that are supermodular, as well as global terms relating many variables, our proposed optimization scheme is able to efficiently minimize the CRF energy to a local minimum. Our experiments show that local minima of the proposed energy lead to better performance, both quantitatively and visually. We begin by describing the proposed exclusion handling and later summarize the remaining model.

3.1. Detection-level exclusion

We first describe how we integrate mutual exclusion at the detection level. In the following, we will identify each detection $d_i^t \in \mathbf{D}$ at location $p_i^t \in \mathbb{R}^2$ with a random variable in a conditional random field (CRF), where t stands for the frame number and i is the index of the detection. Further, let f_d denote the label of d, *i.e*. the assigned target ID of the detection d. Assuming a target size s, it is impossible that two detections originating from the same frame and being at least a distance s apart are caused by the same object. Therefore we introduce an exclusion term

$$\psi_{\mathrm{X}}(f_d, f_{d'}) = \begin{cases} \overline{\psi_{\mathrm{X}}}, & f_d = f_{d'} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

to all edges between simultaneous detector responses

$$(d, d') \in \mathcal{E}_{\mathrm{X}} = \left\{ \left(d_i^t, d_j^t\right) \middle| i \neq j, \left\|p_i^t - p_j^t\right\| > s \right\}. \tag{2}$$

The penalty $\overline{\psi_{\mathrm{X}}}$ is thus incurred if two distant detections are assigned the same trajectory label. For detections that are very close to one another, on the other hand, it is reasonable to accept multiple assignments, since common object detectors sometimes erroneously produce multiple outputs from the same object. This can occur even after non-maxima suppression. The exclusion factors are illustrated in Fig. 2(b).

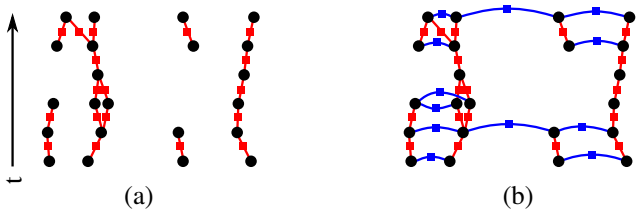

Figure 2. Factor graph of the underlying CRF with black circular nodes representing the random variables (assignment of detections) and square nodes representing the pairwise potentials. For clarity, all unary and high-order potentials are omitted. In addition to simple temporal smoothing factors (red) in *(a)*, we model pairwise exclusion between detections within the same time step (blue, subset shown) to prevent implausible data association *(b)*.

Note that only considering exclusion at the detection level is not enough in order to prevent collisions between targets. In fact, the optimization may otherwise be forced to pick two almost identical trajectories in order to satisfy these inter-object constraints. It is thus crucial not to disregard the path of the actual trajectories.

3.2. Trajectory-level exclusion

Let us now turn to the more challenging task of enforcing exclusion at the level of continuous trajectories. It is obvious that multi-target tracking should take care to prevent situations where two or more targets occupy the same physical space at the same time. Unfortunately, such constraints lead to hard optimization problems.

It has been proposed to encode a collision penalty into a global label cost [4], such that the graph cut framework of [12] can be used for MAP estimation. The penalty considered how much a trajectory overlapped with any other trajectory (whether active or not). To ensure that only one of two overlapping trajectories is suppressed, the penalty was added only to one trajectory in each competing pair. However, the decision which one to penalize was an ad-hoc heuristic, which apparently often fails. Consequently, [4] disabled the collision penalty in all experiments.

Here, we develop a way to seamlessly integrate a co-occurrence potential between two *labels* (*i.e*. trajectories) into the CRF. To simplify the treatment, we will describe the potential in the context of our expansion move-based MAP estimation approach. In particular, we describe the corresponding factor graph for a single α-expansion step, where 0 corresponds to no label change and 1 means a variable is switched to label α. An illustration of the factor graph (without unary and pairwise terms) is depicted in Fig. 3.

Let us first look at the standard per-label cost. Similar to [12], one auxiliary node for each existing label is added and connected to each variable that carries, or may carry, the corresponding label (A_β, A_γ, and A_α in Fig. 3). However, we use a different encoding for the auxiliary variables,

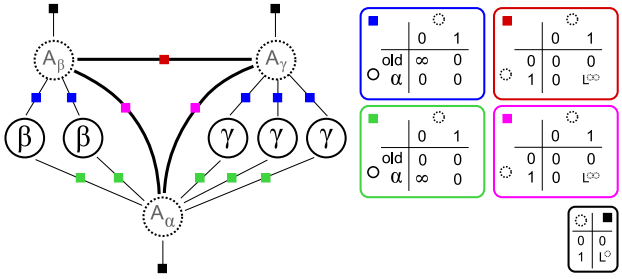

Figure 3. Factor graph encoding of the unary and pairwise label cost before expanding on α. Random variables and their current labels are represented by solid circles, while auxiliary variables are outlined with dashed circles. Solid squares represent unary (black) and pairwise (colored) terms, respectively. The corresponding potentials are depicted on the right with L° and L^∞ being the respective label cost for a single label and a pair of labels. Note that all factors that are unrelated to the label cost are omitted for clarity.

which act as indicator switches for each label: The auxiliary variable contributes the cost L° (black factors) of having a certain label only once if it is switched on, otherwise its associated cost is 0. An infinite pairwise cost[1] (blue and green factors) prevents the indicator from being off when there is at least one node with the corresponding label. While this yields supermodular costs (Eq. (1) already makes the overall cost non-submodular), its purpose will be apparent soon.

We now turn to the pairwise label cost. Having the same graph structure as before, it is possible to insert a connecting factor between each pair of auxiliary variables (red and cyan). The energy should be high if there exist two labels that are unlikely to appear simultaneously. It is therefore reasonable to apply a suitable penalty $L^\infty = \zeta$, if and only if both corresponding auxiliary variables are switched on:

$$h_{\mathrm{X}}(\mathcal{T}_i, \mathcal{T}_j, \mathbf{f}) = \begin{cases} \zeta(\mathcal{T}_i, \mathcal{T}_j), & \exists d, d' : \mathbf{f}_d = i \wedge \mathbf{f}_{d'} = j \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Here $\mathcal{T}_i$ denotes a continuous trajectory of target i. In our case, the co-occurrence penalty is proportional to the spatio-temporal overlap between the two trajectories:

$$\zeta(\mathcal{T}_i, \mathcal{T}_j) = \sum_{t \in \mathbf{O}(\mathcal{T}_i, \mathcal{T}_j)} \zeta^t_{i,j}, \tag{4}$$

which is computed by summing the mutual overlap over all frames in the common lifespan $\mathbf{O}$ of the trajectories. The overlap is approximated with an isotropic sigmoidal function around the center of the target:

$$\zeta^t_{i,j} = \lambda_{\mathrm{X}} \cdot \left(1 - \frac{1}{\exp(-s_a \|\mathcal{T}_i(t) - \mathcal{T}_j(t)\| + s_b)}\right). \tag{5}$$

The two parameters s_a and s_b control the size and the falloff of the sigmoid and are directly related to the application-specific shape of the targets.

[1] In practice a sufficiently large value is used instead.

We emphasize that our formulation of a co-occurrence label cost is general and not restricted to multi-target tracking. It can trivially be transferred to other applications that involve multi-model fitting, such as semantic segmentation or motion estimation. Note that [20], for example, use a co-occurrence cost to prevent unlikely labeling configurations in the context of semantic segmentation. There, however, the cost is overestimated to keep inference tractable. We prefer to model the cost exactly, but can no longer guarantee global optimality of each expansion step.

3.3. Discrete-continuous multiple object tracking

Given a set of detections $\mathbf{D}$ and an over-complete set of potential trajectories $\mathcal{T}$, the goal is to find a data association for $\mathbf{D}$, *i.e.* to assign a unique target ID to each detection or identify it as a false alarm. At the same time, the geometric shape of all active trajectories $\mathcal{T}^* \subseteq \mathcal{T}$ should be fitted to the corresponding detections, such that the residuals between the observation and the tracker output are minimized. To do so, we extend the discrete-continuous multi-target tracking approach of [4] with our exclusion handling.

Let $\mathbf{L} = \{1, \ldots, |\mathcal{T}|, \varnothing\}$ be the set of labels that each random variable $d \in \mathbf{D}$ can attain, and let $\mathbf{f} \in \mathbf{L}^{|\mathbf{D}|}$ be the current labeling. A CRF energy $E(\mathbf{f}, \mathcal{T})$, defined over the discrete labeling $\mathbf{f}$, as well as the continuous trajectories $\mathcal{T}$, is minimized to obtain a plausible solution. To facilitate the optimization, we alternate between updating the discrete variables while keeping the continuous ones fixed, and updating the continuous variables with the discrete ones fixed.

Our complete CRF energy is defined as

$$\begin{aligned} E(\mathbf{f}, \mathcal{T}) = &\sum_d \phi(\mathbf{f}_d, \mathcal{T}) + \\ &\sum_{(d,d') \in \mathcal{E}_{\mathrm{S}}} \psi_{\mathrm{S}}(\mathbf{f}_d, \mathbf{f}_{d'}) + \sum_{(d,d') \in \mathcal{E}_{\mathrm{X}}} \psi_{\mathrm{X}}(\mathbf{f}_d, \mathbf{f}_{d'}) + \\ &\sum_i h_{\mathbf{f}}(\mathcal{T}_i, \mathbf{f}) + \sum_{i, j \neq i} h_{\mathrm{X}}(\mathcal{T}_i, \mathcal{T}_j, \mathbf{f}), \end{aligned} \tag{6}$$

with the following components: The unaries ϕ measure how well the trajectories follow the detector evidence; they are described in more detail in the following section.

The first pairwise term ψ_{S} encourages temporally smooth data association with a standard generalized Potts model (*cf*. Fig. 2(a)). The factors ψ_{S} are defined on pairs of detections in adjacent frames that are spatially close:

$$\mathcal{E}_{\mathrm{S}} = \left\{ \left(d_i^t, d_j^{t+1}\right) \middle| \left\| p_i^t - p_j^{t+1} \right\| < \tau \right\}, \tag{7}$$

where d_i^t denotes the detection i in frame t and p_i^t its (x, y)-location. The second pairwise term ψ_{X} are the detection-level exclusion constraints from Eq. (1).

The first higher-order term (label cost)

$$h_{\mathbf{f}}(\mathcal{T}_i, \mathbf{f}) = h_{\mathrm{ang}}(\mathcal{T}_i) + h_{\mathrm{lin}}(\mathcal{T}_i) + h_{\mathrm{occ}}(\mathcal{T}_i) + h_{\mathrm{per}}(\mathcal{T}_i) \tag{8}$$

models the plausibility of each trajectory in terms of its dynamics and persistence. The details about each component are given in the following section. The second higher-order term h_X is the pairwise co-occurrence label cost from Eq. (3) for trajectory-level exclusion. Note that during inference both higher-order terms are transformed in each α-expansion step to pairwise ones using auxiliary variables, as outlined in Fig. 3. Also note that the energy can only be minimized approximately: finding a global optimum of E w.r.t. $\mathbf{f}$ in polynomial time is only possible for binary submodular energies with $|\mathbf{L}| \leq 2$.

4. Statistical Data Analysis

Energy minimization offers a flexible framework for modeling in vision, and CRF energies additionally give insight into the dependency structure. But aside from the structure, the potentials also need to be specified appropriately. In many cases the potentials (or energy components) are handcrafted, guided by intuition or mathematical convenience. Arguably, it is beneficial to instead derive their functional form from the statistics of the modeled quantities.

Here, we systematically analyze the distribution of various trajectory properties based on eight video sequences (PETS [14] and TUD-Stadtmitte [2]) with ground truth annotations. It is clear that this comparably small amount of data does not cover all possible tracking scenarios. Rather, the goal here is to allow adapting the tracker to a specific application scenario at hand. With the proposed methodology, other researchers or practitioners can easily adjust the approach to their specific application case.

To construct more realistic energies, we analyze the empirical frequencies of the trajectory properties that we model in our CRF, see Fig. 4. Note that due to the limited amount of available ground truth data for multi-target tracking, full CRF learning is not the goal here. Instead, we derive a suitable functional form of the potentials (thick grey curves in Fig. 4). To that end, we study the negative logarithm of the empirical histograms of each property, following the definition of the Boltzmann distribution.

Localization accuracy of the detector. While it is safe to assume that an object detector will not always localize objects perfectly, the question remains what pattern the deviations follow. Fig. 4(a) shows the (negative logarithm of the) empirical distribution of distances between the detector output and the closest ground truth object on the ground plane. To robustify the estimate, only nearest neighbors within 1m are considered. We observe that the energy grows linearly with the distance, suggesting a linear penalty for the data term (respective exponential distribution)

$$\phi(\mathbf{f}_d, \mathcal{T}) = c_i^{\mathsf{t}} \cdot \|p_i^{\mathsf{t}} - \mathcal{T}_{\mathbf{f}_d}(\mathsf{t})\| \quad \text{for} \quad \mathsf{d} = \mathsf{d}_i^{\mathsf{t}}, \tag{9}$$

which we weight with the detection confidence c_i^{t}.

Angular dynamics. Real objects can only move within physical limits. Here we examine the angular velocity of people from their trajectory. Let $x = x(t)$ and $y = y(t)$ be the coordinates of a parametric planar curve and $\dot{x}, \dot{y}$ and $\ddot{x}, \ddot{y}$ its first and second temporal derivatives, respectively. The angular velocity at time t is then given as

$$\dot{\theta}(t) = \frac{\dot{x}(t)\ddot{y}(t) - \dot{y}(t)\ddot{x}(t)}{\dot{x}(t)^2 + \dot{y}(t)^2}. \tag{10}$$

Note that the definition only applies to regular curves, *i.e.* for $\dot{x}(t)^2 + \dot{y}(t)^2 \neq 0 \; \forall t$. This is not a major limitation, since realistic trajectories will usually have a positive velocity. The distribution of $\dot{\theta}$ in Fig. 4(b) suggests representing the penalty with a Cauchy-Lorentz distribution:

$$h_{\text{ang}}(\mathcal{T}_i) = \lambda_{\text{ang}} \sum_{\mathsf{t}} \log\left(1 + \dot{\theta}(\mathsf{t})^2\right). \tag{11}$$

Linear dynamics. In addition to the angular velocity we also examine the linear velocity. Fig. 4(c) shows that people mostly move at a speed of about $1\frac{\text{m}}{\text{s}}$. Deviations from that speed are rare, such that a quadratic penalty is appropriate:

$$h_{\text{lin}}(\mathcal{T}_i) = \lambda_{\text{lin}} \sum_{\mathsf{t}} \left(\sqrt{\dot{x}(\mathsf{t})^2 + \dot{y}(\mathsf{t})^2} - 1000\right)^2. \tag{12}$$

Occlusion length. In many applications, *e.g.* robotics and some surveillance scenarios, targets are observed from a relatively low camera viewpoint. Hence they are periodically occluded, causing the detector (or any other observation model) to fail temporarily. A tracker should nevertheless be able to bridge such short occlusion gaps without spawning false new trajectories. To determine the expected length of such occlusions, we analyze the frequencies of different durations of occlusion (in frames) as shown in Fig. 4(d). Although most occlusions last less than 20 frames, longer ones do occur. We therefore model the penalty for trajectories that are not supported by detections through multiple consecutive frames as a Cauchy-Lorentz distribution:

$$h_{\text{occ}}(\mathcal{T}_i) = \lambda_{\text{occ}} \sum_{j \in \text{gaps}(\mathcal{T}_i)} \log\left(1 + \gamma_j^2\right). \tag{13}$$

Here, γ_j is the number of frames in which trajectory i has no detections close by.

Persistence and length. Assuming that the scene does not contain doors or other openings where objects might disappear, a trajectory will always start and terminate close to the border of the image (or the tracking area). An extensive data analysis of this property is thus not necessary. To prevent fragmented trajectories and allow a buffer entry zone τ, we impose a soft threshold

$$h_{\text{per}}(\mathcal{T}_i) = \lambda_{\text{per}} \cdot \min\left(\tau, \text{dist}(\mathcal{T}_i^{\mathsf{t}^*}, \text{border})\right), \tag{14}$$

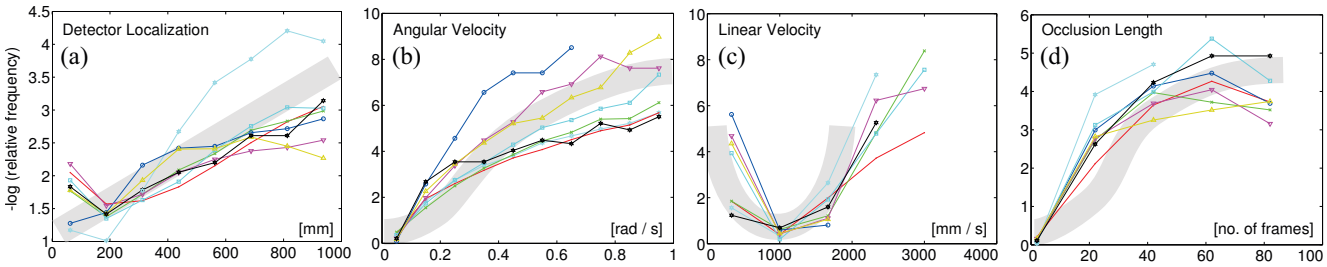

Figure 4. Empirical analysis of various trajectory properties in multiple people tracking, using ground truth data. Thick grey curves denote our suggested models, motivated by their empirical distributions (negative log-frequency shown).

where $\mathtt{t}^*$ stands for birth or death time of a trajectory.

The temporal length of trajectories varies significantly across sequences and does not exhibit a consistent behavior. We therefore do not make any assumptions about it.

5. Implementation

Pruning. To speed up inference we prune the graph in two different ways. We reduce the connectivity by disregarding neighbors from $\mathcal{E}_X$ that lie too far apart. This does not change the CRF energy in the relevant portion of the solution space (*i.e.* near a sensible minimum), because the data term already ensures that such detections will never be assigned the same label. Moreover, the label space of each random variable is reduced to only those (few) trajectory hypotheses that lie within reasonable reach of a detection. Again, this will not change the energy of any remotely plausible solution, for the same reason as above.

Optimization. Like [4, 12] we perform MAP estimation by alternatingly minimizing the energy of the discrete and the continuous variables. The emphasis of this work is on designing a physically and statistically plausible model, with the consequence that the resulting optimization problem becomes harder, even with such an alternation scheme.

Discrete energy minimization is done using α-expansion. Since the energy is non-submodular, we use TRW-S [18] for each binary expansion step. As it is not guaranteed that each expansion step finds a global minimum of the binary sub-problem, we found it beneficial to add a greedy search step in each expansion move: for each label in turn we check whether the energy can be decreased further by entirely removing that label from the current solution (*i.e.* replacing the trajectory by the outlier model). The discrete optimization is implemented using OpenGM [1].

It may seem unnatural to use message passing within α-expansion instead of an st-cut, since message passing algorithms are generally capable of performing inference in multi-label problems. The motivation is that directly running message passing on the multi-label problem is prohibitively slow even for very small graphs due to the global factor in the energy. The factor graph for each expansion move on the other hand is much smaller.

The continuous part of the proposed energy function is not convex and cannot be minimized in closed form. We therefore perform a simplex-based search over the continuous parameters of $\mathcal{T}^*$, starting from a least squares approximation of the objective to find a better minimum. We minimize a simplified energy including only the unary terms ϕ and the continuous label costs h_{ang} and h_{lin}. The solution of this simplified energy minimization step is discarded, if it does not decrease the full CRF energy from Eq. (6). Since the hypothesis space $\mathcal{T}$ is updated in each iteration, the optimization is nevertheless able to escape poor local minima.

The computational effort is similar to the discrete-continuous optimization without mutual exclusion [4]. In practice, we observe running times of 1–2 seconds per frame. Real-time performance appears reachable with a carefully optimized implementation.

6. Experiments

Datasets and metrics. We evaluate our tracker on eight video sequences. Besides the widely used *PETS S2.L1* sequence, we also include four more challenging scenarios from the same dataset. The *PETS* benchmark [14] shows pedestrians walking across an intersection in various directions at variable speed. The number of people varies from a few to as many as 40. In all our experiments, we only use the first camera view point (out of eight recorded). *TUD-Stadtmitte* offers a different setup. Here, a busy pedestrian street is filmed from a low camera angle. In both cases, tracking is performed on a ground plane obtained from camera calibration. Finally, we also test our method on the sequences *BAHNHOF* and *SUNNY DAY* from the ETH Mobile Scene *(ETHMS)* dataset [13], where a busy pedestrian street is filmed from a moving stereo camera. Note that we do not use the available camera calibration and depth maps for these sequences, but rather track people in image space.

A fair quantitative comparison of multi-target tracking methods is challenging. On one hand, tracking by detection critically depends on which object detector is used. On the other hand, different error types and associated error metrics exist, which are not used consistently in the literature. In Tables 1 and 2 we report the widely accepted *CLEAR MOT*

Table 1. Cross-validation results on six sequences.

Method	MOTA	MOTP	MT	ML	FM	ID
baseline 1 [21]	37.4%	64.8%	7	17	104	114
baseline 2 [4]	42.2%	64.1%	11	12	48	65
statistics	45.4%	60.8%	11	12	41	55
det. exclusion	46.7%	63.0%	11	12	38	48
traj. exclusion	46.6%	62.7%	10	12	49	69
combined	51.5%	64.4%	11	13	43	54

[7] metrics evaluated in 3D with a 1m hit/miss threshold. The Multiple Object Tracking Accuracy *(MOTA)* combines all false positives, false negatives, and identity switches into a single number, and Multiple Object Tracking Precision *(MOTP)* measures the average distance between the ground truth and the tracker output. To better assess the quality, we additionally report the numbers of Mostly Tracked (MT, $\geq 80\%$) and Mostly Lost (ML, $\leq 20\%$) trajectories, along with the numbers of track fragmentations *(FM)*, and identity switches *(ID)*. Tab. 3 also shows the numbers for recall and precision. These figures are produced with a 2D evaluation protocol using a publicly available implementation.[2]

6.1. Comparison to a baseline

We systematically compare the individual contributions of our work against two state-of-the-art baselines. To make this comparison as fair as possible, we use publicly available code, ground truth data, and detector evidence throughout our experimentation. *Baseline 1* is a network flow based approach [21] that is solved approximately via dynamic programming. Due to its extraordinary speed, we use the tracklets generated by *baseline 1* as proposal trajectories for our optimization. *Baseline 2* is a recent method based on discrete-continuous optimization [4], but unlike our approach does not properly model inter-object exclusion and uses hand-defined energies that are not derived statistically. We determine the required parameters of all methods by a random search [6] over the parameter space via leave-one-out cross validation. Tab. 1 shows the cross-validation results averaged over all test sequences. We report the results of our full method (*combined*), as well as three intermediate results: only using the statistically motivated energies (*statistics*), adding only the detection-level exclusion factors (*det. exclusion*), and adding only the co-occurrence label cost (*traj. exclusion*). We observe that each individual modeling choice boosts the tracker performance, but it is crucial to handle mutual exclusion at the detection and trajectory level simultaneously, and combined with statistically motivated energies to achieve best possible results; *MOTA* rises by nearly ten percentage points, while the number of identity switches is reduced by $\approx 20\%$. To ease comparison with other approaches, we also give per-sequence results (Tab. 2) using a single set of parameters.

[2] iris.usc.edu/people/yangbo/downloads.html

Table 2. Results of our *combined* method on each test sequence.

Sequence	MOTA	MOTP	MT	ML	FM	ID
S2.L1	90.3 %	74.3 %	18	0	15	22
S2.L2	46.0 %	59.8 %	25	8	105	126
S2.L3	39.8 %	65.0 %	8	19	22	27
S1.L1-2	60.0 %	61.9 %	21	11	19	22
S1.L2-1	26.5 %	60.2 %	6	23	27	34
Stadtmitte	56.2 %	61.6 %	4	0	13	15

Table 3. Quantitative comparison to three state-of-the-art methods on the *ETHMS* dataset [13].

Method	Rcll	Prcn	MT	ML	FM	ID
DP [21]	67.4%	91.4%	50.2%	9.9%	143	4
PIRMPT [19]	76.8%	86.6%	58.4%	8.0%	23	11
Online CRF [24]	79.0%	90.4%	68.0%	7.2%	19	11
Our method	77.3%	87.2%	66.4%	8.2%	69	57

6.2. Further quantitative results

We also evaluate our method on two sequences from the *ETHMS* dataset [13], see Tab. 3. We use the detector output from [19, 24] and the publicly available evaluation script. *DP* is a network flow based approach [21] that is solved approximately via dynamic programming. We slightly tune the parameters for better performance. Due to its extraordinary speed, we use the tracklets generated by *DP* as proposal trajectories for our optimization. State-of-the-art methods for these sequences heavily rely on tracklet linking through significant periods of occlusion, based on appearance and other cues. Since our CRF does not model these, we postprocess our tracker output with a simple extrapolation-based track linking scheme to explore the capabilities of our method when combined with such track linking. While our simplistic linking scheme leads to comparatively many ID switches, the high recall and precision numbers indicate that our discrete-continuous CRF yields a competitive basis for appearance-based occlusion handling.

7. Summary

We proposed a discrete-continuous CRF for multi-target tracking that handles inter-object exclusions at two levels: *(i)* at the data association level based on non-submodular constraints, such that each detection may only explain one target and vice versa; *(ii)* at the trajectory level, where a novel co-occurrence label cost penalizes solutions with overlapping or colliding trajectories. A statistical data analysis was used to derive appropriate CRF potentials. We suggested an expansion move-based optimization scheme to handle the non-submodular energy with global co-occurrence label costs. Our experiments show state-of-the-art results on public benchmarks, with clear improvements from the simultaneous exclusion constraints. Future work may consider incorporating appearance cues into the CRF to better disambiguate targets after long-term occlusions.

Figure 5. Exemplar frames (slightly cropped for display) from the PETS *S2L2* [14], *BAHNHOF*, and *SUNNY DAY* [13] sequences.

Acknowledgments. We would like to thank B. Andres, T. Beier and J. Kappes for releasing OpenGM, as well as for helpful discussions. We also thank T. Pham for pointing out some implementation issues.

References

[1] B. Andres, T. Beier, and J. Kappes. OpenGM: A C++ library for discrete graphical models. *arXiv:1206.0111*, 2012.

[2] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. *CVPR 2010*.

[3] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. *CVPR 2011*.

[4] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. *CVPR 2012*.

[5] J. Berclaz, F. Fleuret, and P. Fua. Multiple object tracking using flow linear programming. *Winter-PETS*, 2009.

[6] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *JMLR*, 13:281–305, 2012.

[7] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *Image and Video Processing*, 2008(1):1–10, 2008.

[8] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. *ICCV 2009*.

[9] W. Brendel, M. Amer, and S. Todorovic. Multiobject tracking as maximum weight independent set. *CVPR 2011*.

[10] I. J. Cox. A review of statistical data association techniques for motion correspondence. *IJCV*, 10(1):53–66, 1993.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR 2005*.

[12] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *IJCV*, 96(1):1–27, 2012.

[13] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. *CVPR 2008*.

[14] J. M. Ferryman and A. Shahrokni. PETS2009: Dataset and challenge. *Winter-PETS*, 2009.

[15] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe. Multi-target tracking using joint probabilistic data association. *IEEE Conf. on Decision and Control*, 1980.

[16] H. Jiang, S. Fels, and J. J. Little. A linear programming approach for multiple object tracking. *CVPR 2007*.

[17] Z. Kalal, J. Matas, and K. Mikolajczyk. P-N Learning: Bootstrapping binary classifiers from unlabeled data by structural constraint. *CVPR 2010*.

[18] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 28(10):1568–1583, 2006.

[19] C.-H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? *CVPR 2011*.

[20] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. *ECCV 2010*.

[21] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. *CVPR 2011*.

[22] J. Vermaak, A. Doucet, and P. Pérez. Maintaining multi-modality through mixture tracking. *ICCV 2003*.

[23] Z. Wu, T. H. Kunz, and M. Betke. Efficient track linking methods for track graphs using network-flow and set-cover techniques. *CVPR 2011*.

[24] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. *CVPR 2012*.

[25] A. R. Zamir, A. Dehghan, and M. Shah. GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs. *ECCV 2012*.

[26] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. *CVPR 2008*.