Multi-Object tracking using Multi-Channel Part Appearance Representation

Thi Lan Anh Nguyen, Furqan M.Khan, Farhood Negin, François Bremond

Multi-Object tracking using Multi-Channel Part Appearance Representation

Nguyen Thi Lan Anh Furqan M.Khan Farhood Negin Francois Bremond
INRIA Sophia Antipolis
2004 Route des Lucioles -BP93 Sophia Antipolis Cedex, 06902 - France
{thi-lan-anh.nguyen|furqan.khan|farhood.negin|francois.bremond}@inria.fr

Abstract

Appearance based multi-object tracking (MOT) is a challenging task, specially in complex scenes where objects have similar appearance or are occluded by background or other objects. Such factors motivate researchers to propose effective trackers which should satisfy real-time processing and object trajectory recovery criteria. In order to handle both mentioned requirements, we propose a robust online multi-object tracking method that extends the features and methods proposed for re-identification to MOT. The proposed tracker combines a local and a global tracker in a comprehensive two-step framework. In the local tracking step, we use the frame-to-frame association to generate online object trajectories. Each object trajectory is called tracklet and is represented by a set of multi-modal feature distributions modeled by GMMs. In the global tracking step, occlusions and mis-detections are recovered by tracklet bipartite association method based on learning Mahalanobis metric between GMM components using KISSME metric learning algorithm. Experiments on two public datasets show that our tracker performs well when compared to state-of-the-art tracking algorithms.

1. Introduction

Multi-object tracking (MOT) has been one of the fundamental problems in computer vision, essential for lots of applications (e.g home-care, house-care, security systems, etc.). The main objective of MOT is to estimate the states of multiple objects while identifying these objects under appearance and motion variation in time. This problem becomes more challenging due to frequent occlusion by background or other objects, object pose as well as illumination variation, etc.

Depending on the time of data association process, tracking algorithms can be categorized into 2 types: local and global tracking. Local trackers [17, 20] associate object detections in current frame with the best matching object trajectories in the past. These methods are capable of online processing based on frame-to-frame association and therefore, could be applied in real-time applications. In general, local trackers use bipartite matching methods for short-term data association where Hungarian algorithm is the most popular method. Although these methods are computationally inexpensive, object identification could fail due to inaccurate detections (false alarms) and only short-term occlusions can be handled. Global trackers [28, 18] can overcome the shortcomings of local trackers by extension of the bipartite matching into network flow. The direct acyclic graph in [28] was formed where vertices are object detections or short tracklets and edges are the similarity links between vertices. In [18], the track of a person forms a clique and MOT is formulated as constraint maximum weight clique graph. The data association solutions for these global tracker are found through minimum-cost flow algorithm. However, global tracking methods also have their obvious drawbacks, such as: high computational cost due to iterative association process to generate globally optimized tracks and with the requirement of entire object detection in a given video.

Recently, some proposed trackers tried to combine both local and global tracking methods in a framework to perform online object tracking. The MOT methods in [1, 16] use the frame-to-frame association to generate tracklets followed by a tracklet association process with a time buffer latency. However, their performance is limited by their object features and tracklet representation. These methods utilize basic features (e.g. 2D information, color histogram or constant velocity) applied on whole body parts and use Gaussian distribution to describes the object. This way of representation could lose important information to discriminate objects and consequently, could fail to track objects in complex scene conditions (such as occlusion, low video resolution or insufficient lighting of environment).

On the other hand, multiple-shot person re-identification methods [11, 27, 15] gained high performances in matching objects from different camera views. In order to match a query person to the closest person in a gallery, these re-identification methods use efficient features and object rep-

978-1-5386-2939-0/17/$31.00

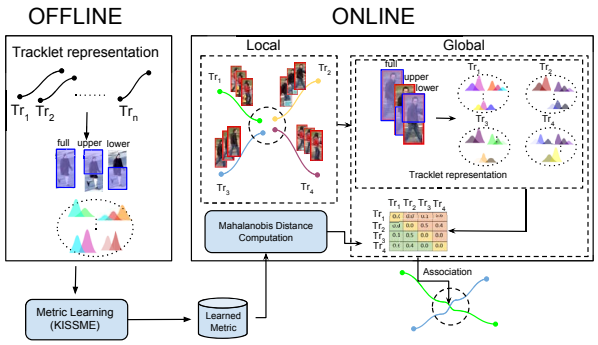

Figure 1. The proposed framework.

resentations. These methods are applied to solve problems that involve pose and camera view setting variation.

In this paper, we propose a robust online multi-object tracking method which extends object representation and methods proposed for re-identification domain to address problems in MOT. This method integrates a local and global trackers in a comprehensive framework. The local tracker generates object trajectories called tracklets. Object features are computed for full and body parts, then, each tracklet is represented by a set of multi-modal feature distribution modeled by GMMs. The global tracker associates tracklets after mis-detections or occlusions based on learning Mahalanobis distance between GMM components. In order to learn this metric, KISSME [8] algorithm is adopted to learn feature transformations between different scenes by directly learning transformation between probability distributions.

The rest of the paper is organized as follow: Section 2 presents the details about the structure and flows of the mentioned two-step comprehensive tracking framework. Section 3 evaluates the robustness of our method by comparing its performance with other state-of-the-art trackers. Finally, section 4 concludes the paper.

2. Proposed framework

Figure 1 illustrates the proposed framework consisting of two offline and online blocks. In the offline block, the framework learns the similarity metric between the tracklets using data in the training set. Once the similarity metric is learned, the two-step online block underlines the interaction between local (frame-to-frame) and global trackers. The local tracker's objective is to find correct object trajectories in the past while, the global tracker tries to find object associations between aggregated tracklets. In the first step, the tracklets are constructed by putting together frame-to-frame tracker's output. For a reliable tracklet representation tracklet filtering is applied by splitting spatially disconnected or occluded tracks and filtering out noisy tracklets. In the second step, in every video segment Δt, the global tracker carries out data association and performs online tracklet matching. Association and matching process happens based on Mahalanobis metric among representations of tracklets stacked in two previous video segments ($2\Delta t$).

2.1. Tracklet Initiation and Filtering

We define tracklet Tr_i between consecutive frames $< m, n >$ as a chain of tracked objects called nodes N_i where i represents the ID of object and N represents the object bounding-box.

$$Tr_i = \{N_i^m, N_i^{m+1}, ..., N_i^{n-1}, N_i^n\} \quad (1)$$

alternatively, a tracklet is a connected sequence of tracked object's bounding-boxes, created using output of the local tracker. In order to provide the global tracker with reliable set of tracklets, the initial tracklets should qualify a filtering process. If due to mis-detection, an inconsistency in tracklet initialization is observed, the tracklet gets re-evaluated. If distance of two bounding-boxes in two consecutive frames was larger than a threshold, a split operation on the tracklet will be followed. In another scenario, if two adjacent tracklets got occluded by each other, split operation is applied on overlapped object detection. And also, if a tracklet's length was smaller than a threshold, the tracklet is considered as noise and is eliminated from further process. Moreover, for every created tracklet, a set of relationships with other tracklets is defined. If the time intervals of the tracklets are overlapped, the relation between those tracklets is called "Neighbour" meaning that they cannot associate with each other. On the contrary, a "Candidate" relation occurs when a tracklet has no time overlap and has close spatial distance with the given tracklet. This tracklet can potentially associate with the given tracklet.

2.2. Tracklet representation

Inspired by person Re-identification approach in [15], appearance of a tracklet is modeled as a multi-channel appearance mixture (appearance model). The representation divides body into three parts: full, upper and lower. Each channel in the mixture model corresponds to a particular body part. Since person Re-identification usually deals with identification of a person from different camera views, it is expected that using Re-id representation becomes even more effective in single-view multi-object tracking problem.

Due to different illumination conditions and arbitrary pose of a person with respect to camera view, the representation should stay invariable to these changes in order to have effective object tracking. By accepting variability as a natural property of appearance, we deal with it as a multi-modal probability distribution of features. To be invulnerable against occlusion, the appearance models are created

independently, one for each part of the detection bounding-box (full, upper and lower part of the bounding-box).

Given a set of nodes (detection bounding-boxes) belonging to tracklet Tr_i, its representation Q is defined as a set of appearance models M_p^i: $Q = \{M_p^i \mid p \in \{full, upper, lower\}\}$. Each appearance model in the set is a multivariate GMM distribution of low-level features of part p. Since fitting a GMM with full covariance matrix to limited number of points and high dimensional features is difficult, the covariance matrices are restricted to be diagonal.

Appearance models help to overcome occlusion, pose variation and illumination problems. Unlike feature pruning methods that are problem specific, we create models with different features without pruning. Although this can cause redundancy in feature representation but the features are computed efficiently to be shared between the parts (upper and lower body regions are defined as 60% of bounding-box of the person). To describe an object we use appearance features that are locally computed on spatial grid of object detection bounding-boxes, including: HOG[3], LOMO[11], HSCD[27] and Color histogram features where LOMO and HSCD features have never been applied in MOT domain. While the framework exploits HOG feature as a shape-based feature to overcome difficulties of pose variation, it benefits from other features to cope with illumination and appearance changes happening in long occlusions.

2.3. Learning mixture parameters

For each body part p of each object with ID *i*, the parameters of the appearance model M_p^i are learned independently. There is no *a priori* knowledge about the number of mixture components –modes– of a person, therefore, both discovery of the modes and description of them using low-level features need to be addressed.

People appearing in a video have different appearance and produce GMMs with variable number of components. Therefore, the number of components are not *a priori* determined and need to be retrieved. In order to infer the number of GMM components for each appearance model automatically, Akaike Information Criterion (AIC) model selection is used. Knowing fixed number of the components, the parameters of a GMM could be learned conveniently using Expectation-Maximization method.

2.4. Similarity metric for tracklet representations

Similarity metric plays an essential role in comparing two candidate tracklets' representations. Similarity of two tracklet representations is defined as the sum of similarities between the corresponding appearance models. Given the distance between two appearance models $d(M_1, M_2)$ of tracklet representations Q_i and Q_j, we can convert this distance into similarity using Gaussian similarity kernel:

$$Sim(Q_i, Q_j) = \sum_{p \in P} \exp\left(- \frac{\overline{d(M_p^i, M_p^j)} - \gamma_{p,j}}{\frac{1}{3}(\beta_{p,j} - \gamma_{p,j})} \right) \quad (2)$$

where $P = \{full, upper, lower\}$. $\overline{d(M_p^i, M_p^j)}$ is max normalized distance between tracklet representations Q_i and Q_j of part p. $\beta_{p,j}$ and $\gamma_{p,j}$ are the maximum and minimum normalized distance between tracklet representation Q_j and representations of all other tracklets, respectively. The factor of $\frac{1}{3}$ in formula makes Gaussian similarity kernel to be zero for tracklet representation Q_i that has maximum normalized distance from tracklet representation Q_j.

The distance between two appearance models is defined as a sum of distance between GMM components weighted by their prior probabilities:

$$d(M_1, M_2) = \sum_{i=1:k_1, j=1:k_2} \pi_{1i} \pi_{2j} d(G_{1i}, G_{2j}) \quad (3)$$

where G_{nk} is the component k of GMM $M_{n \in \{1,2\}}$ with corresponding prior π_{nk}.

To compute the distance between two GMMs we learn Mahalanobis metric between them. For a pair of vectors $x_{ij} = (x_i, x_j)$, squared Mahalanobis distance is defined as:

$$d^2(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j) \quad (4)$$

where $M \succeq 0$ is a positive semidefinite matrix.

Some of popular algorithms to learn matrix M from a set of vector pair $X = \{x_{ij} | i = 1 : m, j = 1 : n\}$ are LMNN [23], ITML [4] and KISSME [8]. However, for our experiments, we use KISSME [8] for its simplicity, low computation cost and effectivess under challenging conditions. KISSME algorithm assumes independent Gaussian generation processes with parameters $\theta^+ = (0, \Sigma^+)$ and $\theta^- = (0, \Sigma^-)$ for positive and negative pairs (x_i, x_j), respectively, based on their difference vector $x_i - x_j$. Given pair associations, the feature-difference covariance matrices can be computed as follow:

$$\Sigma_+ = \frac{1}{N^+} \sum_{x_{ij} = X^+} (x_i - x_j)(x_i - x_j)^T \quad (5)$$

$$\Sigma_- = \frac{1}{N^-} \sum_{x_{ij} = X^-} (x_i - x_j)(x_i - x_j)^T \quad (6)$$

where x_i and x_j are means of two Gaussian distributions and $x_{ij} \in \{X^+, X^-\}$ is ground truth similarity labels, X^+ and X^- are positive and negative sample sets while N^+ and N^- are their sizes, respectively. To obtain samples, we divide every object trajectory in the ground truth into two equal parts, where, each one of the parts is considered as

positive sample of the counterpart and negative samples of a tracklet are all of the other trajectories.

Given the covariance metrices, Mahalanobis metric with matrix $M(M = \Sigma^{+-1} - \Sigma^{--1})$ is learned independently from appearance models of each part and similarity between the Gaussian distribution pairs are retrieved using similarity labels of tracklet pairs.

2.5. Data association

In the online phase, the framework tries to calculate the global matching scores of a tracklet with candidates in its relationship set in every video segment Δt. Similarity matrix S=$\{m_{ij}\}$ is constructed with calculated scores between all of candidates, where i=1..n, j=1..n, and n is the number of tracklets in current time interval [$t-2\Delta$t,t]. If tracklet j is in the candidate list of tracklet i, the similarity of the pair is calculated using Mahalanobis metric $m_{ij} = Sim(Q_i, Q_j)$, otherwise it is set to zero in the similarity matrix. Once the cost matrix is computed, the optimal association pairs, which minimize the data association cost in S, are determined using Hungarian algorithm.

3. Experiments and results

Our approach is evaluated on two challenging benchmark datasets. A local tracker using different object appearance features [2] is selected to validate our framework. We use a publicly available detection and evaluation method to have a fair comparison with other state-of-the-art trackers.

3.1. Datasets

We have evaluated our tracker's performance on two challenging benchmark datasets: MOT challenge (2DMOT2015) and ParkingLot.

- *2DMOT2015* includes 11 sequences, 5783 frames with 721 objects. This dataset shows a large variety of outdoor scenes with strong and frequent person-person occlusions, people moving in random directions captured by fixed or moving low angle cameras and crowded environment (two of the sequences have 197 and 226 moving humans, respectively). Both dataset and person detection information are available on MOTchallenge website[1].

- *ParkingLot* The main challenge of this dataset is occlusion and confusion caused by objects walking closely with similar appearance. We choose Parkinglot1 sequence including 14 objects in 998 frames for testing because it is one of the most popular benchmarks for tracking evaluation and its detection is available on UCF website [2].

[1]https://motchallenge.net/

[2]http://crcv.ucf.edu/data/ParkingLOT/

3.2. Evaluation metrics

We use the common CLEAR MOT consisting of multiple metrics and follow publicly provided toolkit on MOT challenge website for a fair comparison with other approaches. The multiple object tracking precision (MOTP↑) evaluates the intersection area over the union area of the bounding boxes. The multiple object tracking accuracy (MOTA↑) calculates the accuracy composed of false negatives (FN↓), false positives (FP↓), and identity switching (IDSw↓). In addition, the tracking-time metrics are computed: the number of trajectories in ground-truth (GT), the ratio of mostly tracked trajectories (MT↑), the ratio of mostly lost trajectories (ML↓) and the number of track fragments (Frag↓). Here, ↑ indicates that higher scores correspond to better results, and ↓ shows that lower scores correspond to better results.

3.3. System parameters

All parameters have been found experimentally, and remained unchanged for benchmark datasets. The same threshold $\theta = 0.3$ is used for all of the data association process. The size of a video segment is fixed to 15 frames. The minimum size of a tracklet is set to 3.

3.4. Performance evaluation

A quantitative comparison between our approach and thirteen state-of-the-art tracking methods on challenging 2DMOT2015 dataset is shown in table 1. In evaluation part, we also categorize state-of-the-art trackers into two groups: Offline and online tracking. In order to emphasis the robustness of proposed approach which satisfies both requirements of online processing and high tracking performance, we show that our method not only outperforms online methods, but also has comparable performances compared to offline ones.

Our approach outperforms both online and offline methods when metrics ML and FN are used. In detail, our approach misses the least number of objects shown by metric ML and keeps track of highest number of objects, shown by the lowest number of false negatives in metric FN. The results on these two metrics illustrate the impressive improvement of our method compared to others. We reduce nearly one-forth the number of lost objects compared to methods [9, 10, 6, 26, 1] and nearly a half compared to [5] on metric ML. With metric FN, the number of false negatives in our method is reduced at least by 2,379 compared to [14] and the most by 11,421 compared to [5]. According to metric MT, our tracker performs remarkably better than trackers [10, 6, 19, 7, 13, 26, 5, 1] and in total has the second best performance. However, the best tracker [22] evaluated by this metric works only in offline mode.The proposed method achieves comparable results on metric MOTP but is

Methods	Trackers	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDSw↓	Frag↓
Offline	CNNTCM [22]	**29.6± 13.9**	**71.8**	**11.2**	44.0	7,786	34,733	712	**943**
	SiameseCNN [9]	29.0 ±15.1	71.2	8.5	48.4	**5,160**	37,798	**639**	1,316
	ELP [12]	25.0 ±10.8	71.2	7.5	**43.8**	7,345	37,344	1,369	1,804
	MotiCon [10]	23.1 ±16.4	70.9	4.7	52.0	10,404	35,844	1,018	1,061
	CEM [14]	19.3 ±17.5	70.7	8.5	46.5	14,180	**34,591**	813	1,023
	TBD [6]	15.9 ±17.6	70.9	6.4	47.9	14,943	34,777	1,939	1,963
Online	SCEA [25]	**29.1 ±12.2**	**71.1**	8.9	47.3	**6,060**	36,912	604	1,182
	EAMTTpub [19]	22.3 ±14.2	70.8	5.4	52.7	7 ,924	38,982	833	1,485
	OMT_DFH [7]	21.2 ±17.2	69.9	7.1	46.5	13,218	34,657	563	1,255
	RNN_LSTM [13]	19.0 ±15.2	71.0	5.5	45.6	11,578	36,706	1,490	2,081
	RMOT [26]	18.6 ±17.5	69.6	5.3	53.3	12,473	36,835	684	1,282
	GSCR [5]	15.8 ±10.5	69.4	1.8	61.0	7,597	43,633	**514**	**1,010**
	TC_ODAL [1]	15.1 ±15.0	70.5	3.2	55.8	12,970	38,538	637	1,716
	MTS (Ours)	20.6 ±18.7	70.3	**9.0**	**36.9**	15,161	**32,212**	1,387	2,357

Table 1. Quantitative Analysis of our method on MOT15 challenging dataset with state-of-the-art methods. The tracking results of these methods are public on MOTchallenge website. The best values in both online and offline methods are marked in **bold**.

Trackers	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	ID Sw↓	Frag↓
PMPT[20]	79.3	74.1	-	-	-	-	-	-
H2T [24]	88.4	81.9	78.57	0	-	-	21	-
GMCP [18]	90.43	74.1	-	-	-	-	-	-
MTS(Ours)	84.5	74.4	78.57	0	325	925	7	99

Table 2. Quantitative Analysis of our method on ParkingLot1 with state-of-the-art methods.The tracking results of these methods are public on UCF website.

not impressive in on metrics MOTA, FP, IDWs and Frag compared to the others from the state-of-the-art.

However, looking at the table 1, we can see that trackers have best results on some metrics but not on all of the metrics. According to the analysis in [21], trajectory-based metrics, including metrics MT and ML, show the ratios of ground-truth trajectory's life span are covered by a track hypothesis (at least 80% for MT and at most 20% for ML, respectively). MT and ML are not influenced by the number of Frag or IDSw. As a result, these metrics give more information about the coverage of the trajectories rather than the ability of the tracker to reproduce them. On the other hand, results on metrics FP, FN, MOTA and MOTP are sensible to detector errors. Particularly, metrics FP and FN are computed by detector precision and recall, while metric MOTA and MOTP show how much a tracker is able to find target positions and reject false alarms proposed by the detector. Therefore, in terms of tracking performance evaluation, trajectory-based metrics (MT and ML) are proved to be more reliable than the others.

With Parkinglot dataset, we use the MOT evaluation toolkit to compare our tracking performance with publicly annotated data. The results of our tracker and others from state-of-the-art are shown in table 2. Only tracker [24] and ours are evaluated by MT, ML and IDSw. Both methods have the same performances on MT, ML. While [24] has higher results than ours on metrics MOTA and MOTP, our method reduces two-third of IDSw errors. [18] is evaluated only using MOTA and MOTP metrics. The performance of this method is better than ours on MOTA but it performs worse when using MOTP. With [20], on both metrics MOTA and MOTP, our method has better performances in comparison.

4. Conclusions

We have proposed a robust multi-object tracking method which integrates a local and a global tracker into a two-step comprehensive framework. The proposed method works in online mode and is suitable for real-time applications. It also effectively addresses some of the highly challenging problems in MOT such as mis-detection, object appearance changes by occlusion, pose or illumination variations, etc. To do so, the extension of object appearance features and metric learning methods proposed for re-identification domain are effectively adapted to MOT. The effectiveness and robustness of our method are verified by extensive experiments and comparison with state-of-the-art trackers is reported.

Acknowledgement

This work is supported by the Provence-Alpes-Cote d'Azur Region, SafEE and Movement projects.

References

[1] S. H. Bae and K. J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *2014 CVPR*, pages 1218–1225, June 2014.

[2] B. Chau.D.P and Thonnat.M. A multiple feature tracking algorithm enabling adaptation to context variations, 2011. ICDP.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005.

[4] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 209–216, New York, NY, USA, 2007. ACM.

[5] L. Fagot-Bouquet, R. Audigier, Y. Dhome, and F. Lerasle. Online multi-person tracking based on global sparse collaborative representations. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2414–2418, Sept 2015.

[6] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):1012–1025, May 2014.

[7] J. Ju, D. Kim, B. Ku, D. K. Han, and H. Ko. Online multi-object tracking with efficient track drift and fragmentation handling. *J. Opt. Soc. Am. A*, 34(2):280–293, Feb 2017.

[8] M. Kstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *2012 CVPR*, pages 2288–2295, June 2012.

[9] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler. Learning by tracking: Siamese CNN for robust target association. *CoRR*, abs/1604.07866, 2016.

[10] L. Leal-Taix, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese. Learning an image-based motion context for multiple people tracking. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3542–3549, June 2014.

[11] S. Liao, Y. Hu, and S. Z. Li. Joint dimension reduction and metric learning for person re-identification. *CoRR*, abs/1406.4216, 2014.

[12] N. McLaughlin, J. M. D. Rincon, and P. Miller. Enhancing linear programming with motion modeling for multi-target tracking. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 71–77, Jan 2015.

[13] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler. Online multi-target tracking using recurrent neural networks. In *AAAI*, February 2017.

[14] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):58–72, Jan 2014.

[15] F. M.Khan and F. Bremond. Multi-shot person re-identification using part appearance mixture. WACV.

[16] T. L. A. Nguyen, F. Bremond, and J. Trojanova. Multi-object tracking of pedestrian driven by context. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 23–29, Aug 2016.

[17] F. Poiesi, R. Mazzon, and A. Cavallaro. Multi-target tracking on confidence maps: An application to people tracking. *Computer Vision and Image Understanding*, 117(10):1257 – 1272, 2013.

[18] A. Roshan Zamir, A. Dehghan, and M. Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

[19] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro. *Online Multi-target Tracking with Strong and Weak Detections*, pages 84–99. Springer International Publishing, Cham, 2016.

[20] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1815–1821, June 2012.

[21] F. Solera, S. Calderara, and R. Cucchiara. Towards the evaluation of reproducible robustness in tracking-by-detection. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Aug 2015.

[22] B. Wang, L. Wang, B. Shuai, Z. Zuo, T. Liu, K. Luk Chan, and G. Wang. Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.

[23] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, June 2009.

[24] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Z. Li. Multiple target tracking based on undirected hierarchical relation hypergraph. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[25] J. H. Yoon, C. R. Lee, M. H. Yang, and K. J. Yoon. Online multi-object tracking via structural constraint event aggregation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1392–1400, June 2016.

[26] J. H. Yoon, M. H. Yang, J. Lim, and K. J. Yoon. Bayesian multi-object tracking using motion context from multiple objects. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 33–40, Jan 2015.

[27] M. Zeng, Z. Wu, C. Tian, L. Zhang, and L. Hu. Efficient person re-identification by hybrid spatiogram and covariance descriptor. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 48–56, June 2015.

[28] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.