Semi-Online Multiple Object Tracking Using Graphical Tracklet Association

Jiahui Wang, Yulan Guo, Xing Tang, Qingyong Hu, and Wei An

***Abstract*—Online multiple object tracking (MOT) is highly challenging when multiple objects have similar appearance or under long occlusion. In this letter, we propose a semi-online MOT method using online discriminative appearance learning and tracklet association with a sliding window. We connect similar detections of neighboring frames in a temporal window, and improve the performance of appearance feature by online discriminative appearance learning. Then, tracklet association is performed by minimizing a subgraph decomposition cost. Occlusions and missing detections are recovered after tracklet stitching. Our method has been tested on two public datasets. Experimental results have demonstrated the significant performance improvement of our method. Specifically, the proposed method is improved by 8.31% and 12.38% in terms of Multiple Object Tracking Accuracy and Multiple Object Tracking Precision, respectively, as compared to the baseline.**

***Index Terms*—Semi-online, subgraph decomposition, tracklet association.**

I. Introduction

MULTIPLE object tracking (MOT) is a fundamental problem in signal processing and computer vision [1]. Given the object detection results in all video frames, the task of MOT is to estimate the trajectories of all interested objects in a video sequence. Existing MOT algorithms can be divided into two main categories: online [2]–[5] and offline methods [6]–[9]. Online trackers estimate the objects state once detections in a frame are produced. That is, detection responses have to be linked with existing trajectories reliably frame by frame.

Manuscript received July 17, 2018; accepted July 29, 2018. Date of publication September 27, 2018; date of current version October 6, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 61602499, 61471371, and 61605242, in part by the National Postdoctoral Program for Innovative Talents BX201600172, China Postdoctoral Science Foundation, and Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sanghoon Lee. *(Corresponding author: Yulan Guo.)*

J. Wang, Q. Hu, and W. An are with the College of Electronic Science, National University of Defense Technology, Changsha 410073, China (e-mail: wangjiahui16@nudt.edu.cn; huqingyong15@nudt.edu.cn; anwei@nudt.edu.cn).

Y. Guo is with the College of Electronic Science, National University of Defense Technology, Changsha 410073, China, and also with the School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail: yulan.guo@nudt.edu.cn).

X. Tang is with the Jiangsu Department of Water Resources, Nanjing, China (e-mail: daxixi@sina.com).

This letter has supplementary downloadable material available at http://ieeexplore.ieee.org.

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2018.2872403

Offline trackers aim to link discrete detections in an entire video into long and continuous sequences to obtain trajectories of multiple objects. These trackers can be modeled in a graph model [10], [11], where the nodes of the graph represent the detections in all frames and the edges represent the similarities between two detections. Trajectories can then be represented by the paths cascaded through different nodes. The task of MOT is then formulated as a subgraph decomposition problem [12]–[14].

Since detection hypotheses in all frames have to be considered, the time complexity of offline trackers is very high, which limits their real-time applications. Besides, objects can be similar to each other in crowded scenes, making these algorithms being explored in a large state space. In contrast, online algorithms can be applied to real-time applications [2], [3], [15]. However, they are prone to identity switches (IDS) due to false detections, occlusions, pose changes, and camera motion [13], [16]. Choi [17] measures appearance affinity by aggregated local flow and perform global association in the current window. Recently, deep neural networks have been applied to MOT. It is demonstrated that they can significantly improve the tracking performance [14], [18].

In this letter, we propose a semi-online multiple object tracking method. Our method is inspired by an online MOT method [3], where the MOT problem is solved by online association of tracklets and detections. This method determines tracking output according to single-frame detections in most cases. Consequently, wrong correspondences can be produced by false positives under occlusion, resulting in model drifting. To overcome this limitation, we consider the detections in a temporal window (i.e., a short-time interval). In the window, we firstly generate tracklets (i.e., short-time trajectories) to reduce the state space, and evaluate the appearance feature of different tracklets. We then use the adaptive label iterative conditional modes to associate these tracklets into long trajectories [19]. The tracking process is then moved forward by a window step. The procedure of our method is shown in Fig. 1. Experiments have been conducted on two public datasets including Performance Evaluation of Tracking and Surveillance (PETS) [20] and ETH Mobile Scene (ETHMS) [21], with superior performance being achieved in terms of several evaluation metrics, such as Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP) and False Positives (FP).

The major contributions of this letter can be summarized as follows: (1) The shortcomings of false positives and short trajectories faced by online MOT are addressed by fusing global data association methods in a temporal window. Trajectories can be

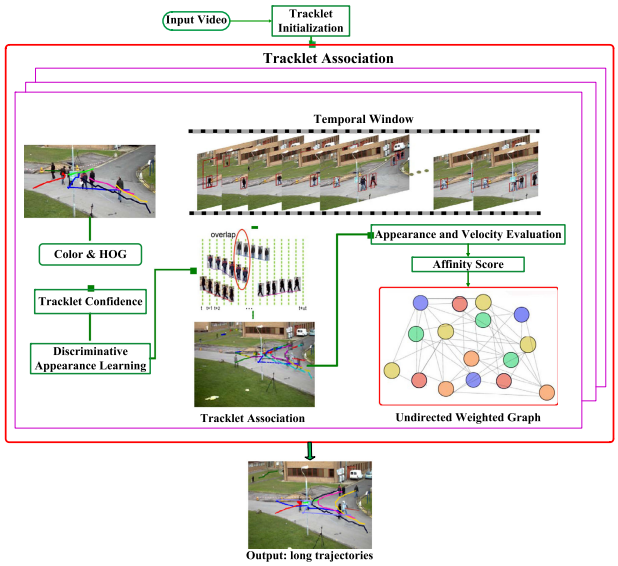


Fig. 1. The procedure of our semi-online MOT method.

recovered from fragments and noisy inputs. Experiments show that a higher tracking precision can be achieved by our method. (2) The descriptors of a tracklet are generated at both ends, and the feature sets which are closest in time are used to calculate the affinity score of tracklets. It is shown that the number of identity switches can be reduced. (3) Discriminant analysis is applied on tracklets to improve the tracklet aggregation performance, and to reduce the calculation time.

II. Appearance Model

In this section, we introduce an online appearance feature extraction method with tracklet confidence estimation and discriminative appearance learning.

A. Tracklet Confidence

A tracklet T_i is represented as the cascade of detections x_t^i in neighboring frames, where x_t^i represents the state of detection i in frame t. Given the set of tracklets $\mathbb{T} = \{T_i\}$, the trajectory hypothesis T can be represented as a set of tracklets ordered in time.

Confidence for each tracklet is calculated as

$$conf\,(T_i) = avg(z_k^i, z_j^i)_{k,j\in[t_s^i,t_e^i]} \times \max((1+\beta\cdot\log((L-\alpha)/L), 0). \quad (1)$$

The first term in (1) denotes the average affinity between detections in an existing tracklet, where (z_k^i, z_j^i) represents two detections z_k and z_j in tracklet T_i. The second term in (1) describes the continuity of tracklet, where $L = |T_i|$ is the length of a tracklet, α is the number of frames where the object is missing, i.e., $\alpha = t_e^i - t_s^i + 1 - L$, t_s^i and t_e^i represent the start and end time of tracklet T_i, respectively. β is a control parameter related to the precision of the detector. β should be set to a large value for a detector with high performance. Tracklet confidence ranges from 0 to 1, we consider a tracklet T_i reliable if $conf\,(T_i) > 0.5$, whereas the remaining tracklets are removed.

B. Online Discriminative Appearance Learning

To learn a discriminative appearance model, we use an image buffer to store 10 latest image patches. For each iteration, we update the image patches of each object by collecting their latest 10 image patches. We normalize each collected sample to 128 × 64 pixels, and then extract its HSV color histogram and HOG feature [22]. Consequently, the training samples are represented as $\mathbb{X} = \{(\boldsymbol{X}_1, \mathrm{y}_1), (\boldsymbol{X}_2, \mathrm{y}_2), \ldots, (\boldsymbol{X}_\mathrm{m}, \mathrm{y}_\mathrm{m})\}$, where X_i and y_i are the feature vector and the identity of a sample, respectively, m is the number of samples. It is necessary to map these high dimensional features into a low dimensional space. Therefore, linear discriminant analysis [23] is used to reduce the dimensionality of features with minimal loss of information. To accommodate appearance changes, the appearance model of each object needs to be updated once the temporal window moves. Then, the appearance model can identify a specific object from other objects more robustly rather than just identifying objects from background.

III. Motion Model

Using apperance feature only is insufficient for tracklet association, therefore, we propose a tracklet motion model to complement the appearance model. Assume that the motion of each tracklet is independent, since these tracklets are usually short, we use linear motion to approximate the velocity of each tracklet. The motion affinity is defined as

$$A_m(T_i,\ T_j) = G(\ p_j^e + v_j^e\Delta t - p_i^s;\ \Sigma_{\Delta t}) \times G(\ p_i^s - v_i^s\Delta t - p_j^e;\ \Sigma_{\Delta t}). \quad (2)$$

That is, the difference between the position predicted by the tracklet and the real position is assumed to follow a Normal distribution $G(\cdot)$ with a variance of $\sum_{\Delta t}$ [24]. In (2), Δt is the time interval between tracklets T_i and T_j, p_j^e is the end position of tracklet T_j, p_i^s is the start position of tracklet T_i, v_j^e is the forward velocity estimated from the start to the end of tracklet T_j, and v_i^s is the backward velocity estimated from the end to the start of tracklet T_i.

IV. Tracklet Association

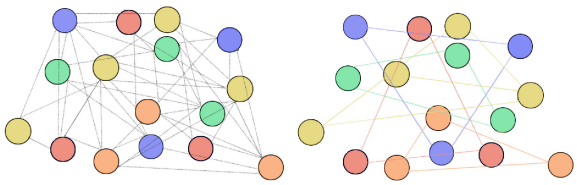
For each tracklet, its appearance feature and velocity are obtained in Sections II and III, which are further used to estimate the affinity between tracklets. Given two tracklets T_i and T_j, the affinity score between these two tracklets is defined as

$$A\,(T_i,\ T_j) = A_a(T_i,\ T_j)\,A_m(T_i,\ T_j), \quad (3)$$

where the motion affinity score $A_m(T_i, T_j)$ is defined in (2), and the appearance affinity score $A_a(T_i, T_j)$ is defined as

$$A_a(T_i,\ T_j) = \frac{\mathbf{W}f(T_i)\cdot\mathbf{W}f(T_j)}{\left\|\mathbf{W}^T f(T_i)\right\|\left\|\mathbf{W}^T f(T_j)\right\|}, \quad (4)$$

where $f\,(T_i)$ and $f\,(T_j)$ are the two feature vectors which are close in time in these two tracklets T_i and T_j, $\mathbf{W}$ is the projection matrix calculated by the discriminate appearance model (Section II). Note that, we compute two appearance feature vectors for each tracklet, i.e., one for the start time and the other for the end time, both of them are calculated by averaging the



(a) Several connection options G (b) Standard sub-graph G_s

Fig. 2. Different tracklets for five objects. Nodes with the same color share the same identity. Edges in the undirected graph (a) show several possible associations between tracklets. (b) is a subgraph of (a), where cracked trajectories belonging to the same object are connected by edges, the weight of an edge represents the affinity of the tracklet pair.

appearance feature vectors in three consecutive frames. To calculate A_a, the time order of two tracklets is first estimated, and the nearest pair of features for these two non-overlapping tracklets are selected. It is based on the observation that, for two cracked tracklets of the same object, the appearance features extracted from detections that are close in time have a high affinity.

Besides, we use a sigmoid function to map the affinity score between two tracklets to $(-1, 1)$:

$$g = \frac{2}{1 + \exp(-\lambda(A - \mu))} - 1, \tag{5}$$

where λ is a decaying factor to control the gradient of the transition curve, μ is a shift factor to control the threshold. We set $\lambda = 1$ and $\mu = 0.25$ in this letter.

A. Undirected Weighted Graph

We now formulate the MOT task as a graph partition problem. The tracklets in a temporal window are represented by nodes V and the pairs of tracklets are connected by edges E. For each pair of nodes $u, v \in V$, we use w_{uv} to represent their affinity weight, $w_{uv} > 0$ indicates that the two tracklets u and v are possible to associate with each other. In this case, an edge e_{uv} is added to connect these two nodes. In contrast, $w_{uv} < 0$ indicates that the two tracklets conflict to each other.

All tracklets in the current window and their affinity can be expressed in a correlation graph G, as shown in Fig. 2(a). We aim to partition the graph G to a standard subgraph, where all nodes in a subgraph share the same identity label, as shown in Fig. 2(b). We link all the tracklets with the same identity in time order, and then infer the states under occlusion and missing detections.

Here, the graph partition problem maximizes the sum of affinity weights w_{uv} assigned to these edges. It can be formulated as a binary linear program:

$$\arg\max \sum_{e_{uv} \in E} w_{uv} e_{uv}, \tag{6}$$

subject to

$$e_{uv} \in \{0, 1\}, \qquad \forall e_{uv} \in E, \tag{7}$$

$$e_{uv} + e_{vt} \leqslant 1 + e_{ut} \quad \forall (u, v), (v, t), (u, t) \in E. \tag{8}$$

The optimal solution for (6) corresponds to the standard subgraph. Formula (8) guarantees the transmissibility in the labels of nodes. That is, if detections u and v have the same label, and detections v and t have the same label, then detections u and t have the same label.

We use the adaptive label iterative conditional modes algorithm [19] to solve the large-scale data association problem defined in (6). The graph partitions are described by using a label vector $I = \{1, 2, 3, \ldots, K\}$, where I represents the identities of nodes, K represents the number of subgraphs. We use the affinity score between tacklets as the distance between nodes. During iteration, conditioned on the current label, each node is assigned the label of its nearest node, or with a new label if it is rejected by all existing nodes. That is, a new label can be assigned when a tracklet belongs to a new incoming object. This method can adaptively determine the number of labels K, i.e., the number of objects. Given a label vector I, the problem is rephrased as an energy function for conditional random:

$$C(I) = \sum_{u,v} w_{uv} \cdot 1_{[I_u \neq I_v]}, \tag{9}$$

where $w_{u,v}$ represents the affinity weight between the pair of nodes u and v, binary variable $1_{[I_u \neq I_v]}$ is equal to 1 if $I_u \neq I_v$. Otherwise, $1_{[I_u \neq I_v]}$ is equal to 0. Note that, if two nodes u and v have a positive affinity, they are expected to have the same label, i.e., $I_u = I_v$, and there is no cost. Otherwise, they are assigned to different labels, and a cost is assigned, which is higher for a larger $w_{u,v}$. As a consequence, minimizing (9) is equivalent to solving (6).

V. Experiments

In this section, we present the experimental results of the proposed semi-online MOT method. We also compare our method with several state-of-the-art methods.

A. Implementation

We implemented the proposed semi-online MOT method in MATLAB on a PC with a 2.2 GHz CPU and 32 GB memory. All parameters were tuned empirically, we set the length of our temporal window to 40 frames.

B. Evaluation Metrics

Similar to [29], [30], MOTA (↑) and MOTP (↑) are used to compare the performance of different MOT algorithms. MOTA combines three different errors, i.e., False Negatives (FN ↓), False Positives (FP ↓), and Identity Switching (IDS ↓). MOTP is calculated as the overlap between the groundtruth tracks and the estimated tracks. It shows the ability of a tracker to estimate the precise object positions. We also used the number of all Groundtruth Trajectories (GT), the number of Mostly Tracked trajectories (MT), the number of Mostly Lost trajectories (ML), Recall (↑), and Precision (↑) for performance evaluation. Note that, for measures denoted by (↑), a higher score indicates better performance.

C. Quantitative Results

PETS2009 S2L1: We first evaluate our method on the PETS2009 S2L1 sequence. PETS2009 contains pedestrians

TABLE I
PERFORMANCE COMPARISON ON SEVERAL SEQUENCES. '-' MEANS THAT THE CORRESPONDING ITEM IS UNAVAILABLE IN THE LITERATURE

PETS (S2L1)	CEM [25]	80.20%	90.60%	12.15%	2.37%	11	**98.4%**	92.4	23	21	1
	CONF [26]	56.30%	79.70%	–	–	11	–	–	23	–	–
	ALEx [27]	71.40%	67.70%	24.26%	26.07%	193	–	–	23	–	–
	OGOMT [28]	80.50%	**98.10%**	–	–	9	–	–	23	23	0
	baseline [3]	83.04%	69.59%	**1.19%**	19.41%	**4**	–	–	23	23	0
	proposed	**93.35%**	81.97%	4.25%	**3.91%**	5	96.07%	**95.61%**	**23**	**23**	0
PETS (S2L2)	CEM [25]	56.90%	59.40%	27.90%	6.04%	99	89.80%	**65.50%**	74	28	12
	CONF [26]	50.00%	51.30%	–	–	–	–	–	74	–	–
	ALEx [27]	27.50%	70.60%	59.78%	4.36%	385	–	–	74	–	–
	OGOMT [28]	66.00%	64.80%	–	–	181	–	–	74	33	**0**
	baseline	**70.12%**	53.92%	**14.99%**	14.35%	**45**	–	–	74	53	1
	proposed	67.21%	**71.30%**	33.04%	**1.83%**	60	**96.98%**	59.67%	**74**	**60**	8
ETHMS BAHNHOF	baseline [3]	72.03%	64.01%	**23.64%**	4.16%	18	–	–	126	93	**3**
	proposed	**76.96%**	**70.29%**	32.19%	**2.28%**	**16**	**86.58%**	**71.27%**	126	96	10
MOT2015 (TUD Campus)	ALEx [27]	35.7%	65.6%	40.67%	13.21%	23	–	–	8	1	1
	CONF [26]	**73.3%**	67.0%	**26.40%**	**0.1%**	**2**	–	–	8	–	–
	baseline [3]	51.23%	70.46%	36.58%	39.58%	3	–	–	8	–	–
	proposed	62.58%	**71.42%**	29.36	12.82%	3	85.56%	63.09%	8	2	1
MOT2015 (TUD Stadtmitte)	ALEx [27]	53.8%	65.6%	24.8%	10.42%	40	86.70%	84.70%	9	5	0
	CEM [25]	**72.2%**	65.5%	**7.7%**	**6.56%**	**4**	–	–	9	7	0
	baseline [3]	45.23	70.46%	27.48%	18.63%	16	–	–	9	–	–
	proposed	67.42	**71.64%**	12.56%	15.32%	6	84.57%	79.65%	9	7	0

walking across an intersection in various directions with variable speeds. For a fair comparison, we used the same detections and groundtruth given by [25] for the experiments. The quantitative results are shown in Table I. We compare our method with those methods using the same tracking-by-detection paradigm as ours [25]–[27]. As expected, our method achieves better performance on the majority of metrics than existing online association methods. Our FP is lower than CEM [25], CONF [26], ALExTRAC [27], OGOMT [28], and the baseline [3]. Besides, the proposed semi-online method can estimate the positions of objects more accurately. For example, the MOTP of our method is improved by 12.38% as compared to the baseline [3]. It is clear that our method achieves remarkable performance on most evaluation metrics.

PETS2009 S2L2: We further test our method on the PETS2009-S2L2 sequence. This sequence is more challenging, with a large number of pedestrians (up to 40) and significant inter-object occlusions. The tracking results are shown in Table I. The proposed method achieves the best performance in terms of MOTP, FN, FP, precision and MT. Note that, the proposed method reduces FP from 14.35% to 1.83% as compared to the baseline [3].

Other Sequences: To demonstrate the generality of our method on the moving platforms, we test our method on the ETHMS dataset [31], TUD_campus and TUD_stadtmitte sequences of MOT2015 dataset [32]. These datasets contain lots of occlusions and interactions due to the low viewing angle of the moving cameras. The evaluation results are shown in Table I. We can see that, our method achieves the best performance in terms of MOTA, MOTP, FP and ML.

D. Window Length Setting

We further tested the performance of the proposed method under different settings of window length. Experimental results achieved on the PETS2009 S2L1 sequence are shown in Fig. 3. It is clear that, our method is similar to online tracking process when the window length is 2. The performance is improved significantly when the window length increases from 2 to 20.

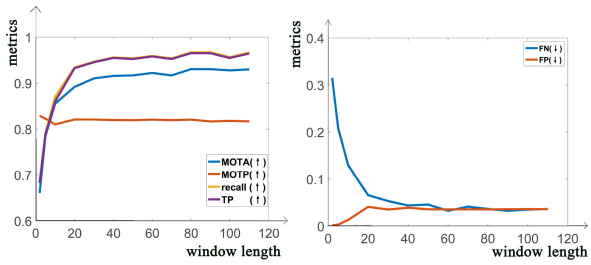


Fig. 3. The performance achieved by the proposed MOT method under different settings of window length.

That is because, by considering all detections in a window, the detections of certain objects in neighboring frames can be associated. Besides, objects under occlusion can be recovered. This clearly demonstrates the effectiveness of our tracklet association method. Then, the performance is saturated when the window length is increased beyond 40. That is because, a window with a length of 40 is sufficiently long to cover most occlusions. The performance achieved on the PETS2009 S2L2 and ETHMS sequences is similar to that achieved on PETS2009 S2L1. To achieve a balance between the accuracy and tracking speed, the window length of our method is set to 40 in this work. Our method takes 0.40, 1.35 and 0.87 seconds per frame on the sequences S2L1, S2L2 and ETHMS, respectively.

VI. CONCLUSION

This letter has presented a method for semi-online multiple object tracking. Initial tracklets are first obtained by connecting similar detections in neighboring frames. The appearance model is then estimated using online discriminative appearance learning and nearest appearance feature extraction. The appearance model is further combined with the motion model to estimate the affinity scores between tracklets. Finally, tracklets are stitched into long trajectories by using a graph model. Our MOT method has been tested on several public datasets. Experimental results have shown a clear improvement of the proposed method over the state-of-the-art.

REFERENCES

[1] Q. Hu, Y. Guo, Z. Lin, W. An, and H. Cheng, "Object tracking using multiple features and adaptive model updating," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 11, pp. 2882–2897, Nov. 2017.

[2] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4705–4713.

[3] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1218–1225.

[4] R. Yao, "Robust model-free multi-object tracking with online kernelized structural learning," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2401–2405, Dec. 2015.

[5] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 1, pp. 260–267.

[6] C. Wang, H. Liu, and Y. Gao, "Scene-adaptive hierarchical data association for multiple objects tracking," *IEEE Signal Process. Lett.*, vol. 21, no. 6, pp. 697–701, Jun. 2014.

[7] B. Wang, G. Wang, K. L. Chan, and L. Wang, "Tracklet association by online target-specific metric learning and coherent dynamics estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 589–602, Mar. 2017.

[8] B. Yang and R. Nevatia, "Multi-target tracking by online learning a CRF model of appearance and motion patterns," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 203–217, 2014.

[9] Y. Yi and H. Xu, "Hierarchical data association framework with occlusion handling for multiple targets tracking," *IEEE Signal Process. Lett.*, vol. 21, no. 3, pp. 288–291, Mar. 2014.

[10] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Subgraph decomposition for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5033–5041.

[11] L. Wen, Z. Lei, S. Lyu, S. Z. Li, and M.-H. Yang, "Exploiting hierarchical dense structures on hypergraphs for multi-object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 1983–1996, Oct. 2016.

[12] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3539–3548.

[13] E. Levinkov *et al.*, "Joint graph decomposition & node labeling: Problem, algorithms, applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1904–1912.

[14] S. Tang, B. Andres, M. Andriluka, and B. p. Schiele, "Multi-person tracking by multicut and deep matching," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 100–111.

[15] Q. Hu, Y. Guo, Y. Chen, and J. Xiao, "Correlation filter tracking: Beyond an open-loop system," in *Proc. British Mach. Vis. Conf.*, 2017, pp. 1–12.

[16] U. Iqbal, A. Milan, and J. Gall, "Pose-track: Joint multi-person pose estimation and tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4654–4663.

[17] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 3029–3037.

[18] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "Poi: Multiple object tracking with high performance detection and appearance feature," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 36–42.

[19] S. Bagon and M. Galun, "Large scale correlation clustering optimization," 2011, arXiv:1112.2903.

[20] J. Ferryman and A. Shahrokni, "Pets2009: Dataset and challenge," in *Proc. 12th IEEE Int. Workshop Perform. Eval. Tracking Surveillance*, 2009, pp. 1–6.

[21] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *Proc. IEEE Conf. Comput. Vis.*, 2009, pp. 1–8.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[23] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[24] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 788–801.

[25] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 58–72, Jan. 2014.

[26] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1820–1833, Sep. 2011.

[27] A. Bewley, L. Ott, F. Ramos, and B. Upcroft, "Alextrac: Affinity learning by exploring temporal reinforcement within association chains," in *Proc. Int. Conf. Robot. Autom.*, 2016, pp. 2212–2218.

[28] H. Possegger, T. Mauthner, P. M. Roth, and H. Bischof, "Occlusion geodesics for online multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1306–1313.

[29] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, no. 1, pp. 1–10, 2008.

[30] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2953–2960.

[31] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "Robust multiperson tracking from a mobile platform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1831–1846, Oct. 2009.

[32] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," 2015, arXiv:1504.01942.