Online Multi-Object Tracking with Structural Invariance Constraint

Xiao Zhou[1,2]
zx2962@stu.xjtu.edu.cn
Peilin Jiang[1,2]
pljiang@mail.xjtu.edu.cn
Zhao Wei[1,2]
aznikline@gmail.com
Hang Dong[1,2]
dhunter1230@gmail.com
Fei Wang[1,2]
wfx@mail.xjtu.edu.cn

[1] National Engineering Laboratory for Visual Information Processing and Application, XJTU, 99 Yanxiang Road, Xi'an, Shaanxi 710054, China

[2] School of Software Engineering, XJTU, 28 West Xianning Road, Xi'an, Shaanxi 710049, China

Abstract

Under the framework of tracking-by-detection, data association is one of the most important issues in multi-object tracking (MOT). Given a video sequence labeled with bounding boxes, data association aims at adopting a graph matching or network flow to maximize (minimize) the sum of the association probabilities (costs), which are generally elaborated by objects' appearance features and motion cues. However, both analogous appearances and moving cameras inevitably increase matching ambiguity and thus make data association intricate and challenging. In this paper, we propose a new data association method to address the online MOT problem by exploiting structural invariance constraint, which is insensitive to both akin appearances and dynamic camera situation. Furthermore, we develop a total probability frame that is able to jointly reason on both appearance and structure cues without adjusting parameters manually. We evaluate our online multi-object tracking algorithm on public MOT Challenge datasets and achieve comparable performance with other state-of-the-art approaches.

1 Introduction

With the growing accuracy of object detection, the framework of tracking-by-detection has been broadly used in addressing multi-object tracking (MOT) problems. In this framework, data association is of great significance and thus arouses widespread concern. According to different inputs, data association is generally categorized into three types: detections-to-detections, tracklets-to-tracklets and detections-to-tracklets. Aiming to join detections into tracklets, the first type handles the association from detections to detections in two adjacent frame images, which belongs to the low-level offline MOT methods under the hierarchical association framework [18]. The second type addresses the matching problem of different tracklets, which is the primary aspect on the study of offline MOT algorithms. The third one,

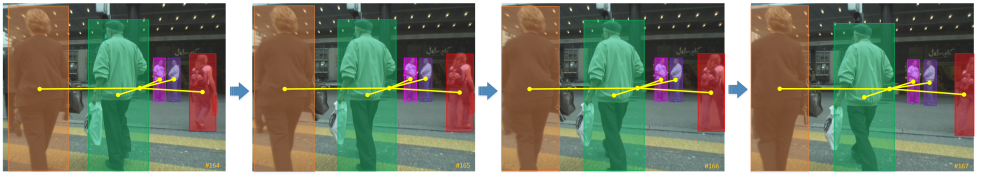

Figure 1: Invariant structure among multiple targets between adjacent frame images from the video sequence ETH-Crossing [23] captured by a moving camera. The yellow lines with a same intersection represent the star-shaped structure among targets. Obviously, the structures in adjacent frames approximately remain invariant and consistent, which inspires us to propose a new data association approach by exploiting the structural invariant constraints.

linking current detections to historical tracklets, is mainly utilized in online MOT approaches to extend tracklets along the timeline.

A typical framework of data association involves contriving pairwise association probability (cost) either manually or automatically [32] and then adopting a graph matching or network flow to maximize (minimize) the sum of association probabilities (costs). Generally, appearance features and motion cues of targets are extracted to construct pairwise association probabilities. Since appearance models can distinguish among different objects very precisely in most circumstances, many studies are dedicated to devising appearance features as discriminatively as possible [2, 6, 33]. However, the appearance information would become unreliable and even invalid when tracking objects with similar looks, for instance, players on the pitch. In such situation, various motion patterns can be adopted to separate different objects. On the one hand, when dealing with a video sequence captured by a fixed camera, the movement of each target is predictable and thus can be formulated as a linear motion model or an autoregressive model [11]. On the other hand, under dynamic camera situation, the movement of each target becomes ambiguous and unpredictable since it involves not only the object's motion itself but also the offset evoked by the fluctuation of the camera. As a result, the conventional motion models fail to describe objects' movements, let alone accurately predicting the positions of missing targets.

Fortunately, the structure of multiple targets between adjacent frame images can nearly remain invariant even in dynamic camera situation. Shown as in Figure 1, we develop a novel representation to measure structural similarities before and after association procedure by exploiting structural invariance constraint. Besides, when coping with multiple cues, previous research usually combines them into an association similarity (cost) linearly and thus needs arduous weight adjustment to obtain better association results. In this work, we propose a total probability frame that couples both appearance features and structure cues independently of any weights, thus avoiding tedious parameters adjustment. We evaluate our proposed online MOT approach using a variety of challenging dataset and achieve comparable performance with state-of-the-art methods. The rest of this paper is organized as follows. We first discuss related work in Section 2. The structural invariance constraint and the total probability frame are described in Section 3. Section 4 presents experimental results.

2 Related Work

Firstly, we review related MOT approaches that pay much attention to motion and structure cues of multiple targets. Andriyenko et al. [1, 28, 29] address data association and trajectory estimation by formulating a discrete-continuous energy function that models many aspects of multi-object tracking. Dicle et al. [11] adopt an autoregressive model to characterize the trajectories of multiple targets that have similar appearances and complicated movements. This approach is excessively dependent on motion cues and thus fail to distinguish objects with ambiguous movement but discriminative appearance.

Yoon et al. [38] utilize the motion context to construct a relative motion network to cope with the unexpected camera motions. However, this method cannot handle abrupt camera motions and fluctuations. In [17], Yoon et al. exploit the structural motion constraints and propose an event aggregation approach to address the MOT problem with moving cameras. This method shows great performance in public datasets.

Apart from motion features and structure cues, two alternatives are usually employed to make a tracker more sophisticated. On the one hand, many works are dedicated to finding more elaborate tools for addressing data association assignment, including minimization of network flow based cost [7, 8, 10, 16], linear programming [11, 19], Hungarian matching [6, 17, 18] and subgraph decomposition [9, 35]. On the other hand, exploiting more discriminative appearance features [2, 3, 39] is increasingly favored. Besides, Rezatofighi et al. [15] modify the association costs with joint probabilities, and decompose the original problem into a series of integer programming, which is more efficient and time-saving.

With the rapid development of deep learning, many MOT algorithms based on convolutional neural networks [22, 34, 36], and recurrent neural networks [30] have been proposed. Due to the powerful capability of learning and extracting image features, these CNN or RNN based approaches outperform most conventional methods with hand-crafted features.

3 Pairwise Association Probability under Structural Invariance Constraint

3.1 Notation

The trajectory of a target is represented by a series of center points of bounding boxes. We denote the kth object in the $t-1$th frame as $\mathcal{O}_k^{t-1}$ and use $\{\mathcal{D}_1^t, \mathcal{D}_2^t, ..., \mathcal{D}_M^t\}$ to represent detections in frame t. A detection in the current frame will be relabeled as $\mathcal{O}_k^t$ if it manages to be associated with the historical object k. We model the structure of all targets in frame t as a star-shaped structure, in which the central node $\bar{\mathcal{O}}^{t-1}$ is described by the spatial center of all targets. Intuitively, we introduce the relative displacement of each target, denoted by $\Delta\mathcal{O}_k^{t-1}$, as the edge of the star-shaped structure, which is computed as:

$$\Delta\mathcal{O}_k^{t-1} = \mathcal{O}_k^{t-1} - \bar{\mathcal{O}}^{t-1}, \bar{\mathcal{O}}^{t-1} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{O}_k^{t-1} \tag{1}$$

Where, N denotes the number of targets. The matching scheme between historical objects and current detections can be formalized as a binary association matrix Φ (or $\mathcal{NA}$, $\mathcal{SA}$). $\Phi(i,j)=1$ indicates that the ith detection manages to be associated with the jth target. Likewise, $\Phi(i,j)=0$ suggests the detection i fails to be associated with the target j.

Due to false alarms and new/leaving targets, a detection is allowed to be associated with one historical target at most and vice versa, which can be formalized as the following constraints:

$$\sum_i \Phi(i,j) \leqslant 1, \sum_j \Phi(i,j) \leqslant 1 \tag{2}$$

3.2 Structural invariance constraint

Ideally, structural invariance constraint demands current detections and historical targets to satisfy one-to-one mapping and the star-shaped structure remains invariant before and after the one-to-one mapping. However, owing to detection noises as well as the situation of targets entering/leaving the field of the view, detections and objects fail to match one-to-one in most circumstances. Therefore, the structure varies regardless of how association procedure runs. Correspondingly, finding the match scheme with the minimal structural variation is the key to addressing the association. It can be described as:

$$\mathcal{R}^* = arg\min_{\mathcal{R}} (StructuralCost(\mathcal{R})) \tag{3}$$

Where, $StructuralCost(\bullet)$ denotes cost function measuring structural variation. $\mathcal{R}$ represents available association schemes between detections and targets. Evidently, the size of solution space is correlated to the problem scale. Specifically, for an association problem between M detections and N targets without considering false positives, there are $N!$ feasible matching results when $M = N$. Furthermore, there are $\frac{N!}{(N-M)!}$ and $\frac{M!}{(M-N)!}$ available solutions when $M < N$ and $M > N$, respectively. Therefore, when we track a large number of objects, traversing all solutions is NP hard and impractical.

3.3 Structural association probability

Restricted by structural invariance constraint, association assignment aims to match detections with targets as much as possible, but at minimal structural variation cost. We define this as structural association. Since the solution space of association assignment is extremely huge, we compute structural association probability (SAP) for each detection-object pair and then maximize the sum of the probabilities to obtain the optimal association event. By doing this, we can avoid searching all available results for the optimal association to achieve minimal structural variation cost.

We denote the structural association probability between the ith detection and the jth target as $P\{\mathcal{SA}(i,j) = 1\}$, which is proportional to its contribution to the invariance of the whole structure. In other words, a detection-object pair facilitating the structural invariance is more likely to be associated than one aggravating the structural alteration.

To obtain pairwise SAP, naive association is defined as a matching assignment that is constrained by "there is one and only one object to match with each detection". Measuring how likely a detection-object pair could be associated naively, naive association probability (NAP) of each detection-object pair is assumed to be independent and proportional to their appearance similarity. Supposing there are M detections in the current frame and N historical targets, the NAP between the mth detection and the nth target can be obtained by:

$$P\{\mathcal{NA}(m,n) = 1\} = \frac{\xi_{mn}}{\sum_{n=1}^{N} \xi_{mn}} \tag{4}$$

Where, $\mathcal{NA}(m,n)=1$ indicates that the mth detection is naively associated with the nth object. ξ_{mn} denotes their appearance similarity. According to the definition of naive association, for the mth detection, $\{\mathcal{NA}(m,n)\,|n=1,2,...,N\}$ constructs collectively exhaustive events. Therefore, based on the law of total probability, we have:

$$P\{\mathcal{SA}(i,j)=1\}=\sum_{n=1}^{N}P\{\mathcal{SA}(i,j)=1|\mathcal{NA}(i,n)=1\}P\{\mathcal{NA}(i,n)=1\} \tag{5}$$

To suppress computational noises, we rewrite the Eq.(5) as:

$$P\{\mathcal{SA}(i,j)=1\}=\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}P\{\mathcal{SA}(i,j)=1|\mathcal{NA}(m,n)=1\}P\{\mathcal{NA}(m,n)=1\} \tag{6}$$

Where, $P\{\mathcal{SA}(i,j)=1|\mathcal{NA}(m,n)=1\}$ represents conditional SAP. The conditional structural association suggests that under the condition of a naive association, $\mathcal{NA}(m,n)=1$ for instance, we implement association procedure between rest detections and objects while minimizing structural variation cost.

We use Φ_{mn} to denote the final association scheme when given a condition pair ($\mathcal{NA}(m,n)=1$). Intuitively, if the final scheme contains a detection-object pair (i,j), i.e. $\Phi_{mn}(i,j)=1$, it can be concluded that the pair (i,j) is contributed to maintaining the structural invariance under the condition of $\mathcal{NA}(m,n)=1$. On the contrary, $\Phi_{mn}(i,j)=0$ indicates that the contribution of pair (i,j) to the structural invariance is very limited. Reasonably, the conditional SAP can be devised as a measure of the contribution to the structural invariance:

$$P\{\mathcal{SA}(i,j)=1|\mathcal{NA}(m,n)=1\}=\begin{cases}\frac{sum(\Phi_{mn})-1}{N-1} & \Phi_{mn}(i,j)=1\\ 0 & \Phi_{mn}(i,j)=0\end{cases} \tag{7}$$

Where, Φ_{mn} satisfies constraints exhibited in Eq.(2) and $\Phi_{mn}(m,n)=1$ as well. $sum(\Phi_{mn})-1$ denotes the number of associated targets apart from the given one in the final association scheme. N represents the number of all historical targets.

So far, once the conditional association scheme Φ_{mn} for each condition pair (m,n) is gained, we can easily calculate the conditional SAP and thus obtain the SAP of each available detection-object pair (i,j) by adopting the Eq.(7) and (6). To address the problem of conditional structural association, we propose a heuristic matching approach, shown in Algorithm 1. The core of this approach is that an available pair provoking the minimal structural variation would be associated preferentially. In specific, we first categorize historical objects into two types: matched and unmatched objects. Then, the positions of unmatched objects in the current frame are estimated from matched objects' positions by minimizing structural variation cost. Next, we search for the nearest pair between detections and objects' prediction, and make a decision: if the intersection over union (IoU) of the pair exceeds a certain threshold, the object would be relabeled as a matched object and the algorithm enters into the next search loop; otherwise, exit the search, end the association, and output the matching scheme, which is shown schematically in Figure 2.

Prediction of unmatched objects' positions in the current frame can be estimated by minimizing structural variation cost, shown in Figure 3. As described in section 3.1, objects' positions need to be normalized as relative coordinates. As a result, the origin of coordinates in each frame is settled by the central position of all targets, denoted by $\bar{\mathcal{O}}^t$. $\Delta\mathcal{O}_i^t$ represents

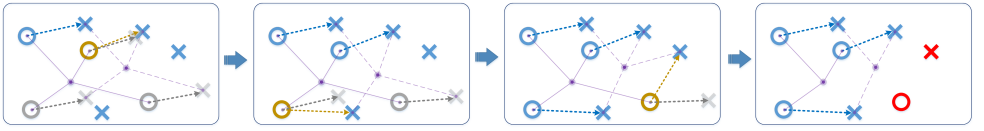

Figure 2: Heuristic matching approach under structural invariance constraint. Blue and gray circles represent matched and unmatched targets, respectively. Gray crosses denote prediction of unmatched targets' positions in the current frame image estimated by minimizing structural variation cost. Detections in the current frame are marked by blue crosses. Solid lines with a same intersection construct the star-shaped structure of targets. Each dash line with an arrow, bridging from an object to a detection, denotes the movement of the object. The brown target coupled with a current detection in each graph represents the nearest pair between detections and the prediction of objects' positions. The first three graphs suggest that when the IoU of the nearest pair exceeds a certain threshold, the object (brown circle) is relabeled as a matched object (blue circle) in the next graph. In the last graph, as the IoU of the nearest pair does not reach the threshold, the object in this pair is relabeled as a missing target, marked by red circle, and the detection is regarded as a new target candidate, marked by a red cross. Correspondingly, the association procedure is ended and the final association scheme is denoted by blue dash lines with arrows.

coordinates of the ith object relative to $\bar{\mathcal{O}}^t$ in frame t, which is calculated by $\Delta\mathcal{O}_i^t = \mathcal{O}_i^t - \bar{\mathcal{O}}^t$. Minimizing the structural variation cost is formalized as:

$$\min_{\hat{\mathcal{O}}_j^t} \sum_{i\in\Omega_M} \|\Delta\mathcal{O}_i^t - \Delta\mathcal{O}_i^{t-1}\|^2 + \sum_{j\in\Omega_U} \|\Delta\mathcal{O}_j^t - \Delta\mathcal{O}_j^{t-1}\|^2 \tag{8}$$

$$\Delta\mathcal{O}_j^t = \hat{\mathcal{O}}_j^t - \bar{\mathcal{O}}^t,\ \bar{\mathcal{O}}^t = \frac{1}{N}\left(\sum_{i\in\Omega_M} \mathcal{O}_i^t + \sum_{j\in\Omega_U} \hat{\mathcal{O}}_j^t\right) \tag{9}$$

Where, Ω_M represents the identity set of currently matched objects, denoted by $\mathcal{O}_i^t$, $i \in \Omega_M$. Likewise, Ω_U represents currently unmatched objects set. Their predicted positions are optimization variables, noted by $\hat{\mathcal{O}}_j^t$, $j \in \Omega_U$. N denotes the number of all historical targets. $\Delta\mathcal{O}_i^{t-1}$ and $\Delta\mathcal{O}_j^{t-1}$ denote relative coordinates of currently matched and unmatched objects, respectively, in $t-1$th frame, as known items. Through a simple mathematical analysis, the analytical solution to this optimization problem is:

$$\hat{\mathcal{O}}_j^t = \mathcal{O}_j^{t-1} + \frac{1}{l_{\Omega_M}} \sum_{i\in\Omega_M} \left(\mathcal{O}_i^t - \mathcal{O}_i^{t-1}\right) \tag{10}$$

Where, l_{Ω_M} denotes the number of currently matched objects, i.e. the size of Ω_M.

In summary, to obtain pairwise SAP, first we utilize Algorithm 1 to gain conditional structural association scheme. Then we adopt Eq.(7) to calculate conditional structural association probabilities. Finally, Eq.(6) is employed to compute pairwise SAP.

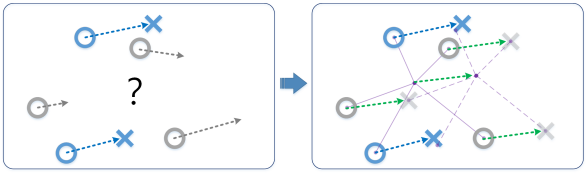

Figure 3: Prediction of unmatched objects' positions by minimizing structural variation cost. Blue circles and crosses represent matched targets and detections, respectively. Solid lines with a same intersection in the right graph represent the star-shaped structure of all targets. Dash lines with a same intersection in the right graph indicate the structural invariance. Gray circles and crosses denote unmatched objects and their predicted positions obtained by minimizing structural variation cost, i.e. Eq. (10)

4 Experiments

In this section, we present a quantitative evaluation of our proposed approach on the standard benchmarks with comparisons to several other methods.

4.1 Implementation details

In order to highlight the key advantage of structural invariance constraint on the data association assignment, color histogram of each target is extracted as the appearance feature and histogram intersection is employed to construct appearance similarities between detections and targets. After obtaining pairwise SAP described in section 3, Hungarian matching algorithm is adopted to gain optimal association event by maximizing the sum of association probabilities. With respect to those detections that are not associated with any historical objects, if their appearance similarities compared to missing targets exceed a certain threshold and their motion in the gap is reasonable and achievable, they will be regarded as the reappearance of missing targets. Otherwise, they are labeled as new targets.

4.2 Datasets and evaluation metrics

We evaluate the performance of our tracking algorithm on the Multiple Object Tracking Benchmark (MOT Challenge) [23]. The MOT Challenge provides a framework for the fair evaluation of MOT algorithms and a set of video sequences with labeled detections. In this paper, we test our tracking algorithm on dataset 2DMOT2015 [23] and MOT16 [27] as well. The data set 2DMOT2015 consists of 11 training and 11 test data sets, in which pedestrian detections are recognized by ACF pedestrian detector [12]. The MOT16 dataset contains 7 training and 7 test sequences, where DPM v5 [14] is adopted to obtain detections.

We adopt the widely used MOT evaluation metrics [5], in which Multiple Object Tracking Accuracy (MOTA) is a comprehensive evaluation with considering three errors: false detections, missed targets and identity switches. Multiple Object Tracking Precision (MOTP) measures the misalignment between the annotated and the predicted bounding boxes. For both MOTA and MOTP, a higher value indicates a better performance. Besides, we also use other evaluation metrics such as Mostly Tracked Targets (MT), Mostly Lost Targets (ML),

Algorithm 1 Heuristic matching approach

Input: Objects $\mathcal{O}$, Detections $\mathcal{D}$,
Conditional pair (m,n)
Output: Conditional matching event Φ_{mn}
Init: Matched and unmatched objects:
$\Omega_M = \{n\}$, $\Omega_U = \{1{:}N\}/\{n\}$
Unmatched detections: $\Psi_U = \{1{:}M\}/\{m\}$

$$\Phi_{mn}(i,j) = \begin{cases} 1 & i = m, j = n \\ 0 & others \end{cases}$$

while $\Omega_U \neq \varnothing$ **do**
  for $j \in \Omega_U$
    $\hat{\mathcal{O}}_j^t = \mathcal{O}_j^{t-1} + \frac{1}{l_{\Omega_M}} \sum_{k \in \Omega_M} \left(\mathcal{O}_k^t - \mathcal{O}_k^{t-1}\right)$
    for $i \in \Psi$
      $Distance(i,j) = \|\mathcal{D}_i^t - \hat{\mathcal{O}}_j^t\|^2$
    end
  end
  $(i^*, j^*) = arg\min(Distance(i,j))$
  if $\text{IoU}\left(B\left(\mathcal{D}_{i^*}^t\right), B\left(\hat{\mathcal{O}}_{j^*}^t\right)\right) > \varphi_r$
    Update: $\Psi_U \leftarrow \Psi_U/\{i^*\}, \Phi_{mn}(i^*, j^*) = 1$
    $\Omega_M \leftarrow \Omega_M \cup \{j^*\}$, $\Omega_U \leftarrow \Omega_U/\{j^*\}$
  else break
  end
end

Table 1: The comparison with other online MOT algorithms on the camera-moving video sequences

	Method	MOTA↑	MOTP↑	MT↑	ML↓
06	TBSS	**46.8**	**75.3**	**14.5%**	**47.1%**
	CDA [3]	39.2	73.4	9.0%	50.7%
	oICF [20]	41.2	71.5	8.1%	53.4%
	EAM [31]	29.9	71.8	4.5%	55.7%
07	TBSS	**40.8**	**73.6**	**13.0%**	35.2%
	CDA	38.8	72.8	11.1%	**31.5%**
	oICF	40.1	73.0	9.3%	35.2%
	EAM	35.0	73.2	9.3%	38.9%
12	TBSS	39.6	**76.7**	12.8%	48.8%
	CDA	38.3	76.3	**16.3%**	58.1%
	oICF	**42.7**	76.3	14.0%	50.0%
	EAM	34.6	75.7	14.0%	**46.5%**
14	TBSS	23.0	**74.5**	**3.7%**	61.6%
	CDA	**26.2**	73.4	**3.7%**	**54.3%**
	oICF	19.2	73.8	3.0%	71.3%
	EAM	16.1	74.4	1.8%	68.3%

False Positives (FP), False Negatives (FN), ID Switches (IDs), the total number of times a trajectory is fragmented (Frag) and the number of frames processed in one second (Hz).

4.3 Evaluation on MOT Challenge Benchmark

First, to verify the superiority of our tracker when coping with video sequences under dynamic camera situation, we exhibit the tracking results of our approach, denoted by TBSS, and other online MOT algorithms on four MOT2016 datasets [27], captured by moving cameras, including MOT16-06, MOT16-07, MOT16-12 and MOT16-14, shown in Table 1. It is important to note that they are all captured by moving cameras. The comparison shows that TBSS achieves the best or the second best performance on the metric of MOTA, MOTP and ML. However, our method does not beat the approaches oICF [20] and CDA_DDALv2 [3] on MOT12 and MOT14. This is because we pay little attention to recognizing the reappearance of a missing target, described in section 4.1. In addition, to demonstrate the performance of structural invariance constraint, the appearance feature used in this work is color histogram, which is not as discriminative as Integral Channel Features (used in oICF [20]) and Deep Appearance Learning (used in CDA_DDALv2 [3]). Therefore, we cannot outperform these two methods in the aspect of missing recovery. As a result, we have more false negatives, which reduce the performance on both MOTA and MT evaluation metrics.

Furthermore, Table 2 shows the tracking results of our approach and other trackers in the dataset 2DMOT2015 [23]. Part *a* exhibits the comparison of our approach and other tracking

Table 2: Results on the 2DMOT2015 test dataset: (a) comparison with the methods based on structural constraints (b) comparison with other sophisticated MOT trackers. Our approach is denoted by TBSS.

	Method	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDs↓	Frag↓	Hz↑
	TBSS	**29.2**	71.3	6.8%	**43.8%**	6068	36779	649	1508	11.5
a	SCEA [17]	29.1	71.1	**8.9%**	47.3%	6060	36912	604	1182	6.8
	RMOT [38]	18.6	69.6	5.3%	53.3%	12473	36835	684	1282	7.9
	TO [25]	25.7	**72.2**	4.3%	57.4%	**4779**	40511	383	**600**	5.0
	LP_SSVM [37]	25.2	71.7	5.8%	53.0%	8369	36932	646	849	41.3
	ELP [26]	25.0	71.2	7.5%	**43.8%**	7345	37344	139	1804	5.7
b	LINF1 [13]	24.5	71.3	5.5%	64.6%	5864	40207	298	744	7.5
	JPDA_m [15]	23.8	68.2	5.0%	58.1%	6373	40084	365	869	32.6
	MotiCon [24]	23.1	70.9	4.7%	52.0%	10404	**35844**	**101**	1061	1.4
	RNN_LSTM [30]	19.0	71.0	5.5%	45.6%	11578	36706	149	2081	**165.2**

methods based on structural constraints, including SCEA [17], RMOT [38]. Part *b* shows the comparison with other sophisticated MOT trackers, including TO [25], LP_SSVM [37], ELP [26], LINF1 [13], JPDA_m [15],MotiCon [24] and RNN_LSTM [30]. TBSS achieves the best performance on MOTA and ML compared with all other trackers. Moreover, TBSS also outperforms the structural constraints based methods on MOTP, FN and Hz.

Finally, the overall performance of our approach and other state-of-the-art methods on the MOT2016 [27] datasets are presented in Table 3. The statistics illustrates that TBSS achieves the best performance on MOTA, FAF, ML and FP. However, in both Table 2 and Table 3, our tracker seems to fragment trajectories frequently, so do other online trackers, including SCEA [17], RMOT [38], RNN_LSTM [30], CDA_DDALv2 [3], oICF [20], EAMTT_pub [31] and OVBT [4]. This could be explained as: for online MOT approaches, conservative strategies are usually adopted to recover missing targets since a bad recovery not only fails to reduce false negatives but also increases false positives and thus deteriorates the overall performance of a tracker. Nevertheless, offline MOT methods address the association from tracklets to tracklets, which aims at recovering the gap between different tracklets. Therefore, offline trackers are incline to link short tracklets into long trajectories, which, consequently, reduces the total number of times a trajectory is fragmented (Frag). Besides, the approach QuadMOT16 [34] outperforms our tracker on the metric of MOTP, as it exploits the strategy of bounding-box regression, which can refine the bounding box of each target and thus enhance the performance on MOTP.

5 Conclusion

We propose a new data association to address the problem of online multi-object tracking by exploiting structural invariance constraint, which is insensitive to the matching noises caused by analogous appearances and dynamic camera situation. Furthermore, we develop a total probability frame combining both appearance and structure cues without any adjustable parameters. Experimental results suggest that our proposed multi-object tracking algorithm achieves comparable performance with other state-of-the-art trackers on MOT Challenge dataset. However, the recovery of missing targets is limited by coarse appearance features

Table 3: Results on the MOT2016 test dataset. Our approach is denoted by TBSS

Method	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDs↓	Frag↓	Hz↑
TBSS	**44.6**	75.2	12.3%	**43.9%**	**4136**	96128	790	1419	3.0
QuadMOT16 [34]	44.1	**76.4**	**14.6%**	44.9%	6388	**94775**	745	1096	1.8
CDA_DDALv2 [3]	43.9	74.7	10.7%	44.4%	6450	95175	676	1795	0.5
oICF [20]	43.2	74.3	11.3%	48.5%	6651	96515	**381**	1404	0.4
LINF1 [13]	41.0	74.8	11.6%	51.3%	7896	99224	430	**963**	4.2
EAMTT_pub [31]	38.8	75.1	7.9%	49.1%	8114	102452	965	1657	**11.8**
OVBT [4]	38.4	75.4	7.5%	47.3%	11517	99463	1321	2140	0.3
LTTSC_CRF [21]	37.6	75.9	9.6%	55.2%	11969	103143	481	1012	0.6

extracted in this paper and thus can be enhanced by exploiting more discriminative appearance cues in future work.

6 Acknowledgements

This work was supported by the National Science and Technology Major Project (Grant No. 2018ZX01008103) and the Foundation Research Funds for the Central Universities (No. 1191329812).

References

[1] Anton Andriyenko and Konrad Schindler. Multi-target tracking by continuous energy minimization. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1265–1272. IEEE, 2011.

[2] Seung-Hwan Bae and Kuk-Jin Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1218–1225, 2014.

[3] Seung-Hwan Bae and Kuk-Jin Yoon. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[4] Yutong Ban, Sileye Ba, Xavier Alameda-Pineda, and Radu Horaud. Tracking multiple persons based on a variational bayesian model. In *European Conference on Computer Vision*, pages 52–67. Springer, 2016.

[5] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008 (1):246309, 2008.

[6] Alex Bewley, Lionel Ott, Fabio Ramos, and Ben Upcroft. Alextrac: Affinity learning by exploring temporal reinforcement within association chains. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 2212–2218. IEEE, 2016.

[7] Asad A Butt and Robert T Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1853, 2013.

[8] Visesh Chari, Simon Lacoste-Julien, Ivan Laptev, and Josef Sivic. On pairwise costs for network flow multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5537–5545, 2015.

[9] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4091–4099, 2015.

[10] Afshin Dehghan, Yicong Tian, Philip HS Torr, and Mubarak Shah. Target identity-aware network flow for online multiple target tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1146–1154, 2015.

[11] Caglayan Dicle, Octavia I Camps, and Mario Sznaier. The way they move: Tracking multiple targets with similar appearance. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2311, 2013.

[12] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (8):1532–1545, 2014.

[13] Loïc Fagot-Bouquet, Romaric Audigier, Yoann Dhome, and Frédéric Lerasle. Improving multi-frame data association with sparse representations for robust near-online multi-object tracking. In *European Conference on Computer Vision*, pages 774–790. Springer, 2016.

[14] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[15] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3047–3055, 2015.

[16] Carsten Haubold, Janez Aleš, Steffen Wolf, and Fred A Hamprecht. A generalized successive shortest paths solver for tracking dividing targets. In *European Conference on Computer Vision*, pages 566–582. Springer, 2016.

[17] Ju Hong Yoon, Chang-Ryeol Lee, Ming-Hsuan Yang, and Kuk-Jin Yoon. Online multi-object tracking via structural constraint event aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1392–1400, 2016.

[18] Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *European Conference on Computer Vision*, pages 788–801. Springer, 2008.

[19] Hao Jiang, Sidney Fels, and James J Little. A linear programming approach for multiple object tracking. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[20] Hilke Kieritz, Stefan Becker, Wolfgang Hübner, and Michael Arens. Online multi-person tracking using integral channel features. In *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*, pages 122–130. IEEE, 2016.

[21] Nam Le, Alexander Heili, and Jean-Marc Odobez. Long-term time-sensitive costs for crf-based tracking by detection. In *European Conference on Computer Vision*, pages 43–51. Springer, 2016.

[22] Laura Leal-TaixÃl', Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *Computer Vision and Pattern Recognition Workshops*, pages 418–425, 2016.

[23] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, April 2015. URL http://arxiv.org/abs/1504.01942. arXiv: 1504.01942.

[24] Laura Leal-Taixé, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. Learning an image-based motion context for multiple people tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3542–3549, 2014.

[25] Santiago Manen, Radu Timofte, Dengxin Dai, and Luc Van Gool. Leveraging single for multi-target tracking using a novel trajectory overlap affinity measure. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[26] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Enhancing linear programming with motion modeling for multi-target tracking. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 71–77. IEEE, 2015.

[27] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, March 2016. URL http://arxiv.org/abs/1603.00831. arXiv: 1603.00831.

[28] Anton Milan, Stefan Roth, and Konrad Schindler. Continuous energy minimization for multitarget tracking. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):58–72, 2014.

[29] Anton Milan, Konrad Schindler, and Stefan Roth. Multi-target tracking by discrete-continuous energy minimization. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2054–2068, 2016.

[30] Anton Milan, Seyed Hamid Rezatofighi, Anthony R Dick, Ian D Reid, and Konrad Schindler. Online multi-target tracking using recurrent neural networks. In *AAAI*, pages 4225–4232, 2017.

[31] R Sanchez-Matilla, F Poiesi, and A Cavallaro. Multi-target tracking with strong and weak detections. In *ECCV Workshops-Benchmarking Multi-Target Tracking*, volume 5, page 18, 2016.

[32] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep network flow for multi-object tracking. 2017.

[33] Francesco Solera, Simone Calderara, and Rita Cucchiara. Learning to divide and conquer for online multi-target tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4373–4381, 2015.

[34] Jeany Son, Mooyeol Baek, Minsu Cho, and Bohyung Han. Multi-object tracking with quadruplet convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5620–5629, 2017.

[35] Siyu Tang, Bjoern Andres, Miykhaylo Andriluka, and Bernt Schiele. Subgraph decomposition for multi-target tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5033–5041, 2015.

[36] Bing Wang, Li Wang, Bing Shuai, Zhen Zuo, Ting Liu, Kap Luk Chan, and Gang Wang. Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association. In *Computer Vision and Pattern Recognition Workshops*, pages 386–393, 2016.

[37] Shaofei Wang and Charless C Fowlkes. Learning optimal parameters for multi-target tracking with contextual interactions. *International Journal of Computer Vision*, 122 (3):484–501, 2017.

[38] Ju Hong Yoon, Ming-Hsuan Yang, Jongwoo Lim, and Kuk-Jin Yoon. Bayesian multi-object tracking using motion context from multiple objects. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 33–40. IEEE, 2015.

[39] Shun Zhang, Yihong Gong, Jia-Bin Huang, Jongwoo Lim, Jinjun Wang, Narendra Ahuja, and Ming-Hsuan Yang. Tracking persons-of-interest via adaptive discriminative features. In *European Conference on Computer Vision*, pages 415–433. Springer, 2016.