A Hierarchical Frame-by-Frame Association Method Based on Graph Matching for Multi-object Tracking

Sourav Garg[1], Ehtesham Hassan[1], Swagat Kumar[1](✉), and Prithwijit Guha[2]

[1] Tata Consultancy Services, New Delhi, India
{sourav.garg,ehtesham.hassan,swagat.kumar}@tcs.com
[2] Indian Institute of Technology Guwahati, Guwahati, Assam, India
pguha@iitg.ernet.in

Abstract. Multiple object tracking is a challenging problem because of issues like background clutter, camera motion, partial or full occlusions, change in object pose and appearance etc. Most of the existing algorithms use local and/or global association based optimization between the detections and trackers to find correct object IDs. We propose a hierarchical frame-by-frame association method that exploits a spatial layout consistency and inter-object relationship to resolve object identities across frames. The spatial layout consistency based association is used as the first hierarchical step to identify easy targets. This is done by finding a MRF-MAP solution for a probabilistic graphical model using a minimum spanning tree over the object locations and finding an exact inference in polynomial time using belief propagation. For difficult targets, which can not be resolved in the first step, a relative motion model is used to predict the state of occlusion for each target. This along with the information about immediate neighbors of the target in the group is used to resolve the identities of the objects which are occluded either by other objects or by the background. The unassociated difficult targets are finally resolved according to the state of the object along with template matching based on SURF correspondences. Experimentations on benchmark datasets have shown the superiority of our proposal compared to a greedy approach and is found to be competitive compared to state-of-the-art methods. The proposed concept of association is generic in nature and can be easily employed with other multi-object tracking algorithms.

1 Introduction

Multiple object tracking is often challenged by the phenomena of cluttered background, unpredictable and non-smooth trajectories, occlusions and variations in appearance on account of illumination, deformations and articulations. The last two decades have seen the exploration of contour, point and region based methods [1] for handling such challenges. However, the recent years have witnessed the tracking-learning-detection [2] based algorithms given the advancement made

G. Bebis et al. (Eds.): ISVC 2015, Part I, LNCS 9474, pp. 138–150, 2015.
DOI: 10.1007/978-3-319-27857-5_13

in the arena of object detection and on-line learning algorithms [3]. In these approaches, a detector is used to locate objects in each frame and then associate them across frames to generate long trajectories. These methods can be broadly classified into batch processing and on-line methods.

Batch processing methods are primarily off-line methods that require detections over all the frames. These detection responses are linked together to form short trajectories called, tracklets which may be fragmented due to occlusions. These tracklets are then globally associated to generate longer trajectories [4–7]. In some cases, the long term trajectories are built directly from individual detection responses as in [8–10]. In either case, the global association is crucial in these methods and a number of approaches [4,11,12] have been proposed in the literature to achieve this. The computational requirement for these methods are huge as they require iterative associations for generating globally optimized trajectories. It is therefore difficult to apply these methods to real-time applications and leads to latency when used with a sliding time window [5].

On the other hand, on-line methods [13–16] can be used for real-time applications as the trajectories are build sequentially by resolving associations on a frame-by-frame basis using only current and past information. While these methods are comparatively simpler, they suffer from frequent ID switches and drifting under occlusions.

In this work, our main focus is to improve tracking performance under occlusions. A complete categorization of such occlusions in to 14 different static and/or dynamic forms can be found in [17]. We aim to improve the performance of on-line methods by using a hierarchical framework that uses spatial layout consistency and an inter-object relationship model to resolve associations among the detected objects between two consecutive frames. As a first step, easy to resolve targets are associated using spatial layout consistency. This is done by finding a MRF-MAP solution for a probabilistic graphical model using a minimum spanning tree over the object locations and finding an exact inference using belief propagation. The targets which remain unassociated in the first stage are processed through the second stage where a Kalman Filter based relative motion model is used to predict the occlusion state for each target. This along with the information of immediate neighbors is used to resolve the identities of objects under static or dynamic occlusion.

The main contribution made in this paper are as follows. First, a new hierarchical framework is proposed for resolving frame-by-frame associations among the detection responses that is non-greedy, fast and non-iterative. Second, in contrast to existing graphical methods [18,19] which use dense representations of objects, we use a sparse representation where the target locations are connected through a minimum spanning tree and the association is resolved through MRF-MAP belief propagation. Third, Unlike methods [20] that use relative motion model for tracking objects, we use predict the occlusion state using the same. This along with the information about immediate neighbors is used to detect IDs of objects under static or dynamic occlusion.

The rest of paper is organized as follows. A brief review of related work is presented in Sect. 2. The proposed approach is described in Sect. 3. The experimental results are presented in Sect. 4. Finally, we conclude the work in Sect. 5 and sketch the future extensions.

2 Related Work

In the related works, significant amount of research contributions exist in the field of multi-object tracking which have attempted different issues and challenges. Nevertheless, in this section we focus on some of the recent and significant works which are related to the problem of association resolution in multiple object tracking which is the major focus of this paper.

In the recent works, use of probabilistic graphical models for tracking has shown some interesting results. In [4,19] novel formulations of CRF has been applied for tracking in multi-object scenario. In [4], authors proposed online learning of CRF models for solving the multi-target tracking which uses tracklet pairs as node and correlation among pairs for edge potential computation. Subsequently in [5], the authors have extended the concept in multiple instance learning framework for learning a non-linear motion map. In [19], authors proposed single CRF model for joint tracking and segmentation of multiple objects. They solve a multi-label problem where node potentials are computed with detection responses and likelihood of super-pixels belonging to a target, and edge potentials are computed as overlap between super-pixels in spatial and temporal domain. The approach is similar to Poiesi et al. [18] which uses a particle filter based tracker and applies a MRF for resolving the IDs between multiple objects where each particle carrying an ID acts as a node in MRF model.

All of these methods use dense object representations of an object such as multiple particles/super-pixels, while we use sparse detections as the nodes with edges defined over their spatial connectivity in our MRF based graphical model.

The association based approaches defined using a probability distribution as an optimization problem are solved using energy minimization as in [4–6,19] or a MAP solution as in [16,18]. Depending on the type of underlying graph, kinds of constraints involved in the formulation and the number of variables, these optimization routines differ from each other. We make use of Belief Propagation as a MAP inferencing technique to find an optimal assignment for our probability distribution. The underlying graph in our case being a Minimum Spanning Tree allows for an exact inference in two iterative steps.

The use of motion model for predicting the trajectories due to lack of a proper assignment is done in different ways. [4,5] use CRF model for learning the motion model, [18] uses particle filter and means shift algorithm, [6,19] use motion cues in the energy minimization formulation. We use Kalman Filter for individual motion and relative motion between tracker pairs. The relative motion model also present in [5] is based on long term trajectories by defining head-close and tail-close tracklet pairs for discriminating them among the other possible pairs, whereas, we define it based on Kalman Filter relative motion learning for predicting occlusions and determining occluder-occludee pairs.

We define states for the tracked object for it being completely visible or partially or fully occluded by either other objects or background. A similar definition for occlusion scenarios is also mentioned in [19] where occlusion is defined using visibility and depth of the objects, cues from which become penalties in the energy minimization problem. On the other hand, we define these states based on the relative motion model as a result of which we explicitly handle the occlusion recovery from the occluder-occludee pairs.

3 Proposed Approach

In order to explain our approach, we will make use of the following notations. A given video sequence is represented by the symbol $I_k,\ k = 1, 2, \ldots, N$ indicating that the video has a total of N frames. Each detected object is represented by a bounding box (BB) and is labeled with a global ID. Each detected object for a given frame I_k is represented by the symbol $D_i^k,\ i = 1, 2, \ldots, n$, where n is the number of objects detected in the frame. Each detector D_i^k is represented by a rectangle $\{c_i, w_i, h_i\}$, where c_i is the two-dimensional image coordinate of the center of the rectangle. w_i and h_i are the width and the height of the rectangle respectively. The corresponding tracker windows available in the frame is represented by the symbol $T_j^k,\ j = 1, 2, \ldots, m$. The trackers are the bounding boxes obtained from those in the previous frame using a local tracker based on SURF matching, color histograms or simply, motion predictors.

The objective of our work is to grow trajectories of detected objects by resolving frame-by-frame associations in an on-line fashion. The various components of the proposed scheme is shown in Fig. 1. It has three modules: Detection, Association and Tracking. The detection module uses object detection algorithm to find bounding boxes for all objects available in each query frame. These detections are associated with the detections from previous frame in the association module. The association module makes use of a two stage hierarchical approach to resolve these associations. The first stage makes use of spatial layout consistency to resolve identities of easier targets. The second stage makes use of a Kalman filter-based relative motion model to deal with difficult cases of occlusion. This is the main contribution of our approach. Finally, there is the histogram and SURF matching based tracking module to track some objects locally which cannot be identified otherwise especially the ones which reappear after occlusion. These modules are explained next in this section.

The Detection Module – An object detector is run on every frame to find out the detection responses. These responses act as an observation for the graphical model to find out the associations between trackers and detections. New trackers are initialized for the unassociated responses that occur consecutively for more than three times.

The Association Module – This module resolves the association among the detections between two consecutive frames. This is done by using a two step process. In the first step, easier targets are identified by using spatial layout

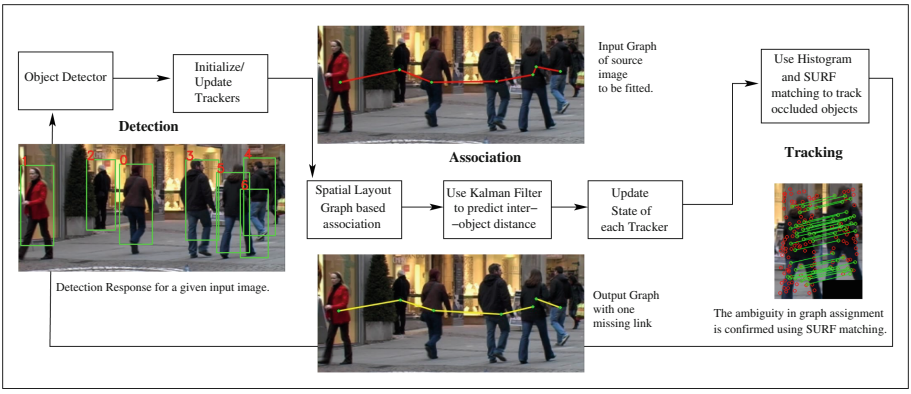


Fig. 1. Block diagram depicting various components of our approach. It can be broadly categorized into Detection, Association and Tracking modules.

consistency. The difficult cases like occlusion and missed detections are dealt in the second stage by using a relative motion model and is described next.

Spatial Layout Consistency – can be used to identify wrong or missed detections by exploiting the fact that the arrangement of objects may not change immediately between consecutive frames. The arrangement of objects is captured in the form of a graph which is formed by fitting a minimum spanning tree (MST) over the centers of tracker windows.

Considering the nodes as the random variable X_i over the bounding box locations c_i, the edges of the tree is represented by constructing an adjacency matrix A_{ij}. We define a joint probability distribution over the graph (MST) G with each of its node represented by X_i and edge by A_{ij}

$$P(G) = 1/Z \prod_{i,j} \psi(\mathbf{X}_i, \mathbf{X}_j) A_{ij} \tag{1}$$

where Z is the normalizing factor. The function $\psi(X_i, X_j)$ is the potential function for an edge between nodes X_i and X_j and is defined as

$$\psi(X_i, X_j) = \exp\left(-w_1 \Delta\theta_{ij} - w_2 \Delta\mathbf{r}_{ij}\right) \phi_1 \phi_2 \Delta s_1 \Delta s_2 \tag{2}$$

where $\Delta\theta_{ij}$ is the change of angle between the edges and $\Delta\mathbf{r}_{ij}$ is the change in edge length between the source and destination graph. ϕ_1 and ϕ_2 are the measures of appearance matching coefficient (Bhattacharya Coefficient in this case) for respective nodes as compared to the source nodes. Δs_1 and Δs_2 are the measures of change in scale of the nodes compared to the source nodes.

Source Graph – The underlying source graph (minimum spanning tree) as shown in Fig. 2(a) is the spatial layout with random variables defined for each node that represents a tracked object. The edges are specified by the adjacency matrix. The use of tree is intuitively justified as we do not really want to model

the edges between the pairs of targets which are far off from each other. Also in real life scenarios it is highly likely that moving objects like humans or vehicles interact and change their motion according to other similar moving objects in their immediate neighborhood. This in turn leads to simplification of the optimization problem that we solve using Belief Propagation in two iterative steps as a MAP inferencing technique for the probabilistic graphical model.

Observations – The objects detected by the detector in the next frame I_{k+1} are the observations which will be used for resolving associations through graph matching. These observations are shown as green rectangles in Fig. 2(b). The yellow rectangles are the dummy observations which are obtained from the tracker locations in the previous frame. The dummy observations help in reducing the false assignments and facilitate one-to-one matching. These dummy variables are deliberately given weights which are 4 times less than the actual observations, so that there is high chance of assigning a true observation than dummy than a false one to the random variable in a respective order. Hence, there is no scope for many-to-one associations. Also, one-to-many associations do not occur because the optimization formulation searches for only the maximum value assignment. Thus, the two-sided mutex constraints are implicitly taken care of.

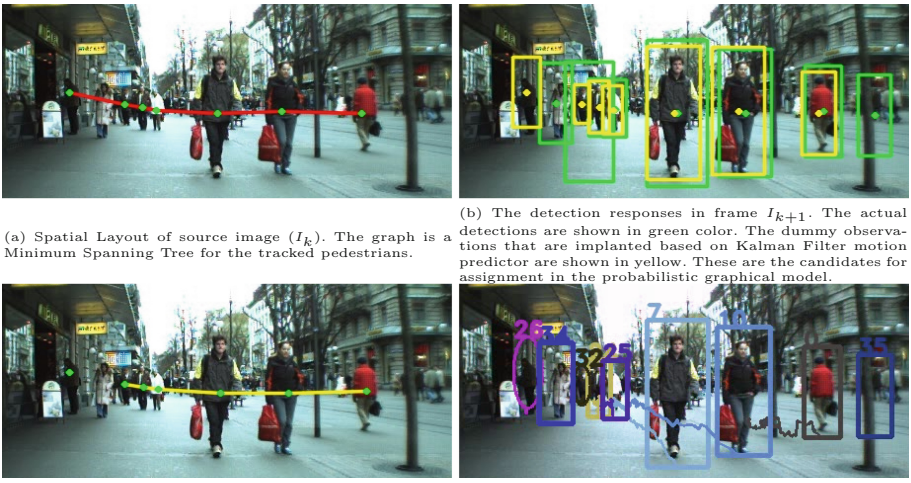
Resultant Graph – Now a MRF-MAP solution is sought for the graph matching taking the source graph as the prior and the current observations as the observation likelihood. The resulting graph is shown in Fig. 2(c). This helps in getting rid of the rightmost detection window which wrongly detected a tree as a human. Secondly, several missed detections were recovered such as objects with IDs 25 and 32. These missed detections are shown using ellipses in Fig. 2(c) as the tracker location is obtained only from prediction. The right most detected window is assigned a new object ID 35.

Relative Motion Model and Occlusion State of the Target – As stated earlier, the more difficult cases of static and/or dynamic occlusions are resolved by using relative motion model and neighborhood information of each object in a group. We use relative motion model to predict if an object is going to be occluded in the immediate future. This is unlike [20] where the relative motion model is used for tracking objects. We define a Kalman Filter K_{ij} for each tracker pair (T_i, T_j) using the inter-object distance d_{ij} as the state variable. The relative inter-object motion is approximated using a constant velocity state model given by the following equation

$$\begin{bmatrix} d_{ij}^{k+1} \\ \dot{d}_{ij}^{k+1} \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} d_{ij}^{k} \\ \dot{d}_{ij}^{k} \end{bmatrix} + \mathbf{w}_k \tag{3}$$

where, $\dot{d}_{ij}^{k}$ is the corresponding velocity and $\mathbf{w}_k$ is the process noise with covariance matrix given by

$$\mathbf{Q}_k = \begin{bmatrix} \Delta t^2 & \Delta t \\ \Delta t & 1 \end{bmatrix} \tag{4}$$



(a) Spatial Layout of source image (I_k). The graph is a Minimum Spanning Tree for the tracked pedestrians.

(b) The detection responses in frame I_{k+1}. The actual detections are shown in green color. The dummy observations that are implanted based on Kalman Filter motion predictor are shown in yellow. These are the candidates for assignment in the probabilistic graphical model.

(c) Resultant Spatial Layout is the output of MRF-MAP based selected edges. The left-most edge is not selected as detection response is not available for that.

(d) The confirmed pedestrians are shown as a rectangle and the predicted ones are shown as an ellipse.

Fig. 2. Resolving IDs for detection responses through graph matching (Color figure onliine).

And, the measurement model is given by

$$\left[d_{ij}^{k}{}^{m}\right] = \left[1\ 0\right] \begin{bmatrix} d_{ij}^{k} \\ \dot{d}_{ij}^{k} \end{bmatrix} + \mathbf{v}_k \tag{5}$$

where, $\mathbf{v}_k$ is the measurement noise with covariance matrix

$$\mathbf{R}_k = \left[\sigma_{\Delta d_{ij}}^2\right] \tag{6}$$

The predicted state $d_{ij}^{k+l},\ l = 1, 2, 3, \ldots$ is used to determine the state of the tracking object as explained next. A target can be any of the following states:

$$S_i = \begin{cases} 1, \text{if target is available and is confirmed} \\ 2, \text{if target is occluded by another target} \\ 3, \text{if target is occluded by the background} \\ 4, \text{if target cannot be confirmed by any means} \end{cases} \tag{7}$$

Any target whose ID is resolved and confirmed by spatial layout consistency (the first step) is said to be in state $S_i = 1$. If the target matches with the dummy observation then it can have any of the remaining three states. It belongs to inter-object occlusion if the following condition is satisfied.

$$S_i = 2 \iff \exists\ l\ \ \in \{1, 2, 3\} : d_{ij}^{k+l} < \epsilon \tag{8}$$

where, d_{ij}^{k+l} represents the predicted values of the inter-object distance d_{ij}^{k}. The target belongs to the state of object-background occlusion if the following condition is satisfied.

$$S_i = 3 \iff \phi_i < \delta \tag{9}$$

where ϕ_i is the appearance matching value computed through Bhattachharya Coefficient for color distributions. The target belongs to the fourth state if it is not confirmed whether it is present or not, that is, there is no sufficient evidence to prove its belongingness to any of the other three states.

Merge and Split – The dynamic occlusions of objects are usually dealt by considering the concepts of merge and split [21]. The merges are implicitly dealt with in our approach due to the spatial layout consistency as explained in Fig. 3. The two objects that are about to merge are constrained to be apart such that it is better for the missing target to match with its dummy observation than with its occluder. Hence, ID switches can be easily avoided during a merge.

On the other hand, splits are handled with the knowledge of state of the target. Any occluded target, when reappears, is searched in the neighborhood of the occluder. If the appearance of the target is similar to the new detection response in the neighborhood of occluder then it is confirmed to be the occludee. This way we are able to avoid unnecessary fragments for targets which are occluded even for longer durations. This is explained in the Fig. 4. In this figure, the trajectories of objects 1 and 3 merge together and get occluded as shown in Figs. 4(a)–(c). The pedestrian 1 is correctly identified by using neighborhood relationship information as shown in Fig. 4(e).

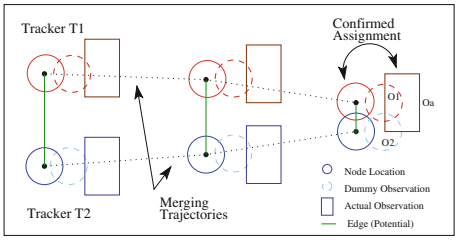
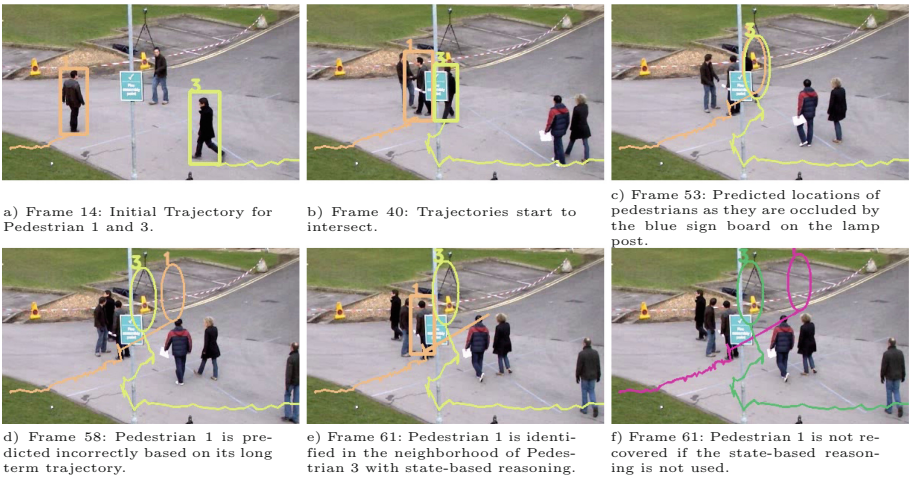


Fig. 3. The availability of a dummy observation facilitates correct assignment of detection response during merging of trajectories.

Tracking – Association of targets is sometimes limited by the availability of detection responses and the detector's accuracy and precision. This makes it necessary to localize some of the targets by means of appearance based local region tracking. We use SURF correspondences for localizing targets that are in state $S_i = 4$.

4 Experiments and Results

The performance of the proposed algorithm is tested on three datasets, namely, ETH [22], TUD [12] and PETS 2009 [23] datasets. The ETH dataset comprises of two sequences, that is, Bahnhoff and Sunny Day, both having a dynamic camera movement, but a similar view angle. Videos of both TUD and PETS dataset are recorded using a static camera. The TUD dataset has a view from a low-lying



a) Frame 14: Initial Trajectory for Pedestrian 1 and 3.

b) Frame 40: Trajectories start to intersect.

c) Frame 53: Predicted locations of pedestrians as they are occluded by the blue sign board on the lamp post.

d) Frame 58: Pedestrian 1 is predicted incorrectly based on its long term trajectory.

e) Frame 61: Pedestrian 1 is identified in the neighborhood of Pedestrian 3 with state-based reasoning.

f) Frame 61: Pedestrian 1 is not recovered if the state-based reasoning is not used.

Fig. 4. Occlusion recovery based on knowledge of correct states of target. Pedestrian 1 is identified again once it recovers from occlusion by making use of the information it was in the neighborhood of Pedestrian 3 before occlusion.

camera with humans of relatively larger size whereas, the PETS dataset has a camera placed far from the crowd with a slanted top view. Both ETH and TUD datasets have frequent dynamic occlusions while PETS dataset has inconsistent trajectories and small human sizes. These datasets are publicly available. The performance of the proposed algorithm is compared with several state-of-the-art algorithms. The proposed algorithm was implemented on an Intel Core $i7$ CPU with 2.90 GHz clock speed and 8 GB RAM. The average computation time for processing each frame depends on the number of objects in each frame. For instance, the per frame computation time for ETH dataset is around 100 ms as there are high number of objects per frame. For TUD dataset, the per frame computation time is 70 ms. The PETS dataset takes very less time, that is, 38 ms, because of less number of trajectories and smaller number of objects being tracked. The computational complexity of our algorithm is polynomial in number of trajectories in a given frame. Apart from that, few peaks occur for pedestrians of larger size due to SURF matching. There is no off-line processing involved and hence, no latency is noted. The computation time for the detection of humans is not taken into account in any of these cases.

Evaluation Metrics – In order to compare the performance of our algorithm with others, we have used the evaluation metrics as defined in [24–26]. We make use of 12 evaluation parameters for comparing our method with other algorithm as shown in Table 1. The parameters have the usual meaning and are not described here due to space constraints. The average rank of each tracker is computed as described in [26] to indicate the overall performance.

The performance of our algorithm is compared with four other state-of-the-art methods on three datasets, namely, ETH, TUD and PETS 2009 as mentioned

above. Out of these four, two methods [5,6] use global optimization to resolve associations. In [5], authors have used temporal window of 8 frames to make the process on-line but, with a latency of 2 s. It is also mentioned that there is a trade-off between accuracy and latency in their approach. The work reported in [7] is one of the most recent approach in the category of frame-by-frame association. In this work, the authors have reported results with four different settings with each one of them having different evaluation parameters and performing well in different scenarios. We have chosen to compare our results with the one condition that does not use global optimization. This is closest to our approach, though our results are competitive even for other settings. Finally, we have compared our method to a greedy approach [27] where the associations between detectors and trackers are resolved by using the Hungarian algorithm. The affinity matrix is constructed by using histogram matching and overlap between the object windows.

The performance comparison of the above methods with the proposed method is provided in Table 1. In terms of average ranking, our algorithm gives best performance for PETS dataset, second best performance for TUD dataset and worst performance for ETH dataset. The poor performance on ETH dataset is mostly due to lack of a better appearance model than due to camera motion as proven by the high number of trajectory fragments. This means that the objects

Table 1. Comparison of results with state of the art methods on standard datasets. The average rank shows that our approach has the best rank for PETS, second best for TUD and last for ETH.

Dataset	Method	MOTP	MOTA	Recall	Precision	FAF	GT	MT	PT	ML	Frag	IDS	FPS	Avg Rank
PETS	Nevatia et al. 2014 [5]	-	88.38	93.0	95.3	0.268	19	89.5	10.5	0.0	13	0	14	2.8
	Bae et al. 2014 [7]	69.59	83.04	-	-	0.204	23	100.0	0.0	0.0	4	4	5	2.6
	Anton et.al. 2014 [6]	80.2	90.6	-	-	0.074	23	91.3	4.35	4.35	6	11	1	2.9
	Garg et al. 2015 [27]	86.5	90.4	97.46	93.65	0.380	19	94.73	5.27	0.00	27	22	20	2.7
	Proposed approach	84.30	91.24	95.3	95.6	0.234	19	89.5	10.5	0.0	12	2	26	2.0
TUD	Nevatia et al. 2014 [5]	-	84.0	87.0	96.7	0.184	10	70.0	30.0	0.0	1	0	10	1.6
	Anton et.al. 2014 [6]	65.5	71.1	-	-	0.513	9	77.8	22.2	0.0	3	4	1	2.5
	Garg et al. 2015 [27]	80.0	73.4	88.5	85.8	0.90	10	80.00	20.0	0.0	9	5	10	2.4
	Proposed approach	81.7	79.1	80.5	98.7	0.067	10	60.0	40.0	0.0	7	3	14	2.0
ETH	Nevatia et al. 2014 [5]	-	70.6	79.0	90.4	0.637	125	68.0	24.8	7.2	19	11	10	1.8
	Bae et al. 2014 [7]	64.0	72.0	-	-	0.035	126	73.8	23.8	2.4	38	18	2	2.0
	Garg et al. 2015 [27]	72.5	62.0	86.0	78.4	7.65	124	75.0	20.2	4.8	70	31	10	2.4
	Proposed approach	70.1	58.7	77.4	80.6	1.422	124	56.5	35.4	8.1	62	14	10	2.9

which reappear after occlusion is not alloted the same ID as the appearance matching fails to identify the target.

We also calculated results for a greedy association based approach [27] that solves frame to frame associations between detection and tracking results. This is done by using an affinity matrix calculated by histogram matching and overlap coefficients. The matrix is solved for one to one associations using Hungarian algorithm. This can be considered as a bare minimum framework required for resolving frame-to-frame associations. The table shows that the proposed algorithm helps in reducing the ID switches and trajectory fragmentation to a greater extent as compared to the greedy approach. As the concepts are generic in nature, this can be applied in conjunction with other existing methods to improve their performance.

5 Conclusion and Future Work

In this paper, we have made an attempt to improve the performance of frame-by-frame association methods in multiple object tracking by using concepts like spatial layout consistency and neighborhood relationship information of objects in a group. The spatial layout consistency is imposed through MRF-based graph matching scheme which is used to resolve the association of easily detected objects in a frame. The difficult cases of occlusions and missed detections are dealt with by using a relative motion model which is used to predict if the objects are going to be occluded in the near future. The association for such cases are resolved using information of immediate neighbors of each object tracked in a group. Through experiments on standard datasets, it is shown that the proposed approach works better when compared to the greedy method of association. The results are also competitive as compared to the other state of the art methods. Moreover, our approach doesn't have any latency or off-line processing and uses higher level information to take the right decision for an object's state instead of every time relying on the trajectories. In our future work, we would combine our current approach with a better a appearance model which could be used for tracking individual targets.

References

1. Li, X., Hu, W., Shen, C., Zhang, Z., Dick, A., Hengel, A.V.D.: A survey of appearance models in visual object tracking. ACM Trans. Intell. Syst. Technol. **4**, 58 (2013)
2. Smeulders, A., Chu, D., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: an experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. **36**, 1442–1468 (2014)
3. Zhang, X., Yang, Y.H., Han, Z.: Object class detection: a survey. ACM Comput. Surv. **46**, 10 (2013)
4. Yang, B., Nevatia, R.: An online learned CRF model for multi-target tracking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2034–2041. IEEE (2012)

5. Yang, B., Nevatia, R.: Multi-target tracking by online learning a CRF model of appearance and motion patterns. Int. J. Comput. Vis. **107**, 203–217 (2014)
6. Milan, A., Roth, S., Schindler, K.: Continuous energy minimization for multitarget tracking. IEEE Trans. Pattern Anal. Mach. Intell. **36**, 58–72 (2014)
7. Bae, S.H., Yoon, K.J.: Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1218–1225. IEEE (2014)
8. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. IEEE Trans. Pattern Analy. Mach. Intell. **33**, 1806–1819 (2011)
9. Ben Shitrit, H., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 137–144. IEEE (2011)
10. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1201–1208. IEEE (2011)
11. Brendel, W., Amer, M., Todorovic, S.: Multiobject tracking as maximum weight independent set. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1273–1280. IEEE (2011)
12. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1265–1272. IEEE (2011)
13. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multiperson tracking-by-detection from a single, uncalibrated camera. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 1820–1833 (2011)
14. Shu, G., Dehghan, A., Oreifej, O., Hand, E., Shah, M.: Part-based multiple-person tracking with partial occlusion handling. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1815–1821. IEEE (2012)
15. Song, X., Cui, J., Zha, H., Zhao, H.: Vision-based multiple interacting targets tracking via on-line supervised learning. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 642–655. Springer, Heidelberg (2008)
16. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. Int. J. Comput. Vis. **75**, 247–266 (2007)
17. Guha, P., Mukerjee, A., Subramanian, V.K.: Ocs-14: You can get occluded in fourteen ways. In: 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), pp. 1665–1670 (2011)
18. Poiesi, F., Mazzon, R., Cavallaro, A.: Multi-target tracking on confidence maps: an application to people tracking. Comput. Vis. Image Underst. **117**, 1257–1272 (2013)
19. Milan, A., Leal-Taixé, L., Schindler, K., Reid, I.: Joint tracking and segmentation of multiple targets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5397–5406 (2015)
20. Yoon, J.H., Yang, M.H., Lim, J., Yoon, K.J.: Bayesian multi-object tracking using motion context from multiple objects. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 33–40 (2015)
21. Guha, P., Mukerjee, A., Subramanian, V.K.: Formulation, detection and application of occlusion states (oc-7) in the context of multiple object tracking. In: 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS 2011), pp. 191–196. IEEE (2011)

22. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: Robust multiperson tracking from a mobile platform. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 1831–1846 (2009)
23. Pets 2009 (2009). http://www.cvg.reading.ac.uk/PETS2009/
24. Li, Y., Huang, C., Nevatia, R.: Learning to associate: hybridboosted multi-target tracker for crowded scene. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2953–2960. IEEE (2009)
25. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. J. Image Video Process. **2008**, 1 (2008)
26. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. ArXiv e-prints (2015)
27. Garg, S., Rajesh, R., Kumar, S., Guha, P.: An occlusion reasoning scheme for monocular pedestrian tracking in dynamic scenes. In: 12th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS 2015) (2015)