Online Multi-Object Tracking by Quadratic Pseudo-Boolean Optimization

Long Lan[†‡], Dacheng Tao[‡], Chen Gong[‡], Naiyang Guan[†], Zhigang Luo[†]
[†] College of Computer, National University of Defense Technology
[‡] Centre for Quantum Computation & Intelligent Systems, FEIT,
University of Technology, Sydney
long.lan@nudt.edu.cn, dacheng.tao@uts.edu.au,
goodgongchen@gmail.com, {ny_guan, zgluo}@nudt.edu.cn.

Abstract

Online multi-object tracking (MOT) is challenging: frame-by-frame matching of detection hypotheses to the correct trackers can be difficult. The Hungarian algorithm is the most commonly used online MOT data association method due to its rapid assignment; however, the Hungarian algorithm simply considers associations based on an affinity model. For crowded scenarios, frequently occurring interactions between objects complicate associations, and affinity-based methods usually fail in these scenarios. Here we introduce quadratic pseudo-Boolean optimization (QPBO) to an online MOT model to analyze frequent interactions. Specifically, we formulate two useful interaction types as pairwise potentials in QPBO, a design that benefits our model by exploiting informative interactions and allowing our online tracker to handle complex scenes. The auxiliary interactions result in a non-submodular QPBO, so we accelerate our online tracker by solving the model with a graph cut combined with a simple heuristic method. This combination achieves a reasonable local optimum and, importantly, implements the tracker efficiently. Extensive experiments on publicly available datasets from both static and moving cameras demonstrate the superiority of our method.

1 Introduction

Multi-object tracking (MOT) is an important computer vision topic, the goal being to locate objects in successive frames to produce intact trajectories. However, MOT is particularly challenging, since frequent interactions, anomalous motion, and similar appearances of tracked objects are common in many real-world scenarios, and each of these challenges must be overcome to improve performance. Many solutions have been proposed in this field. Among them, tracking-by-detection is a widely accepted MOT method. Tracking-by-detection tracks objects based on given detection hypotheses; however, in practice, even the best detector produces false or missing detections. A key issue, therefore, is how to effectively associate these imperfect hypotheses and build complete trajectories.

There are two main types of detection-based tracking. The first is the batch method [Huang *et al.*, 2008; Milan *et al.*, 2014; Yang and Nevatia, 2014; Zamir *et al.*, 2012], in which whole frames are pre-processed in advance to produce fragmented trajectories (tracklets), with which data is associated. The other is the online method [Bae and Yoon, 2014; Breitenstein *et al.*, 2011; Poiesi *et al.*, 2013; Shu *et al.*, 2012], which is designed for real-time applications with tracking performed on only previous and current frames. The batch method has recently attracted interest since cues from future frames impart a significant effect in complex scenarios (i.e., those with many occlusions and collisions), and optimization based on global information delivers better results. The online method usually improves tracking when used with other techniques such as particle filtering [Breitenstein *et al.*, 2011] and KLT [Benfold and Reid, 2011].

With respect to data association, the Hungarian algorithm, which efficiently assigns hypotheses to the correct tracker, is usually used in online methods [Bae and Yoon, 2014; Breitenstein *et al.*, 2011; Shu *et al.*, 2012]. The performance of the Hungarian algorithm is closely related to the affinity model, and many related methods have sought to develop a discriminative appearance model and robust motion model to improve data association. However, Hungarian-derived methods ignore favorable information between objects. Objects frequently interact in crowed scenes, which inevitably worsens affinity models and the Hungarian algorithm's optimal assignment is problematic in this scenario. Several recent studies have formulated their MOT model by introducing interactions; for instance, [Yang and Nevatia, 2014] regards interactive tracklets as the pairwise terms of CRF and redefines the data association cost. [Milan *et al.*, 2013] proposes label cost CRF to deal with detection- and trajectory-level interactions. The convincing performance of these works prove the significance of utilizing interactions. However, all these works focus on global associating, thus are not the online case. Many classical combinatorial optimization approaches have recently been successfully applied to MOT. For example, [Zamir *et al.*, 2012] formulates MOT as a generalized minimum clique problem, and in a related study [Dehghan *et*

al., 2015] updates MOT as a maximum multi-clique problem. In both works, tracklets from the same object are assumed to be from the same clique, and tracking are translated to find the maximum cliques. [Tang *et al.*, 2015] considers MOT as a subgraph decomposition problem, with subgraphs representing feasible data associations and MOT finding the optimal subgraph. Although these approaches show their efficacies in handling one to one association, they fail to extend to include interactions.

In this paper, we focus on online MOT modeling, and formulate it as quadratic pseudo-Boolean optimization (QPBO) problem. Online MOT is challenging, since neither future frames nor global optimization is available. We study the online case based on the following two facts. First, recently proposed DPM detector [Felzenszwalb *et al.*, 2010] provides more accurate detection hypotheses, and appearance models built on these part-based detections are discriminative. Second, interactions are informative in online MOT, and carefully incorporated these interactions significantly improve online tracking performances. We use the advanced detector to obtain hypotheses and introduce QPBO to explore useful interactions. We consider our online MOT as a QPBO problem due to its flexibility in describing interactions. In our model, we formulate two frequently occurring interactions, i.e. collision interaction and overlapping interaction, as the pairwise potentials of QPBO and extend our online model to handle complex scenes. We note that our online tracker performs local data association and any iterative or global associations are not implemented. To speed up our method, a novel optimization algorithm is proposed, which enables QPBO to find the optimum efficiently. To our knowledge, both our model and optimization have not previously been studied.

2 The proposed online MOT method

In this section, we discuss several useful interactions that occur in online MOT and devise our online MOT model to take them into account. Since our method focuses on data association between trackers and detection hypotheses, we record all hypotheses and trackers' states. $D^t = [d_1^t, d_2^t, \ldots, d_m^t]$ denotes m detection hypotheses at the t^{th} frame. Each d_i^t is an individual hypothesis, $i \in \{1, 2, \ldots, m\}$. During the matching process of our online model, only the detection hypotheses of the current frame are used. We denote the trackers $T = [T_1, T_2, \ldots, T_n]$, where n is the number of trackers, and each $T_j = \{\mathcal{X}_j, \mathcal{V}_j, \mathcal{F}_j\}$, $j \in \{1, 2, \ldots, n\}$, where $\mathcal{X}_j = \{d_i^t | \delta^t(i,j) = 1, t^s \leq t \leq t^e\}$ represents the hypothesis set that contains all the detections matched to tracker T_j, and $\delta^t(i,j) = 1$ means d_i^t has been linked to T_j, and equals 0 otherwise. t^s, t^e are the start frame and ending frame, respectively. $\mathcal{V}_j$ signifies the tracker velocity, which can be estimated using the collected hypothesis sequences using the Kalman filter. $\mathcal{F}_j$ denotes the specific SVM classifier, which is constructed similar to [Shu *et al.*, 2012] using the appearance features. These specific classifiers are fed online and the classifier is updated when the matched detection is not heavily occluded, i.e. 70% of the detection is visible. Based on this framework, an affinity model between all trackers and hypotheses can easily be obtained, as detailed in Section 4.

As mentioned before, data association in MOT is a classical combinatorial optimization problem. It can, therefore, be easily formulated as a graph, the most common methods being to regard detection hypotheses or tracklets as vertices and affinities as edges [Dehghan *et al.*, 2015; Tang *et al.*, 2015]. However, we consider our online MOT as a quadratic pseudo 0-1 optimization problem, and we construct the graph differently. The graph vertices represent combinations of trackers with hypotheses. Given m detection hypotheses and n trackers, there exist m*n vertices in the graph and we denote the vertex set with Ω. The graph size is varied according to the active trackers that are not terminated and hypotheses of the incoming frame. Each vertex $v_p = \delta(i,j)$, $p \in \Omega$ taking either 1 or 0 to indicate whether or not the combination of d_i and T_j is successful. For simplicity, we take v_p as tuple $< d_i, T_j >$. We minimize the following energy function to achieve the data associations of our QPBO-based online MOT model:

$$E(v) = v^T u + \alpha v^T F(\mathbf{1} - v) + \beta v^T G v, \quad v \in \{0,1\}^{|\Omega|}. \tag{1}$$

Where $v = \{v_p | p \in \Omega\}$ is a binary vector defined on all Ω vertices. $u = \{u_p | p \in \Omega\}$ represents the unary term with each element denoting the cost of the corresponding vertex. The unary term is derived independently from our affinity model detailed later. $\mathbf{1}$ is an all one vector of the same length as v; Symmetric matrices $F = \{f_{pq} | p, q \in \Omega\}$ and $G = \{g_{pq} | p, q \in \Omega\}$ represent pairwise potentials constructed using the mentioned two interactions, where F is constructed from collision interactions, and G is constructed from overlapping interactions. Both potentials can be seen as the cost of violating interactions. α, β balance the two pairwise terms.

2.1 Tracking assumption

Two tracking assumptions are addressed by almost all MOT methods when tracking by detection. The first is that one tracker links to at most one detection for each step. The second assumption emphasizes that one detection cannot be occupied by two different trackers simultaneously. To obey the first tracking assumption, each tracker in our model only take its best matched hypothesis as matchable. But, different from Hungarian algorithm, we reassure each exclusive match by nearby trackers. That is to say, nearby trackers can deny the match. We introduce collision interaction to achieve this goal. Due to the stricter conditions of matching, one object is likely to be tracked by two trackers. We thus relax the second assumption to allow two active trackers to match to one hypothesis and merge them into one tracker. We consider the second assumption together with overlapping interaction.

2.2 Collision interaction

One MOT challenge is how to distinguish closely positioned similar-appearing objects in crowded scenes, since motion models and traditional appearance models tend to fail in this scenario. We introduce collision interactions to describe this challenging situation and devise a classifier for each pair of closely aligned trackers. Specifically, when two trackers are detected within a close and reasonable distance, we consider

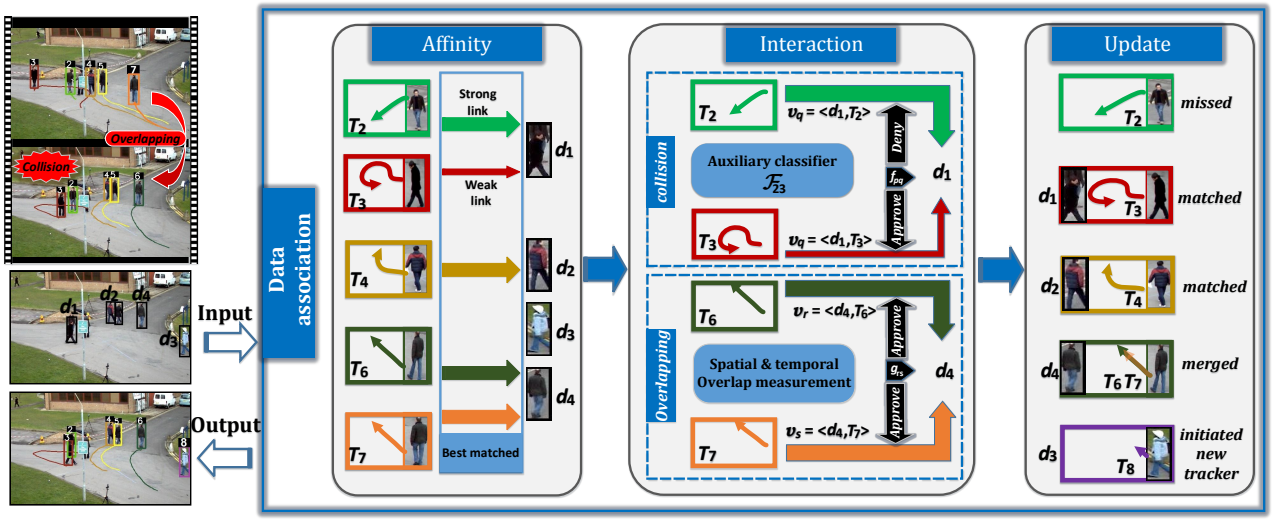

Figure 1: A system flow of the proposed online MOT. Affinity provides the best macthed pairs between trackers and incoming detections. Then, two types of interactions are considered to improve matching. Lastly, trackers update to produce the tracking output. The similar-appearing T_2 and T_3 come to close and have a collision. The collision interaction reassures links using auxiliary classifier $\mathcal{F}_{23}$ and correctly matches d_1 to T_3 in spite of affinity matching d_1 to T_2 with a strong link. T_6 and T_7 share heavy overlaps and both strong link to d_4. The overlapping interaction approve d_4 to match to T_6 and T_7 simultaneously, and merge them into one single tracker.

there is a collision and a classifier is generated with training samples collected only from these two trackers. This classifier is naturally more discriminative for separating the two trackers than a global classifier. To prevent similar appearing trackers linking to incorrect hypotheses, the following cost is provided:

$$f_{pq} = \begin{cases} c_{max}, & \mathcal{F}_{jj'}(d_i) = 0 \wedge col(T_j, T_{j'})) \\ 0, & otherwise. \end{cases} \quad (2)$$

Here p, q are the indices of $v_p =< d_i, T_j >$ and $v_q =< d_i, T_{j'} >$, respectively. $\mathcal{F}_{jj'}(d_i)$ denotes the binary output of the auxiliary classifier $\mathcal{F}_{jj'}$ on hypothesis d_i, $col(\cdot)$ checks whether there is a collision between the two trackers of T_j and T'_j. In our model, $col(\cdot)$ returns true when the distance between two trackers is less than three times width of their last matched detection. c_{max} signifies an infinite cost to which we set a sufficiently large value 10^5 in practice. In consideration of matching d_i to T_j, Eq. (2) heavily penalizes the situation where $\mathcal{F}_{jj'}(d_i)$ classifies d_i to $T_{j'}$.

2.3 Overlapping interaction

In collision interaction, our trackers are prudent to match to hypotheses candidates, which is prone to initiate new trackers. In such cases, one object is likely to be tracked by two or more trackers. As emphasized in [Milan *et al.*, 2013], heavily overlapped trackers always originate from the same objects and should be suppressed. However, most studies have neglected this situation to simplify optimization. We address this case using pairwise term G in our QPBO model as follows:

$$g_{rs} = \begin{cases} O(T_j, T_{j'}), & if\ d_i = d_{i'} \\ 0, & otherwise. \end{cases} \quad (3)$$

Here r, s are the indices of $v_r =< d_i, T_j >$ and $v_s =< d_{i'}, T_{j'} >$, respectively. $O(T_j, T_{j'})$ is used to measure the spatial-temporal overlap cost of two trackers T_j and $T_{j'}$ defined:

$$O(T_j, T_{j'}) = -\log\left(O_S(T_j, T_{j'})\, O_{TN}(T_j, T_{j'})\right). \quad (4)$$

$O_S(T_j, T_{j'})$ and $O_{TN}(T_j, T_{j'})$ is the spatial overlap ratio and temporal non-overlap ratio, respectively. $O_S(T_j, T_{j'})$ calculates the ratio of overlapping areas of trackers T_j, $T_{j'}$ in terms of all their shared time periods. Here, each tracker is smoothed by interpolating. $O_{TN}(T_j, T_{j'})$ calculates the ratio of temporal non-overlap number of T_j, $T_{j'}$ to their total detections. β balances this item. To prevent two unrelated trackers from merging together, we use a large β to guarantee two merged trackers are sufficiently overlapped.

3 Optimization using QPBO

The quadratic Boolean optimization has been shown to be a generalization of other combinatorial optimization problems. Since our model emphasizes more MOT interactions, these complicated constraints make it difficult to apply our problem to a particular combinatorial optimization. In this paper, we consider our model as a QPBO. Since $v_i u_i = v_i u_i v_i$ for every $v_i \in \{0, 1\}$, our model in Eq. (1) can be reshaped as:

$$\begin{aligned} E(v) &= \min_{v \in \{0,1\}^\Omega} \sum v^T M v + \alpha v^T F \mathbf{1} \\ &= \min_{v_p, v_q \in \{0,1\}} \sum_{p,q} m_{pq} v_p v_q + \alpha \sum_{p,q} f_{pq} v_p \end{aligned} \quad (5)$$

Where $M = diag(u) - \alpha F + \beta G$, $diag(u)$ denotes a diagonal matrix with diagonal elements equal to vector u. According to [Boros and Hammer, 2002], Eq. (5) is submodular

and can only be efficiently solved in low-order polynomial time if $m_{pq} \leq 0$, $\forall (p,q) \in \Omega^2$. However, in our model, most elements are positive. The general non-submodular case of Eq. (5) is NP hard. We solve our model using graph cuts [Boros and Hammer, 2002] combined with simple greedy heuristic searching [Merz and Freisleben, 2002; Liu and Tao, 2014] for efficiency. Graph cuts are considered one of the most efficient QPBO algorithms; however, as noted in many works, a graph cut may only provide part of the optimum (global) if the model is not submodular and leave the rest unsolved. For MOT applications, only limited objects appear simultaneously in a short period of time, therefore, the application size is relatively small; in fact, there are less than 50 simultaneous trackers in our experiment and the constructed graph has around 2500 vertices in maximum. Embracing this observation, We subsequently use a simple greedy method [Merz and Freisleben, 2002; Tao *et al.*, 2007; 2009] to obtain the remaining solutions, which initiates the value of each vertex to 0.5 and approximates to local solution by heuristically adjusting its value to 0 or 1. This method is known to be very efficient for the small size of an unconstrained quadratic Boolean optimization problem.

4 Implementation details

As mentioned above, our online model is based on detection hypotheses; we use the deformable part-based detector [Felzenszwalb *et al.*, 2010] to obtain hypotheses for each frame, and all our hypotheses are treated equally. We initiate a new tracker when a hypothesis is not assigned to any active trackers and add it to the active trackers set. On the other hand, we terminate an active tracker when it is not matched to any hypotheses over 50 frames. α and β of Eq. (1) are two trade off parameters in our model, we set $\alpha = 1$ and $\beta = 100$ empirically. Eq. (1) also refers affinity model to obtain u, we detail our affinity model here.

Our online model only focuses on data associations between trackers and hypotheses; thus, since the affinity model only measures how well trackers match hypotheses, we take tracker and hypothesis' affinity as their link probability. The link probability is estimated using four cues (including appearance, motion, size and reliability) as follows:

$$\begin{aligned} P_{ij}^{link} &= P\left(\delta\left(i,j\right)=1|d_i,T_j\right) \\ &= \Lambda_A * \Lambda_M * \Lambda_S * \Lambda_T, \end{aligned} \tag{6}$$

where

$$\begin{aligned} \Lambda_A\left(d_i,T_j\right) &= \frac{1}{Z}\exp\left\{\frac{\mathcal{F}_j\left(d_i\right)}{\sigma_A}\right\}, \\ \Lambda_M\left(d_i,T_j\right) &= G(p_{T_j}+\mathcal{V}_j*\Delta t_j - p_{d_i};0,\sigma_{ij}), \\ \Lambda_S\left(d_i,T_j\right) &= G(w_{T_j}-w_{d_i};0,\sigma_{ij}), \\ \Lambda_T &= \pi^{l_j+\Delta t_j}. \end{aligned}$$

$\Lambda_A(d_i,T_j)$ measures appearance similarity; $\mathcal{F}_j(d_i)$ is the classifier score of $\mathcal{F}_j$ on d_i, we adopt the well-proven LBP and color features [Shu *et al.*, 2012] in our appearance model and build the classifier similar to [Shu *et al.*, 2012]. Since we build a specific classifier for each tracker only after a sufficient number of samples have been collected, in our experiments $|\mathcal{X}_j| \geq 5$. In cases with insufficient samples, we replace with correlation. We set $\sigma_A = 0.1$ empirically. Z is the normalization factor, similar to [Huang *et al.*, 2008]. $\Lambda_M(d_i,T_j)$ is used to measure motion; p_{T_j} represents the position of T_j (the last linked detection replaced), the same as p_{d_i}. Δt_j is the time gap between the current frame and T_j. $G(\cdot)$ represents Gaussian distribution function. $\Lambda_S(d_i,T_j)$ measures the size change of tracker and hypothesis; w_{d_i} denotes the width of hypothesis d_i and w_{T_j} denotes the width of tracker T_j. Motion variance σ_{ij} has the same formulation throughout and we introduce parameter λ to tune the variance.

$$\sigma_{ij} = \lambda * \min(w_{T_j}, w_{d_i}). \tag{7}$$

λ is an important parameter to our model and its selection will be detailed in experiment section. Λ_T measures the reliability of T_j, it includes two considerations: long trackers are more reliable and trackers unmatched for a long time however are less reliable. We take π as the missed detection rate of the DPM detector and set $\pi = 0.1$. To make long trackers more confident, we denote $|T_j|$ as the length of T_j and set $l_j = max(0, 10 - |T_j|)$. Trackers longer than 10 are considered confident enough. Δt_j measures the time gap between T_j and current frame. Large gap means the corresponding trackers are likely to leave the scene and less confident. We eventually obtain the link cost:

$$c_{ij} = -\log(P_{ij}^{link}). \tag{8}$$

To avoid our model violating the first tracking assumption, for each tracker T_j, we find the best matched hypothesis d_k where $k = \arg\min_i c_{ij}$. Then our unary terms in Eq. (1) can be readily achieved by setting $u_p = c_{kj} - h$ when $v_p =< d_k, T_j >$, and $u_p = c_{max}$ otherwise. Here h is a constant, which can be treated as a threshold to enforces the best matched pair to associate. We set $h = 100$ and it has no obvious influence to the tracking performances.

5 Experiments

In this section, we firstly validate the effectiveness of two interactions mentioned in Section 3. Then we compare the performance of the proposed online MOT model with several state-of-the-art methods on five publicly available datasets. Subsequently, we further prove the proposed method is more robust than the Hungarian algorithm in complex scenarios. Computational cost is analyzed at last.

5.1 Evaluation metrics

For quantitative evaluation, we use the widely adopted CLEAR MOT metrics [Bernardin and Stiefelhagen, 2008] (including MOTA(↑) and MOTP(↑)) and trajectory-based metrics (TBM) [Yang and Nevatia, 2014] (including Mostly Tracked MT(↑), Fragments FM(↓) and Identity Switch IDS(↓))[1]. CLEAR MOT metrics measure the results based on entire video and calculate MOTA and MOTP frame-by-frame. MOT accuracy (MOTA↑) evaluates accuracy in the presence of false positives, false negatives, and IDS. MOT

[1]Here "↑" means the larger, the better on the corresponding metric, while "↓" denotes the smaller, the better.

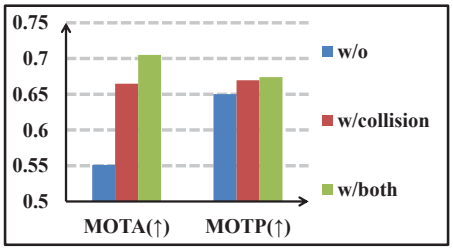

Figure 2: The effectiveness of two interactions (overlapping and collision) to tracking results. w/o: model outputs without any of the two interactions; w/collision: model outputs with collision interaction; w/both: model outputs with both overlapping and collision interactions.

precision (MOTP↑) evaluates the intersecting area of the tracking output and the ground truth. TBM is an important supplemental metric for better evaluation of MOT, which measures tracking results based on trajectories and is used to estimate the completeness of each trajectory. Specifically, MT↑ evaluates the most tracked trajectories that are successfully tracked at least 80%. IDS↓ counts the number of times that a tracked trajectory changes its matched identity. FM↓ counts the number of times that a trajectory in ground truth is interrupted by the tracking output. For fair comparison, we report the TBM of our method based on the existing MOT evaluation tool provided by [Yang and Nevatia, 2014].

5.2 Validation of our method

The reason that our method achieves the encouraging performance is due to the incorporation of two types of interaction: overlapping and collision. To demonstrate that the introduced interactions are indeed helpful to our online MOT model, we investigate the performance gain brought by each of the two interactions on *ETH-bahnhof* and *ETH-sunny* datasets [Ess *et al.*, 2008]. Firstly, we remove the two interactions from our tracking scheme and observe the performance on MOTA and MOTP. After that, collision interaction is incorporated to test its effectiveness to improve the performance. Finally, we introduce the overlapping interaction to form the completed proposed method, and see whether overlapping interaction is able to further enhance the tracking performance. The contributions of the two interactions are illustrated by Fig. 2.

5.3 Comparison with the state-of-the-arts

The datasets adopted here include *PETS-S2L1*[2], *TUD-Crossing*, *TUD-Campus* [Andriluka *et al.*, 2008], *ETH-bahnhof*, and *ETH-sunny* [Ess *et al.*, 2008]. These datasets provided a wide range of challenges including occlusion, crowded scenarios, and moving backgrounds. The ground truth of all datasets are easy to obtain, and many related works have reported their results on these datasets, allowing the straightforward comparison of our tracker with other state-of-the-art methods.

The qualitative and quantitative results of our method on the above datasets are presented by Fig. 3 and Tables 1∼4, respectively. Table 1 reveals that we obtain the best MOTA on *PETS-S2L1*, which indicates our online fed classifiers effectively captures the changing appearance of each pedestrian. As showed in the first row of Fig. 3, our tracker 2 successfully tracks the pedestrian for more than 500 frames. Tables 2 and 3 present the tracking performances on *TUD-Crossing* and *TUD-Campus* datasets. It can be observed that the proposed method achieves the best MT and very encouraging MOTA, demonstrating our trackers' robustness to the heavy collisions. Tracker 2 showed in the third row of Fig. 3 re-identifies the pedestrian even after the pedestrian being occluded up to 45 frames. The tracking results on moving background datasets *ETH-Bahnhof* and *ETH-Sunny* are showed in Table 4. Compared with batch methods [Kuo *et al.*, 2010; Yang and Nevatia, 2014], our model is prone to produce more FM and IDS. This is predictable, since online methods are known to perform poorly when handling long-term occlusions, and the trajectories are tend to split without further global association. However, the proposed model achieves the best results when compared with other online methods. Tracker 4 showed in the last row of Fig. 3 perfectly deals with the camera moving and detection missing, and renders a complete trajectory of the pedestrian who exists in all frames of the sequence.

Table 1. Tracking performance on *PETS-S2L1*.

Method	MOTA	MOTP	MT	FM	IDS
Milan [2014]	90.6	80.2	91.0	6	11
Dehghan [2015]	90.4	63.1	95.0	3	0
Yang [2014]	–	–	90.0	13	0
Bae [2014]	83.0	69.6	100.0	4	4
Breitenstein* [2011]	79.7	56.3	–	–	–
Bae* [2014]	77.4	69.0	100.0	12	10
Proposed*	95.1	71.7	95.0	5	3

Table 2. Tracking performance on *TUD-Crossing*.

Method	MOTA	MOTP	MT	FM	IDS
Zamir [2012]	91.6	75.6	–	–	0
Dehghan [2015]	91.9	70.0	76.9	–	2
Tang [2015]	80.9	78.0	61.5	1	1
Breitenstein* [2011]	84.3	71.0	–	–	2
Proposed*	87.4	75.2	84.6	1	3

Table 3. Tracking performance on *TUD-Campus*.

Method	MOTA	MOTP	MT	FM	IDS
Segal [2013]	82.0	74.0	62.5	3	0
Tang [2015]	83.3	76.9	62.5	1	0
Breitenstein* [2011]	73.3	67.0	–	–	2
Proposed*	86.6	71.4	87.5	2	1

Table 4. Tracking performance on *ETH-bahnhof & ETH-sunny*.

Method	MOTA	MOTP	MT	FM	IDS
Kuo [2010]	–	–	58.4	23	11
Yang [2014]	–	–	68.0	19	11
Bae [2014]	72.0	64.0	73.8	38	18
Poiesi* [2013]	–	–	62.4	69	45
Bae* [2014]	67.9	60.0	68.3	57	23
Proposed*	70.5	67.4	70.0	34	25

Methods marked with * are online models.

[2]http://www.cvg.reading.ac.uk/PETS2009/a.html

Figure 3: Examples of our tracking results on the test sequences *PETS-S2L1* (1st row), *TUD-Crossing* (2nd row), *TUD-Campus* (3rd row), *ETH-bahnhof* (4th row), and *ETH-sunny* (5th row).

5.4 Robust to motion variances

Due to the unavoidable missing detections, in crowded scenes hypothesis is usually re-detected in a distance far from its previous position. Besides, in moving background datasets, the linear motion prediction is imprecise [Bae and Yoon, 2014; Liu *et al.*, 2014]. To overcome these unexpected motions, a natural way is to build a more discriminative appearance model. The Hungarian algorithm simply uses the information of individual tracker to build the appearance model. As a result, they cannot successfully deal with the issues mentioned above. However, we build our appearance model by further exploring the interaction cues, which effectively distinguish different objects and enable our trackers to link to distant hypotheses accurately. Fig. 4 shows the comparison results of the proposed method with the Hungarian algorithm under different motion variances $\lambda \in \{2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}\}$ using the same affinity model. Large λ means that trackers are allowed to link the hypotheses that are far apart. Here, we use four key metrics introduced above, i.e. MOTA, MOTP, MT and IDS, to show the improvement brought by our strategy. The reported MOTA, MOTP and MT are averaged over the outputs on the five datasets, and IDS is summed over the outputs on all five datasets. It is obvious that the Hungarian

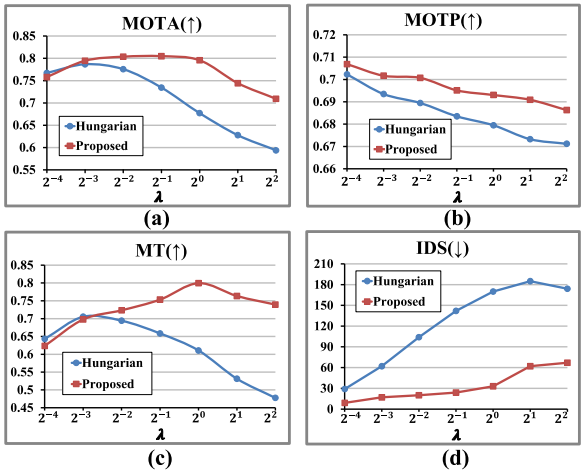

Figure 4: Tracking results under different motion variances.

algorithm prefers smaller motion variance $\lambda = 2^{-3}$ in term of MOTA (see Fig. 4(a)) and MT (Fig. 4(c)). It fails under large λ where many IDS is produced (see in Fig. 4(d)). However, the proposed method is robust to the large motion variance in terms of MOTA, MT and IDS. The large λ enables our method successfully track more completed trajectories than the Hungarian algorithm (see in Fig. 4(c)). Embracing this observation, we set $\lambda = 1$ for all our previous experiments.

5.5 Runtime analysis

Since the incorporation of interactions, our online method spends about 23% more time than the Hungarian algorithm on the five test sequences. Such overhead includes the cost of data association and interaction computation. In fact, due to the efficient solver to QPBO, the data association cost accounts for less than 5% of the overhead. In a nutshell, our online method remarkably improves MOT tracking performance without significantly increasing the run time. We perform our experiments on a 3.45GHz PC with 6.0 GB memory with codes implemented in C++. Our non-optimized codes run at around 3-12 fps on the different sequences, and the run time mainly depends on the number of simultaneous trackers. The processing time does not include the cost of detection procedure.

6 Conclusion

In this paper, we formulate online MOT as a QPBO problem. As a general combinatorial optimization method, QPBO proves to be more flexible for solving MOT. We introduce two frequently occurring interactions into QPBO to comprehensively analyze the link probability of data association. Compared to a unary affinity model, the interactions help our trackers associate hypotheses more accurately in complicated scenarios. Extensive experiments demonstrate the effectiveness of our model, and the quantitative analysis of each interaction reveals their respective value. Furthermore, we propose an efficient optimization for our QPBO-based online tracking model, with the proposed combination of a

graph cut and a simple heuristic search, thus solving our non-submodular problem efficiently in polynomial time.

7 Acknowledgment

This work is partially supported by Research Fund for the Doctoral Program of Higher Education of China, SRFDP (under grant No. 20134307110017), Scientific Research Plan Project of NUDT (under grant No. JC13-06-01) and Australian Research Council Projects (DP-140102164, FT-130101457 and LE-140100061).

References

[Andriluka *et al.*, 2008] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, pages 1–8, 2008.

[Bae and Yoon, 2014] Seung-Hwan Bae and Kuk-Jin Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *CVPR*, pages 1218–1225, 2014.

[Benfold and Reid, 2011] Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, pages 3457–3464, 2011.

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

[Boros and Hammer, 2002] Endre Boros and Peter L Hammer. Pseudo-boolean optimization. *Discrete applied mathematics*, 123(1):155–225, 2002.

[Breitenstein *et al.*, 2011] Michael D Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1820–1833, 2011.

[Dehghan *et al.*, 2015] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, page 2, 2015.

[Ess *et al.*, 2008] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, pages 1–8, 2008.

[Felzenszwalb *et al.*, 2010] 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.

[Huang *et al.*, 2008] Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, pages 788–801. 2008.

[Kuo *et al.*, 2010] Cheng-Hao Kuo, Chang Huang, and Ramakant Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *CVPR*, pages 685–692, 2010.

[Liu and Tao, 2014] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[Liu *et al.*, 2014] Xinwang Liu, Lei Wang, Jian Zhang, Jianping Yin, and Huan Liu. Global and local structure preservation for feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 25(6):1083–1095, 2014.

[Merz and Freisleben, 2002] Peter Merz and Bernd Freisleben. Greedy and local search heuristics for unconstrained binary quadratic programming. *Journal of Heuristics*, 8(2):197–213, 2002.

[Milan *et al.*, 2013] Anton Milan, Kaspar Schindler, and Stefan Roth. Detection-and trajectory-level exclusion in multiple object tracking. In *CVPR*, pages 3682–3689, 2013.

[Milan *et al.*, 2014] Anton Milan, Stefan Roth, and Kaspar Schindler. Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):58–72, 2014.

[Poiesi *et al.*, 2013] Fabio Poiesi, Riccardo Mazzon, and Andrea Cavallaro. Multi-target tracking on confidence maps: An application to people tracking. *Computer Vision and Image Understanding*, 117(10):1257–1272, 2013.

[Segal and Reid, 2013] Aleksandr V Segal and Ian Reid. Latent data association: Bayesian model selection for multi-target tracking. In *ICCV*, pages 2904–2911, 2013.

[Shu *et al.*, 2012] Guang Shu, Afshin Dehghan, Omar Oreifej, Emily Hand, and Mubarak Shah. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, pages 1815–1821, 2012.

[Tang *et al.*, 2015] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Subgraph decomposition for multi-target tracking. In *CVPR*, pages 5033–5041, 2015.

[Tao *et al.*, 2007] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1700–1715, 2007.

[Tao *et al.*, 2009] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):260–274, 2009.

[Yang and Nevatia, 2014] Bo Yang and Ramakant Nevatia. Multi-target tracking by online learning a crf model of appearance and motion patterns. *International Journal of Computer Vision*, 107(2):203–217, 2014.

[Zamir *et al.*, 2012] Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *ECCV*, pages 343–356. Springer, 2012.