Improving Global Multi-target Tracking with Local Updates

Anton Milan[1] Rikke Gade[2] Anthony Dick[1]
Thomas B. Moeslund[2] Ian Reid[1]

[1]University of Adelaide, Australia [2]Aalborg University, Denmark

Abstract. We propose a scheme to explicitly detect and resolve ambiguous situations in multiple target tracking. During periods of uncertainty, our method applies multiple local single target trackers to hypothesise short term tracks. These tracks are combined with the tracks obtained by a global multi-target tracker, if they result in a reduction in the global cost function. Since tracking failures typically arise when targets become occluded, we propose a local data association scheme to maintain the target identities in these situations. We demonstrate a reduction of up to 50% in the global cost function, which in turn leads to superior performance on several challenging benchmark sequences. Additionally, we show tracking results in sports videos where poor video quality and frequent and severe occlusions between multiple players pose difficulties for state-of-the-art trackers.

Keywords: Multi-target tracking, data association

1 Introduction

Tracking multiple objects in a dynamic environment is crucial for visual scene understanding. Some of the most relevant applications for this task include driver assistance, visual surveillance, and sports analysis. The problem itself consists of localising each target in every single time instance *as well as* correctly maintaining each target's identity over time. This latter task is often referred to as *data association* and can be solved by existing methods as long as all targets remain sufficiently far apart from one another. However, challenges arise when several targets come close together causing intersecting or intertwined trajectories. In such situations, recovering each individual's identity has a combinatorial complexity in the number of tracks and measurements, and thus quickly becomes infeasible. In addition, the task is complicated further by noisy sensor data with imprecise localisation, false alarms, and missing measurements.

Most current approaches to multi-target tracking are based on *tracking by detection* [1–6]. Here, tracks are formed by linking detections obtained independently in each frame in a preprocessing step. This helps to avoid tracker drift, but usually depends on a pre-defined target model which is trained offline. When tracking by detection, more accurate results have been obtained by so-called *global* methods that consider a batch of several frames (or even an entire

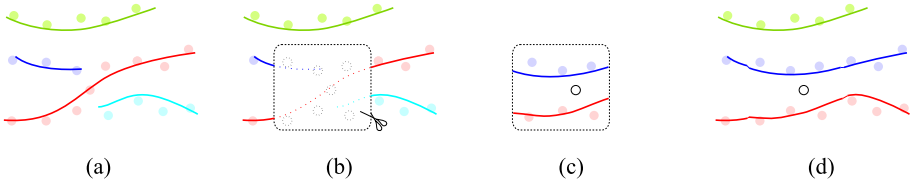

Fig. 1. Overview of our optimisation algorithm. Given a possibly erroneous solution *(a)*, we locate each error *(b)* and perform a local optimisation within its neighbourhood *(c)*. The newly obtained solution is inserted back into the original one if and only if it increases the overall likelihood considering all remaining frames and targets *(d)*.

video sequence) jointly as opposed to determining the state based only on previous observations [7, 8]. The rationale here is that potential ambiguities may be resolved more easily once more evidence is acquired. However one must accept a delay in the output as a tradeoff for better accuracy.

Although tracking by detection approaches achieve state-of-the-art results, they struggle in those situations where the detector provides little to no evidence for the presence of a target. Detector failures may arise for numerous reasons, such as low image contrast, partial or complete occlusions, or abrupt and significant change in appearance due to lighting, posture, or object size. Even though short detection dropouts in certain, unambiguous areas can usually be bridged robustly by global optimisation techniques, correctly resolving data association remains challenging in cases where several targets merge on the image plane obstructing each other's line of sight. Long term occlusions or ambiguities are even more challenging, as the number of feasible association combinations increases with the time interval considered.

We propose to exploit the power of *model-free visual trackers* to 'untangle' tracks in such challenging situations (see Fig. 1 for an illustration). Model-free trackers do not rely on pre-existing detections, instead building an online model of target appearance based purely on an instance of the target appearing in a single frame. The performance of visual object trackers has increased dramatically in recent years [9] making them robust to appearance change and partial occlusion, which is a desirable property for solving the problem at hand. Moreover, we propose a strategy to integrate model-free visual object tracking into a multi-target tracking setting. Although visual trackers have, in one way or another, been previously used in combination with multiple target tracking [10–12], we present a rather different strategy to couple the two approaches.

In particular, our main contributions are as follows:

- We propose a scheme to explicitly detect challenging situations in multi-target tracking and address these in a way that builds on recent progress in both single and multi-target tracking.
- We apply model-free visual trackers to several targets simultaneously in order to resolve difficult situations locally.

- We integrate visual trackers into a multi-target tracking framework to find improved optima of the objective function by making local changes.
- We demonstrate the validity of our approach on particularly challenging sports videos.

We argue that our approach is able to drive the optimisation much quicker towards improved local minima leading to a substantial increase in performance both visually and quantitatively. Our experiments show superior performance on several challenging benchmark sequences.

2 Related Work

The popularity of multi-target tracking in computer vision has increased dramatically in the recent past leading to a large amount of related literature. In this section we will concentrate on the most important work related mainly to offline multi-target tracking approaches. Despite their limitation of a delayed output, offline approaches to multi-target tracking have become increasingly popular due to their superior accuracy. The main difference to online (or recursive) approaches, such as Kalman filters [13, 14] or particle filters [7, 8] is that instead of processing each frame as soon as it is obtained, the optimisation of an objective function is performed on a batch of consecutive frames simultaneously. These methods are usually more robust at dealing with false positives or occlusions.

Offline multi-target tracking. The main difference between approaches lies in the exact formulation of the objective function and its optimisation strategy. Jiang *et al.* [15] solve an integer linear program using LP-relaxation to obtain a (nearly) optimal solution. However, the number of targets in the scene needs to be fixed a-priori. Zhang *et al.* [1] reformulate the task as a network flow problem, which can be solved in polynomial time using min-cost flow algorithms. Occlusions are handled by inserting target hypotheses in a greedy fashion. Their approach served as a starting point for a similar strategy [16], which followed a greedy optimisation scheme and was thus much more efficient. Another globally optimal approach, which explicitly models merged measurements is presented in [17]. Individual tracks are however resolved using a simple shortest paths strategy, which may result in intersecting paths. More recently, Liu *et al.* [5] use a network-flow approach to recover long-term trajectories of sports players using context-aware motion models, while Butt and Collins [18] integrate high-order dynamic terms. A coupling of object detection and tracking has been proposed in [19, 20] with a quadratic and linear objective, respectively. Further formulations to solve for data association include graph-based approaches, such maximum weight independent set [21] set-cover [22] and generalised minimum clique graphs [4].

A slightly different way to solve the task is to concentrate on reconstructing trajectories rather than on data association and only implicitly handle the latter. A regularly discretised space allows one to pose the problem as an integer

linear program, which is solved to global optimality by LP-relaxation [23] or by the k-shortest paths algorithm [2]. To overcome the limitation imposed by the discrete grid, a purely continuous state space is used in [6]. However, such an accurate description of the complex task leads to a highly non-convex optimisation problem which is minimised locally by gradient descent augmented with heuristic discontinuous jumps. A more elegant discrete-continuous energy was later proposed in [24], where both trajectory estimation and data association are handled simultaneously by minimising a single objective.

The main motivation for designing such complex objective functions [19, 21, 6, 24] is to describe the problem at hand as accurately as possible. Although they obtain state of the art results, they are difficult to optimise and often become trapped in local minima. In practice, this manifests in tracking errors such as fragmented trajectories or confused target identities. In this work we focus on overcoming these errors by applying recent results from single target tracking.

Model-free tracking. Recent advances in visual single object tracking [25–27] have also adopted the tracking by detection paradigm. However, rather than train the detector offline, so-called *model-free* trackers train a classifier to separate the target from its background, using positive and negative training examples gathered while tracking. This has the advantage of requiring only initialisation in a single frame and of training a detector specifically for the current appearance of the target. Several methods have been applied to the task, including multiple instance learning [28], structured output learning [26], metric learning [29] and kernel methods [27].

In general, model-free tracking methods are successful over short time periods but their performance degrades over longer time spans, or when target appearance changes significantly. By using them to correct short term errors in long term tracks obtained by global methods, we play to the strength of these two different approaches. Because the model-free tracker operates only on the output of the global tracker, it is independent of its implementation and can therefore be combined with any of the above tracking frameworks. The final result is still obtained by optimising the global objective function; the short term tracks are simply used to generate plausible hypotheses, which the optimiser can use to break out of local minima.

Other recent work has also demonstrated the use of single target visual trackers within multi-target tracking. In [10], contours of multiple objects of arbitrary shape are represented using level-sets and an underlying generative model determines location, depth ordering and segmentation of each target. Similarly, a level-set tracker is also applied in the context of pedestrian tracking from a moving camera in [11], where sparse person detections are augmented with the temporally varying target contours provided by the low-level tracker. Izadinia *et al.* [30] detect pedestrians using the deformable part-based model [31] and in addition to tracking entire people, trajectories of their individual body parts are recovered. In [12], multi-target tracking is based on both, detections from an offline object detector and a visual tracker output. The decision on which cue to

Algorithm 1: Tracking multiple targets by local and global optimisation

input : Initial global solution S (Sec. 3.1)
output: Final trajectories
while $\neg$ *converged* **do**
  Find next error Ξ in current solution S (Section 3.2)
  Optimise locally within spatio-temporal neighbourhood of Ξ to obtain $\hat{S}$ (Sec. 3.3, 3.4)
  Stitch partial solution $\hat{S}$ into global solution S, if global cost is reduced (Sec. 3.5)
end

use is made based on a pre-trained model using several features such as detector response or optic flow. Zhang *et al*. [32] also propose to couple several individual trackers by enforcing to preserve the spatial structure between all targets over time. While this may help to resolve data association in certain cases where objects tend to exhibit similar motion patterns, it is not generally applicable to arbitrary people tracking, in particular sports videos with abrupt and erratic target motion.

Our method is different from previous work in the following aspects: (i) We exploit the power of visual trackers explicitly in difficult situations. To this end, we localise difficult situations in the space-time volume and use the output of multiple coupled single target trackers to generate a strong set of local hypotheses. (ii) We present a local data association scheme for single target trackers. To avoid clumping and identity switching between individual trackers, we follow a simple, yet effective technique based on bipartite graph matching. (iii) We integrate the output of single target trackers into a global energy minimisation method. To avoid potential drift caused by online learned trackers, the local solution is verified in the global context using a robust multiple target objective.

3 Multi-target tracking by energy minimisation

In this work, we follow the recent trend and address multi-target tracking by minimising a highly complex energy function. We use the discrete-continuous formulation proposed in [24]. Note, however, that our method is generic and does not rely on any specific formulation of the underlying objective function.

A weakness of any non-convex global objective is that it may become trapped in local minima, which results in fragmented or incorrectly associated tracks. To remedy this, we propose to focus explicitly on those solution regions that are most likely to be erroneous and to guide the optimisation toward alternative solutions using single target tracking with local data association. The entire algorithm is summarised in Algorithm 1, and the individual steps are illustrated in Figure 1.

We now describe in more detail each of the steps in the algorithm.

3.1 Global data association

In our formulation, multi-target tracking is performed by optimising a discrete-continuous objective, where both data association and trajectory estimation are solved for by minimising a single energy function. Given a set of target detections $\mathbf{D} = d_1, \ldots, d_D$ within a video sequence of F frames, the goal is to find the most likely solution for assigning a unique target identifier to each detection, and at the same time to estimate a continuous trajectory for each target.

Following the notation of [24], we represent the state space by two sets of variables. A discrete set $\mathbf{f} = f_i, \ldots, f_D$ determines the data association, where each variable takes on a label l from the label set $\mathbf{L} = \{1, ..., L, \varnothing\}$ which corresponds to a specific target (or a false alarm). A set of continuous variables $\mathcal{T}$ describes the shape of all trajectories under consideration, where each trajectory is represented by piecewise polynomials.

The discrete part of the energy is posed as a graphical model with unary and pairwise potentials and label costs [33]:

$$E(\mathbf{f}) = \phi_d + \psi_{d,d'} + h_{\mathbf{f}} + h^{\mathsf{X}}_{\mathbf{f}}, \tag{1}$$

while the continuous part controls the trajectories:

$$E(\mathcal{T}) = \phi_{\mathcal{T}} + h_{\mathbf{f}} + h^{\mathsf{X}}_{\mathbf{f}}. \tag{2}$$

In a nutshell, the unary (or data) terms ϕ measure how well the trajectory hypotheses fit the observations, the pairwise terms ψ enforce spatio-temporal smoothness in the labelling, and the label cost models a prior on individual trajectories ($h_{\mathbf{f}}$), such as target dynamics or track persistence, as well as on pairs of tracks ($h^{\mathsf{X}}_{\mathbf{f}}$) to suppress implausible solutions with strongly overlapping trajectories. The complete energy is then minimised by alternately fixing one set of variables at a time, generating the initial solution S. For more details, we refer the reader to [24].

3.2 Error detection

Given an initial solution hypothesis S, our goal is to localise errors within this solution and correct them. Several types of local error may exist, including split tracks, swapped identities and merged trajectories. The importance of each error type is application specific, but to demonstrate our approach, we focus only on the most obvious error type that is also convenient to detect, namely an interrupted trajectory. Under the assumption that the scene does not contain any doors or large scene occluders where people may disappear indefinitely, a target that enters the field of view must ideally remain tracked until it leaves the scene. Therefore, any trajectory $\mathcal{T}_i$ that terminates prematurely and not close to the image border is considered a candidate for improvement. In practice, this is likely to overestimate the number of locations at which errors may occur. This does not detract from the final solution as, in the case of a genuine track

endpoint, all hypothesised track joins are likely to result in a higher overall cost, and therefore the initial solution will be unchanged.

Let $\mathbf{x}_i^t$ be the (x, y) location of target i in frame t. Further, let t_i^* denote either the first or the last frame in which target i exists. An error $\Xi = \{x_i, y_i, t_i^*\}$ is possibly present at the spatio-temporal location $\mathbf{x}_i^{t_i^*}$ if and only if $1 < t < F$ *and* $\beta(\mathbf{x}_i^{t_i^*}) > \tau$. $\beta(\cdot)$ computes the distance to the closest image border and τ is a margin where trajectories are allowed to terminate, which is set to 100 pixels in our experiments.

3.3 Choosing the local optimisation region

To optimise the solution locally, we consider a spatio-temporal window around each detected error. In particular, we optimise over the temporal window $\mathbf{\Omega} = \{t - k, \ldots, t + k\}$, where k is fixed to 10 frames in our experiments. This time span is usually long enough to resolve an ambiguity, but still local enough to rely on the output of a visual tracker.

It remains to determine which existing trajectories are involved in the current error and should be re-estimated within $\mathbf{\Omega}$. On one hand, it is desirable to reduce the current problem to the smallest possible subset to enable efficient optimisation. On the other hand, discarding too many concurrent trajectories may lead to conflicts in the later step when the two solutions are to be merged. In a typical setting, trajectories that are far apart from one another are independent. A reasonable trade off therefore is to only consider a small subset of trajectories $\mathcal{T}^* \subset \mathcal{T}$ which is within a neighbourhood $\mathbf{\Sigma}$ of the error Ξ. To determine the neighbourhood, we create a short auxiliary trajectory $\hat{\mathcal{T}}$ by tracking the target back and forth within the temporal window $\mathbf{\Omega}$, initialised from the error Ξ. To reduce the state space for the optimisation while at the same time not ignoring important dependencies, we consider only those detections d_i that are within a certain radius of $\hat{\mathcal{T}}$ during the local optimisation (*cf*. Fig. 1 (c)). Formally, the set of target candidates is reduced to

$$\hat{\mathbf{D}} = \{d_i | t_i \in \mathbf{\Omega}, \|\mathbf{d}_i - \hat{\mathcal{T}}^{t_i}\| < 2s\}, \tag{3}$$

where $\mathbf{d}_i$ denotes the spatial and t_i the temporal location of detection i, respectively, and s is the target size.

3.4 Local optimisation

In principle, any existing method can be used to find a plausible solution within the spatio-temporal neighbourhood of the detected error. To guide the optimisation into more promising regions, we exploit single target visual trackers in combination with local data association to generate likely trajectory hypotheses. To this end, we initiate a tracker $\mathbb{T}_i$ from each terminating point of each trajectory $\mathcal{T}_i$ in $\mathcal{T}^*$ that is involved in the error Ξ. In our experiments we employ a recent tracker by Henriques *et al*. [27]. We use the implementation distributed by the authors. In practice, its high robustness and speed make it feasible to

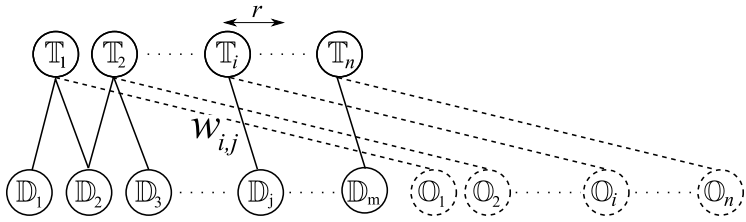

Fig. 2. Example of the bipartite graph that must be solved for each frame. Each tracker $\mathbb{T}_i$ is connected to all detections $\mathbb{D}_j$ that lie within its search radius and to one occlusion node $\mathbb{O}_i$.

quickly generate many short term track hypotheses, although again other single target trackers could also be used. The resulting tracks form a set of strong candidates for selection in the optimisation procedure.

Traditional single target tracking-by-detection algorithms consider only the single detection in each frame with maximum classification score. In order to solve ambiguous situations where several trackers may detect the same target, we extend this approach by including other possible targets. We include all non-overlapping detections whose classification score is more than 10% of the maximum classification score for the individual tracker. The task of local data association is then to optimise the associations between a set of individually trained trackers $\mathbb{T}$ and the set of detections $\mathbb{D}$ for each frame. By representing this association problem in a bipartite graph we are able to find an optimal solution using the Hungarian algorithm. In order to handle occlusions we also introduce an occlusion node for each tracker, which accounts for fully occluded targets. Figure 2 illustrates an example of the bipartite graph for one frame.

Each tracker $\mathbb{T}_i$, $i = 1, \ldots, n$, is initialised, and linked to m detections $\mathbb{D}_1, \ldots, \mathbb{D}_m$ and its respective occlusion node $\mathbb{O}_i$. Edges between trackers and detections with a distance larger than the search radius r have weights zero and are therefore omitted in Figure 2. The size of r is chosen as the mean of the height and width of the target. The weight assigned to each edge combines the appearance measure given by the classification score and a proximity measure that penalises large spatial jumps between consecutive frames:

$$w_{i,j} = s_{i,j} \cdot p_{i,j}, \tag{4}$$

where $s_{i,j}$ is the classification score for tracker i evaluated on target j, scaled to $[0, 1]$. $p_{i,j}$ is a linear proximity measure between the last detection of tracker i and target j and is defined as

$$p_{i,j} = \frac{1}{r} \max\left(0, r - \|\mathbb{T}_i^{t-1} - \mathbb{D}_j\|\right). \tag{5}$$

The proximity measure is used as a simple random walk motion model. Particularly in sports the motion may be abrupt, therefore, we choose this zero displacement model rather than assuming constant velocity.

Edges connecting a tracker to its occlusion node are assigned a low weight, which is empirically chosen as 8% of the maximum classification score in order to be lower than real detections.

3.5 Combining local and global solutions

To stay consistent with the overall formulation, we minimise the same discrete-continuous objective function as is used to evaluate the quality of the complete solution on the spatio-temporal subset $\{\mathbf{\Sigma}, \mathbf{\Omega}\}$ using track hypotheses from Section 3.4. After the optimisation, the resulting solution $\hat{S}$ replaces the original solution within $\{\mathbf{\Sigma}, \mathbf{\Omega}\}$ if the overall energy $E(\hat{S} \cup \tilde{S})$ is decreased. $\tilde{S}$ is obtained by simply removing all partial trajectories from S that lie within the spatio-temporal neighbourhood $\{\mathbf{\Sigma}, \mathbf{\Omega}\}$ of the error.

4 Experiments

Datasets. We demonstrate our approach on eight different sequences. The first set consists of six publicly available videos including the PETS 2009 benchmark [34][1] and TUD Stadtmitte [35]. All videos show pedestrians in a single view but they exhibit a large variation in person count, camera viewpoint and motion patterns. Since the camera calibration is available for this dataset, we perform tracking on the ground plane in world coordinates.

As well as evaluating on standard benchmarks, we also demonstrate the performance on difficult sports tracking data. In particular we show tracking on two sequences in the challenging sport of Australian Rules Football (AFL), in which there is regular and frequent crowding of players and contact between them, making it a very difficult tracking problem. We make this new dataset including the ground truth annotations and the detections used in this work publicly available[2].

Metrics. Quantifying performance of multiple target tracking is a notoriously difficult task [36]. Ambiguities in annotations, assignments strategies and metric descriptions prohibit a purely objective evaluation. Here we follow the most widely used strategy and report several metrics for all our experiments. Next to standard precision and recall figures we report the CLEAR MOT metrics [37], which consists of tracking accuracy (MOTA) and tracking precision (MOTP). The former combines three error types: false positives, missed targets and identity switches, into a single number such that zero errors corresponds to 100%. The latter measures the localisation error of the tracker w.r.t. the annotated ground truth. Moreover, we also show the number of correctly recovered trajectories as proposed in [38]. A target is considered mostly tracked (MT), if it is correctly detected in over 80% of frames within its time span. Similarly, a mostly

[1] Sequences: S2L1, S2L2, S2L3, S1L1-2, S1L2-1

[2] http://research.milanton.net/data

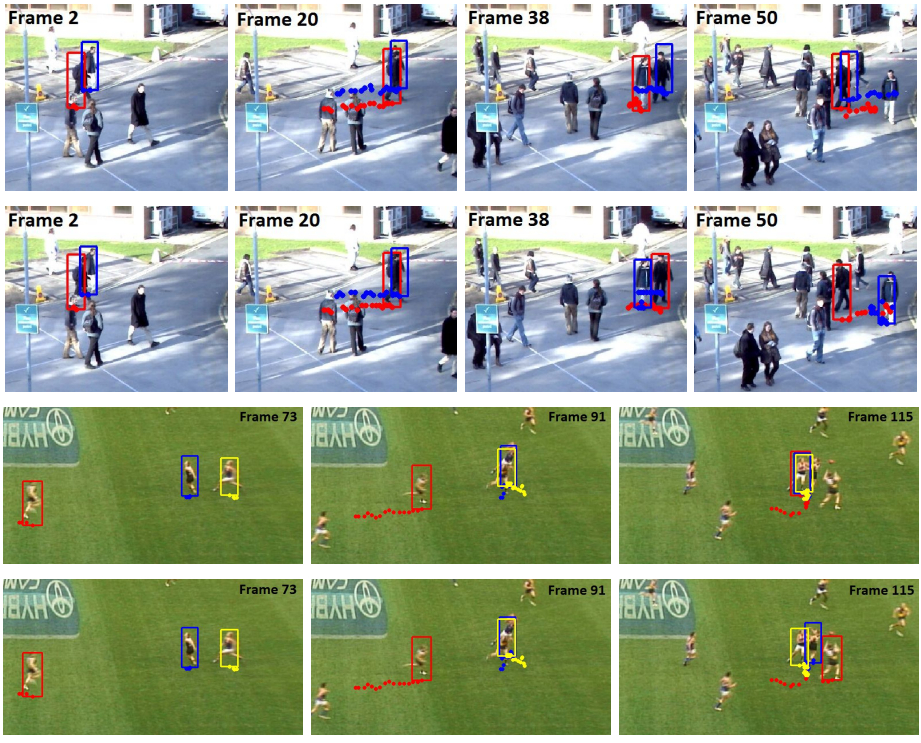

Fig. 3. Visual trackers without (1st and 3rd rows) and with (2nd and 4th rows) local data association. See text for details.

lost (ML) trajectory is only recover in 20% of frames or less. Finally, the numbers of track fragmentations and identity switches are stated for completeness.

Before presenting the overall tracking performance of our system, we discuss the importance of local data association for single target trackers and illustrate the potential of our locally driven optimisation scheme measured by the reduction of the total cost. We then provide an extensive quantitative evaluation on various challenging sequences and compare our results to several state-of-the-art methods.

4.1 Local data association

Let us first qualitatively demonstrate the effect of local data association using multiple model free trackers in situations with a high presence of occlusions. Figure 3 shows two comparisons between tracking with and without local data association. The first sequence is from the PETS 2009 S2L2 dataset, and the second one is a challenging situation from an AFL game. The images are cropped for better visibility. In all cases model-free trackers are initialised for each target depicted with bounding boxes in the left most image. The 1st and 3rd rows

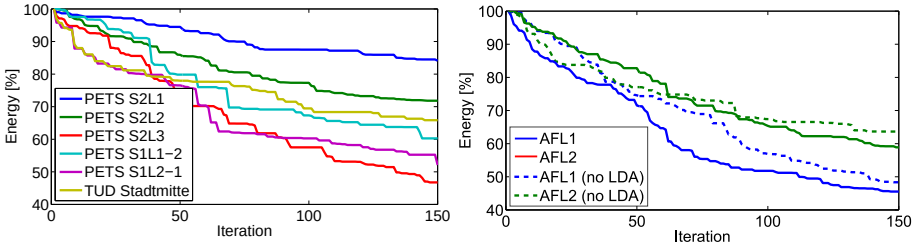

Fig. 4. Minimising a global energy function by focusing on local optimisation windows. The dotted plots on the right hand side depict the energy minimised by our scheme by only using independent single target trackers *without* local data association (no LDA).

show the results of running the two, respectively three trackers individually, without local data association. In rows two and four the results are obtained by including the Hungarian data association described in Section 3.4. The results show clearly that the individual trackers are prone to drift in settings with multiple persons. In row 1 the identity switches and the blue target is lost after occlusion. By including local data association these situations are resolved, and the two targets are correctly tracked even after full occlusions. In the 3$^{\text{rd}}$ row without data association the three trackers clump together and follow the same person which yields the highest classifier score for each of the trackers after the occlusion. The local data association shown in the 4$^{\text{th}}$ row again resolves this situation and keeps tracking the three individual persons while maintaining their correct identities.

To quantify the importance of multiple tracker reasoning, we minimise the global objective with and without explicit local data association (no LDA) for proposal generation. Experimental results are reported in the following sections.

4.2 Energy minimisation

To verify the potential of our approach, we compare the magnitude of the initial solution S of the overall energy to the final solution obtained after including local tracks. Figure 4 shows the relative energy decrease for various sequences. The energy is scaled in each case such that the initial point, which is obtained by [24], corresponds to 100%. It is important to note that we minimise the exact same energy without introducing new detections. By focusing on erroneous regions and by exploiting model-free trackers for better hypothesis generation, our proposed local optimisation can find a lower global cost in nearly every iteration and an overall reduction of over 50% in some cases. Dotted lines for the AFL sequences show the energy reduction using proposals of independent single target trackers without local data association. One iteration takes approximately one second to compute on a standard PC. We set the maximum number of iteration to 150 in all our experiments.

Table 1. Quantitative results on two AFL sequences. Best result across all methods is highlighted in bold face for each measure.

Method	MOTA	MOTP	MT	ML	Frag.	ID sw.	Precision	Recall
FFP Detector [39]	–	–	–	–	–	–	65.4%	55.0%
SMOT [40]	16.7%	60.8%	2	3	**38**	**14**	59.8%	52.0%
DCO [24]	29.7%	63.3%	3	**2**	93	97	70.9%	56.3%
ours (no init)	32.0%	**64.1%**	6	**2**	54	54	67.4%	64.5%
ours (no LDA)	39.0%	63.6%	6	**2**	44	27	72.1%	64.2%
ours (full)	**41.4%**	63.6%	**7**	**2**	39	22	**73.2%**	**65.8%**

4.3 Quantitative evaluation

AFL sports data. We first demonstrate quantitative performance of our approach on the two sports sequences. To obtain candidate detections, we trained a person detector based on fast feature pyramids [39] using only one single image as training data resulting in moderate precision and recall. Table 1 shows the detector's performance as well as tracking results from two recent multi-target tracking methods. The similar-appearance multiple object tracker (SMOT) [40] is specifically designed to address situations shown in these sports sequences with similar target appearance by relying only on motion similarity and using a generalised linear assignment to reconstruct long-term tracks. While this method shows excellent performance with no or little detection noise, it struggles to correctly infer plausible trajectories in a realistic challenging setting. The second baseline is a recent energy minimisation-based method (DCO) [24], which can eliminate many false positive detections. However, due to the complex formulation of its objective, the optimisation reaches only a moderate local minimum with many short tracks leading to a high number of interrupted trajectories and identity switches.

The second part shows three variants of our proposed method. The first one (no init) is our optimisation strategy starting from the trivial solution, where each detection is considered an error (or equivalently a single-frame track). Note that we are able to outperform other methods by appying our customized optimisation scheme. The second strategy (no LDA) uses [24] as initialisation but does not involve local data association for hypothesis generation as described in Section 3. Finally, by applying our full method using localised optimisation with visual trackers, we are able to further minimise the objective function, which is also reflected in the superior tracking performance.

Public benchmark. Our second set of experiments involves a public tracking benchmark. Table 2 shows a quantitative comparison of our proposed strategy to previous methods: A network flow-based approach solved with dynamic programming (DP) [16], globally optimal tracking on a discrete grid (KSP) [2] and the same energy minimisation formulation as before [24]. All numbers are computed using code provided by the authors, publicly available detections, ground

Table 2. Comparison to previous methods on a standard benchmark (PETS, TUD). The results are averaged over six sequences.

Method	MOTA	MOTP	MT	ML	Frag.	ID sw.	Precision	Recall
HOG/HOF Det. [41, 42]	–	–	–	–	–	–	79.5%	62.2%
DP [16]	46.0%	**64.7%**	8	11	165	204	91.7%	55.8%
KSP [2]	41.7%	62.8%	8	20	**10**	**18**	91.6%	46.8%
DCO [24]	55.7%	63.6%	11	**9**	49	43	93.0%	61.6%
ours	**56.9%**	64.1%	**13**	10	40	48	**93.4%**	**62.8%**

truth and evaluation scripts[3]. Note that the slightly higher absolute number of ID switches is a result of incorrectly bridging or extending interrupted trajectories. However, the positive effect of recovering more tracks (MT) and thereby increasing the recall outweighs, yielding higher overall accuracy.

Although we outperform state-of-the-art methods on this benchmark, the improvement is less prominent than in the AFL case. One reason for this behaviour may be that the detection quality is poorer on the sports sequences due to large deformations and small target size (*cf*. Tab. 1 and Fig. 5), yielding a more complex optimisation problem with more local minima. It is also possible that [24] finds a solution much closer to the global optimum on the public sequences, which may indicate that it is well suited to the benchmark but shows limitations on novel data.

4.4 Qualitative results

Finally, Figure 5 illustrates qualitative results on three sequences. Each row shows three frames from AFL1, AFL2, and PETS S2L2, respectively. Note that our method is able to correctly identify nearly all targets even in extremely challenging conditions with substantial levels of multiple occlusions. Also note how the potential of using visual model-free trackers within a traditional multi-person tracking setting is unfolded in situations with extensive pose variation, such as demonstrated by the cyan (ID 33) and the blue (ID 20) targets in the first and second row, respectively. Please refer to the supplemental video for further visual results.

5 Conclusion

We proposed a simple yet effective method to optimise highly complex objectives for multiple target tracking by focusing explicitly on correcting errors locally. A local data association technique combined with a set of visual object trackers is able to drive the optimisation into much better minima reducing the energy by over 50% and consequently leading to superior solutions. We demonstrate the

[3] Note that the corrected numbers are reported for [24], which differ from the original publication.

Fig. 5. Exemplar frames from two AFL clips and the PETS S2L2 sequence. Note that using model-free trackers allows one to maintain the identity of a player even during severe deformations and pose changes (*cf*. the blue target (ID 20) in the second row).

validity of our approach on particularly challenging sports sequences and public benchmark data achieving state of the art performance.

In future work we plan to more thoroughly investigate different error types and their influence on the final solution. It may also be possible to design even more accurate and more complex objective functions that better approximate the true state but still remain tractable using our local optimisation strategy.

Acknowledgements

We gratefully acknowledge the financial support of the Australian Research Council through Laureate Fellowship FL130100102 to IDR.

References

1. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR 2008
2. Berclaz, J., Fleuret, F., Türetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. IEEE T. Pattern Anal. Mach. Intell. **33**(9) (September 2011) 1806–1819

3. Yang, B., Nevatia, R.: An online learned CRF model for multi-target tracking. In: CVPR 2012. 2034–2041
4. Zamir, A.R., Dehghan, A., Shah, M.: GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs. In: ECCV 2012. Volume 2. 343–356
5. Liu, J., Carr, P., Collins, R.T., Liu, Y.: Tracking sports players with context-conditioned motion models. In: CVPR 2013. 1830–1837
6. Milan, A., Roth, S., Schindler, K.: Continuous energy minimization for multitarget tracking. IEEE T. Pattern Anal. Mach. Intell. **36**(1) (2014) 58–72
7. Vermaak, J., Doucet, A., Pérez, P.: Maintaining multi-modality through mixture tracking. In: ICCV 2003. 1110–1116
8. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV 2009
9. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: CVPR 2013. 2411–2418
10. Bibby, C., Reid, I.: Real-time tracking of multiple occluding objects using level sets. In: CVPR 2010. 1307–1314
11. Mitzel, D., Horbert, E., Ess, A., Leibe, B.: Multi-person tracking with sparse detection and continuous segmentation. In: ECCV 2010. Volume 1. 397–410
12. Yan, X., Wu, X., Kakadiaris, I.A., Shah, S.K.: To track or to detect? An ensemble framework for optimal selection. In: ECCV 2012. Volume 5. 594–607
13. Kalman, R.E.: A new approach to linear filtering and prediction problems. Transactions of the ASME–Journal of Basic Engineering **82**(Series D) (1960) 35–45
14. Julier, S.J., Uhlmann, J.K.: A new extension of the kalman filter to nonlinear systems. In: International Symposium on Aerospace andDefense Sensing, Simulation and Controls. (1997) 182–193
15. Jiang, H., Fels, S., Little, J.J.: A linear programming approach for multiple object tracking. In: CVPR 2007
16. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR 2011
17. Henriques, J.a., Caseiro, R., Batista, J.: Globally optimal solution to multi-object tracking with merged measurements. In: ICCV 2011
18. Butt, A.A., Collins, R.T.: Multi-target tracking by lagrangian relaxation to min-cost network flow. In: CVPR 2013
19. Leibe, B., Schindler, K., Van Gool, L.: Coupled detection and trajectory estimation for multi-object tracking. In: ICCV 2007
20. Wu, Z., Thangali, A., Sclaroff, S., Betke, M.: Coupling detection and data association for multiple object tracking. In: CVPR 2012
21. Brendel, W., Amer, M.R., Todorovic, S.: Multiobject tracking as maximum weight independent set. In: CVPR 2011
22. Wu, Z., Kunz, T.H., Betke, M.: Efficient track linking methods for track graphs using network-flow and set-cover techniques. In: CVPR 2011
23. Berclaz, J., Fleuret, F., Fua, P.: Multiple object tracking using flow linear programming. In: 12th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (Winter-PETS). (December 2009)
24. Milan, A., Schindler, K., Roth, S.: Detection- and trajectory-level exclusion in multiple object tracking. In: CVPR 2013
25. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N Learning: Bootstrapping binary classifiers from unlabeled data by structural constraint. In: CVPR 2010
26. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. In: ICCV 2011. 263–270

27. Henriques, J., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. In: ECCV 2012. Volume 4. 702–715
28. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: CVPR 2009
29. Li, X., Shen, C., Shi, Q., Dick, A., Hengel, A.v.d.: Non-sparse linear representations for visual tracking with online reservoir metric learning. In: CVPR 2012. 1760–1767
30. Izadinia, H., Saleemi, I., Li, W., Shah, M.: $(MP)^2T$: Multiple people multiple parts tracker. In: ECCV 2012. Volume 6. (2012) 100–114
31. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE T. Pattern Anal. Mach. Intell. **32**(9) (2010) 1627–1645
32. Zhang, L., van der Maaten, L.: Structure preserving object tracking. In: CVPR 2013. 1838–1845
33. Delong, A., Osokin, A., Isack, H.N., Boykov, Y.: Fast approximate energy minimization with label costs. Int. J. Comput. Vision **96**(1) (January 2012) 1–27
34. Ferryman, J., Shahrokni, A.: PETS2009: Dataset and challenge. In: 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS). (December 2009)
35. Andriluka, M., Roth, S., Schiele, B.: Monocular 3D pose estimation and tracking by detection. In: CVPR 2010
36. Milan, A., Schindler, K., Roth, S.: Challenges of ground truth evaluation of multi-target tracking. In: 2013 IEEE CVPR Workshops (CVPRW). (June 2013) 735–742
37. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The CLEAR MOT metrics. Image and Video Processing **2008**(1) (May 2008) 1–10
38. Li, Y., Huang, C., Nevatia, R.: Learning to associate: Hybridboosted multi-target tracker for crowded scene. In: CVPR 2009
39. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE T. Pattern Anal. Mach. Intell. (2014) To appear.
40. Dicle, C., Sznaier, M., Camps, O.: The way they move: Tracking multiple targets with similar appearance. In: ICCV 2013
41. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR 2005. 886–893
42. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: CVPR 2010