Improving Multi-target Tracking via Social Grouping

Zhen Qin Christian R. Shelton
University of California, Riverside
{zqin001,cshelton}@cs.ucr.edu

Abstract

We address the problem of multi-person data-association-based tracking (DAT) in semi-crowded environments from a single camera. Existing tracklet-association-based methods using purely visual cues (like appearance and motion information) show impressive results but rely on heavy training, a number of tuned parameters, and sophisticated detectors to cope with visual ambiguities within the video and low-level processing errors. In this work, we consider clustering dynamics to mitigate such ambiguities. This leads to a general optimization framework that adds social grouping behavior (SGB) to any basic affinity model. We formulate this as a nonlinear global optimization problem to maximize the consistency of visual and grouping cues for trajectories in both tracklet-tracklet linking space and tracklet-grouping assignment space. We formulate the Lagrange dual and solve it using a two-stage iterative algorithm, employing the Hungarian algorithm and K-means clustering. We build SGB upon a simple affinity model and show very promising performance on two publicly available real-world datasets with different tracklet extraction methods.

1. Introduction

We consider grouping dynamics and show that modeling social grouping behavior (SGB) can improve multi-person data association based tracking (DAT) performance. The general multi-target tracking problem is to provide the trajectories and identities of multiple targets in video sequences. It is an extensively explored topic in computer vision for its importance in applications like automated surveillance, video retrieval system, and human-computer interaction. DAT (also known as the tracklet-linking problem) is an emerging category of multi-target tracking which considers frames over an extended time window. Different affinity models and optimization methods have been proposed to link conservatively and reliably extracted tracklets (short tracks) into longer ones to form the final tracking result. Though not as suitable for time-critical applications, this approach is more useful for other application scenarios since it can more easily handle occlusions and detection failures. However, due to challenges (mainly from low-level processing errors) such as imperfect detection, low resolution, abrupt motion, and illumination and appearance changes, problems such as false response, track fragmentation, and identity switching still arise. Furthermore, the optimum provided by optimization methods like the Hungarian algorithm is usually sensitive to outliers, so a slightly bad visual cue or a small change of parameter may lead to a different solution.

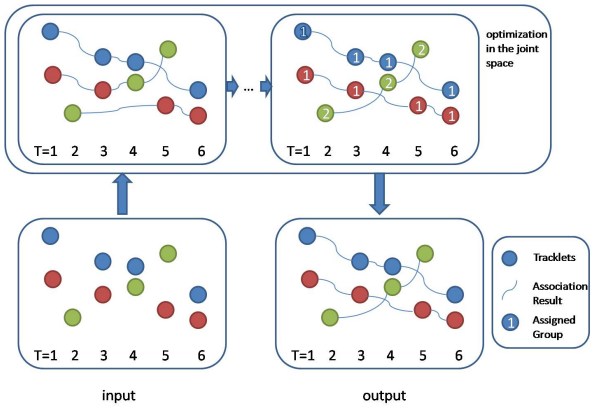


Figure 1: An illustration of our approach: Tracklet colors are ground truth and numbers are our social groupings. We optimize over tracklet-tracklet linkings and tracklet-grouping assignments to exploit social grouping behavior.

In this work, we look beyond purely visual cues and model social grouping behavior to mitigate ambiguities in the tracklet linking problem. When there are few people, multi-person tracking is a relatively easy task since there are few occlusions or interactions, but in semi-crowded scenarios, there are many grouping behaviors we may utilize to disambiguate. Sociology and current computer vision research results show that people tend to walk in groups, when they usually stay close to each other and have similar speeds and trajectories [18]. So in this work, we not only measure the quality of the tracklet-tracklet linkings based on visual cues as in previous work, but we also take the quality of

tracklet grouping into consideration. The main evidence used to link tracklets is still visual cues, but we provide a principled way of using social grouping behavior to regularize the solution to prevent outliers from dominating the tracking result.

We propose a general nonlinear global optimization framework in the joint tracklet-tracklet linking and tracklet-grouping assignment space. A simple illustration is shown in Fig. 1. The contributions and novelty of our work are

- We explicitly consider social grouping behavior to improve data association based tracking. This formulation may be applied to different affinity models, optimization algorithms, and detection methods.
- Our framework is robust to parameters settings and the basic affinity model and automatically selects the number of groups.

2. Related work

Multi-target tracking. For multi-target tracking, time-critical approaches usually use particle filtering algorithms for state estimation [29] [10] [17]. For less time-critical tasks, with the help of state-of-art tracklet extraction methods such as human detector approaches [9] or tracking error approaches [23], researchers look at extended time periods and link tracklets to recover full tracks[4][26]. This is known as DAT or deferred logical inference.

To obtain more reliable affinity scores or linking probabilities between tracklets, [15] learns a discriminative affinity model from a feature pool, boosting of training data from similar scenes. [12] uses a similar boosting framework but focuses on the appearance cues. To effectively infer the best matching given the affinity measurements among tracklets or detection responses, different optimization methods such as Linear Programming [11], Quadratic Boolean Programming [14], the Hungarian algorithm [20][9], K-shortest path [5], MWIS[6], set-cover [25], and approximated dynamic programming[21] have been proposed. In this work, we choose the Hungarian algorithm for simplicity.

[28] and [23] are the most similar to this work. The former looks beyond pairwise tracklet assignment and considers motion dependency that restricts abrupt motion change, but only focuses on the individual track level. The latter builds upon basic affinity model and consider long-term dependency. It finds suspicious tracks with high error rate within themselves after initial matching and does re-sampling and re-matching, but it also focuses on single tracks without using other tracks as social context.

Social behavior in tracking. Using pedestrian trajectories to infer social behavior is not rare, both in the computer vision and sociology community[8][16]. Recently social behavior has caught more attention in tracking community. Typical social factors include a pedestrian's destination, desired speed, and repulsion from other individuals, as well as social grouping behavior. [18] proposed a more effective dynamic model based on such information, [2] uses social structures to improve tracking in crowded scenarios. [19] and [27] infer grouping for better trajectory prediction and behavior prediction respectively. [6] also considers other tracks as contextual constraints for the solution. But to the best of our knowledge, our work is the first to directly and generally consider the quality of social grouping behavior as a higher-level reasoning evidence to improve DAT performance while tracking. Also, we consider grouping from a clustering point of view. Though clustering views have been applied for tracking such as in [24] and [22], they usually focus on flow analysis or feature clustering for people counting. But our framework works for semi-crowded environments and focuses on individuals to improve tracklet linking performance.

3. Modeling social grouping behavior

We build out formulation of social grouping behavior model from the basic tracklet-linking problem. Then we propose a two-stage iterative algorithm that optimizes the joint tracklet-tracklet linking and tracklet-grouping assignments, resulting in steps that can be solved efficiently by off-the-shelf algorithms. Finally we describe how to choose K, the number of clusters used in this algorithm.

3.1. Problem formulation

We aim to recover the trajectories of an unknown number of targets considering consistency of both visual and social cues within a time interval [0,T]. We are given a set of n tracklets (possibly including false alarms) $\boldsymbol{\tau} = \{\tau_1, \tau_2, \ldots, \tau_n\}$ within [0,T]. Each tracklet is a sequence of short but reliable state estimations (in our case, the position and size of targets) across some time interval. The task is to determine which tracklets correspond to the same target. This can be represented as a correspondence matrix ϕ, such that

$$\phi_{ij} = \begin{cases} 1 & \text{if tracklet } j \text{ immediately follows tracklet } i, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

with the added constraints that $\sum_j \phi_{ij} = 1$ and $\sum_i \phi_{ij} = 1$, indicating each tracklet should only follow and be followed by one other tracklet (except for the first and last tracklets of each track, as which we will address in Sec. 3.3). We let Φ be the set of valid correspondence matrixes.

We define a pairwise transition score matrix M (See Sec. 4.2 for details), in which M_{ij} denotes the negative log-likelihood that tracklet j should be the first instance linked after tracklet i. We call this the basic affinity model. Traditional DAT can be formulated as

$$\arg\min_{\phi \in \Phi} \sum_{ij} \phi_{ij} M_{ij}. \tag{2}$$

This is an assignment problem which can be solved optimally by the Hungarian algorithm in polynomial time.

We want to take the quality of social grouping behavior into consideration to help eliminate visual ambiguities within the video and improve tracking performance. The goal is to maximize the consistency of both visual and social grouping cues. For social grouping evaluation, we assume people form groups of an optimal number of K (see Sec. 3.3 for the selection of K) and, within each group, there is a group mean trajectory G_k. Then we measure how well each tracklet sticks to its group and add this term to the original objective function. Thus, the SGB DAT problem is formulated as

$$\begin{aligned} &\arg\min_{\phi\in\Phi,\psi\in\Psi,G} \quad \sum_{ij}\phi_{ij}M_{ij} + \alpha\sum_{ik}\psi_{ik}D(\tau_i, G_k) \\ &\text{s.t.} \quad \forall i,j,k \quad \phi_{ij}(\psi_{ik}-\psi_{jk}) = 0, \end{aligned} \tag{3}$$

where ψ is a social grouping matrix:

$$\psi_{ik} = \begin{cases} 1 & \text{if tracklet } i \text{ is assigned to group } k, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Again there is an added constraint that $\sum_k \psi_{ik} = 1$ and we let Ψ be the set of valid social grouping matrixes. This constraint can be naturally satisfied by modeling the tracklet-grouping assignment task as a clustering problem.

The constraint in Eq. 3 naturally asserts that if two tracklets are linked (they belong to the same target), they should be assigned to the same group. $D(\tau_i, G_k)$ is the distance measure between tracklet i and group trajectory G_k.To balance the contributions of visual cues and social grouping behavior, α is chosen as the weighting parameter. For each dataset in experiments, we choose α by coarse binary search within *only one time window* and keep it constant for all the others. We call Eq. 3 the primal problem.

3.2. A two-stage iterative algorithm

To solve Eq. 3, a nonlinear optimization problem with equality constraints, we apply Lagrange theory and get

$$\begin{aligned} L(\phi,\psi,G,\mu) = &\sum_{ij}\phi_{ij}M_{ij} + \alpha\sum_{ik}\psi_{ik}D(\tau_i,G_k) \\ &+ \sum_{ijk}\mu_{ijk}\phi_{ij}(\psi_{ik}-\psi_{jk}), \end{aligned} \tag{5}$$

in which the μs are the Lagrange multipliers indexed by i,j and k. So the dual of this problem is

$$\begin{aligned} &\max q(\mu) \\ &\text{where} \quad q(\mu) = \min_{\phi\in\Phi,\psi\in\Psi,G} L(\phi,\psi,G,\mu). \end{aligned} \tag{6}$$

The corresponding linking result ϕ (also ψ and G if interested) of the optimization result is the output of the method.

For a fixed μ, let

$$(\phi^\mu,\psi^\mu,G^\mu) = \arg\min_{\phi\in\Phi,\psi\in\Psi,G} L(\phi,\psi,G,\mu). \tag{7}$$

To solve Eq. 6, we use a line search strategy guided by the subgradient:

$$\frac{\partial L(\phi^\mu,\psi^\mu,G^\mu,\mu)}{\partial\mu_{ijk}} = \phi^\mu_{ij}(\psi^\mu_{ik}-\psi^\mu_{jk}). \tag{8}$$

Therefore to find the subgradient, we must solve Eq. 7. We do so by a two-stage block coordinate-descent algorithm, using a penalized Hungarian algorithm and a penalized K-means clustering.

The first stage minimizes over ϕ (the tracklet linking result) from Eq. 5 with ψ and G fixed as

$$\begin{aligned} \phi^\mu &= \arg\min_{\phi\in\Phi}\sum_{ij}\phi_{ij}M_{ij} + \sum_{ijk}\mu_{ijk}\phi_{ij}(\psi_{ik}-\psi_{jk}) \\ &= \arg\min_{\phi\in\Phi}\sum_{ij}\phi_{ij}[M_{ij} + \sum_{k}\mu_{ijk}(\psi_{ik}-\psi_{jk})]. \end{aligned} \tag{9}$$

Comparing with Eq. 2 this amounts to adding a penalty term to the matrix scores, based on the Lagrange multipliers. So Eq. 9 can again be effectively solved by the Hungarian algorithm as is commonly done for Eq. 2.

The second stage minimizes over ψ and G based on Eq. 5, with ϕ fixed as $(\psi^\mu, G^\mu) =$

$$\begin{aligned} &\arg\min_{\psi\in\Psi,G}\alpha\sum_{ik}\psi_{ik}D(\tau_i,G_k) + \sum_{ijk}\mu_{ijk}\phi_{ij}(\psi_{ik}-\psi_{jk}) \\ =&\arg\min_{\psi\in\Psi,G}\sum_{ik}\psi_{ik}[\alpha D(\tau_i,G_k) + \sum_{j}(\mu_{ijk}\phi_{ij}-\mu_{jik}\phi_{ji})]. \end{aligned} \tag{10}$$

where the equality can be achieved by changing the sums' indexes. We note the form of the resulting function is that of a clustering problem like K-means. The first step of K-means clustering is observation assignment step, in which each observation is assigned to its closest mean (center). In this case, we view each tracklet τ_i as an observation while each G_k (group mean trajectory) is a center. The distance function $D(\cdot,\cdot)$ function is the distance measure and the penalty/reward term $\sum_j(\mu_{ijk}\phi_{ij}-\mu_{jik}\phi_{ji})$ adjusts this distance measure. The second step of K-means is the center update step, as described in Sec. 4.4. Since K-means converges to local minimum, we run several random initial grouping assignments and use the output of the one with the minimum value for Eq. 6 as the final result.

For Eq. 9, the intuition is that if two tracklets are not assigned to the same group but are linked, we add the $\sum_k \mu_{ijk}(\psi_{ik}-\psi_{jk})$ term to drive them to different tracks, if the visual cues are not very strong. For Eq. 10, we add the penalty term that considers linking results to drive tracklets

into different groups. Therefore, we are guided explicitly to a better global optimization result. Notice that for the typical initial value of μ equal all zero, the solution for Eq. 9 is the same as the original result for Eq. 2, which is a reasonable starting point.

In summary, for each initial grouping assignment, optimization for Eq. 6 over μ is based on subgradient ascent. For each fixed μ, there is an alternating minimization process based on block coordinate-minimization across the tracklet-tracklet linking space and tracklet-grouping assignment space. The latter involves K-means which has iterations itself in the tracklet-grouping assignment space. Finally, we keep the tracklet-linking result for the run with the minimum value of $q(u)$.

3.3. Selection of K

Until now we have not discussed the selection of K for clustering. However, it is a key issue to guarantee automation. Here we propose a penalized version of our algorithm to select K automatically. The total solution cost of Eq. 3 decreases with K (as the second term becomes smaller with more clusters). Therefore, to prevent overfitting the number of clusters, we place a linear penalty (weight β, fixed for all our experiments) on the number of clusters, K. A summary of our method is shown in Alg. 1. Note that parallelizing our algorithm is trivial, since each iteration of the outer loop is independent. In experiments, building upon the basic affinity model, our unoptimized implementation takes 1 to 10 seconds to converge to a local maximum for each run on a video window of 5 seconds.

Algorithm 1: SGB Algorithm

Data: Tracklet set $\boldsymbol{\tau}$
Result: Linking Result ϕ_{Final}

for $K \leftarrow 1$ **to** K_m **do**
  for $i \leftarrow 1$ **to** N **do**
    $\mu \leftarrow 0$, $\phi^{K,i} \leftarrow 0$
    initialize $\psi^{K,i}$ and $G^{K,i}$ randomly
    while *Not local maximum for Eq. 6* **do**
      $\mu \leftarrow$ subgradient ascent: Eqs. 7 and 8
      while $\phi^{K,i}$ *or* $\psi^{K,i}$ *changes* **do**
        Update $\phi^{K,i}$: Eq. 9
        while $\psi^{K,i}$ *changes* **do**
          Update $\psi^{K,i}$: Eq. 10
          Update $G^{K,i}$ according to $\psi^{K,i}$
        end
      end
    end
  end
  $Cost^{K,i} \leftarrow$ primal cost ($\phi^{K,i}$, $\psi^{K,i}$, $G^{K,i}$): Eq. 3
end
$(K^*, i^*) \leftarrow \arg\min_{K,i} Cost^{K,i} + \beta K$
$\phi_{Final} \leftarrow \phi^{K^*,i^*}$

4. Implementation

We explain some implementation details of our system.

4.1. Tracklet extraction

Our framework can use different affinity models and detection methods. For our experiments, we use two very different frameworks. The first one is the error-based tracklet extraction method of [23]. This method uses traditional detection methods without assuming a human detector. This method benefits from easy implementation and detection of multiple class objects, but suffers from inaccurate detection.

We build our second tracklet extraction framework based on a popular human detector[7], combining nearest neighbor association and template matching to extract conservative tracklets. Given detection responses, we link detection response pairs only at consecutive frames which have very similar color, size and position. Additionally, the newly added detection must be similar to the first detection in the tracklet, thus avoiding within-tracklet ID switches caused by gradual changes. We find this simple strategy produces almost zero ID switches within tracklets and good recall performance.

4.2. Basic affinity model

Social grouping behavior regularizes our solution and alleviates the need for a highly tuned affinity model. However, the basic affinity model must produce *reasonable* measurements, M_{ij}. Here we build a simple affinity model considering three features which are commonly used and shown to be among the most important features[15][13]:

$$M_{ij} = f_t(\tau_i, \tau_j) + \gamma_1 f_{appr}(\tau_i, \tau_j) + \gamma_2 f_{motion}(\tau_i, \tau_j). \tag{11}$$

The time constraint is

$$f_t(\tau_i, \tau_j) = \begin{cases} 0 & \text{if } 0 < \Delta t_{ij} < t_{max}, \\ \infty & \text{otherwise,} \end{cases} \tag{12}$$

where Δt_{ij} is the time gap between the end of tracklet i and the start of tracklet j. So tracklet linking is only possible when tracklet j takes place later than tracklet i and within the maximum allowed frame gap t_{max}. For the appearance model $f_{appr}(\tau_i, \tau_j)$, we use the Bhattacharyya distance between the average color histograms within the tracklets[23]. We employ the HSV color space and get a 24-element feature vector after concatenating 8 bins for each channel. The motion model $f_{motion}(\tau_i, \tau_j)$ measures the motion smoothness, which is defined in both forward and backward direc-

tion by a negative log-Gaussian:

$$f_{motion}(\tau_i, \tau_j) = -\ln G(p_i^{tail} + v_i^{tail}\Delta t_{ij} - p_j^{head}, \Sigma_j) - \ln G(p_j^{head} - v_j^{head}\Delta t_{ij} - p_i^{tail}, \Sigma_i). \quad (13)$$

p_i and v_i represent the refined positions and velocities of τ_i for both the beginning and ending part. A smaller value for all these three models between two tracklets indicates a bigger likelihood that they should be linked. We select the γs (feature weights) by *only looking at one time window*. We show in Sec. 5 that this simple model can provide reasonable linking results.

4.3. Augmentation of pairwise assignment problem

As mentioned before, $\phi \in \Phi$ enforces $\sum_i \phi_{ij} = 1$ and $\sum_j \phi_{ij} = 1$, but they should not hold for the first and last tracklet of each track respectively. In other words, traditional pairwise assignment algorithms like the Hungarian algorithm are not able to identify initialization or termination of tracks when applied to the M matrix directly. A simple example is where one person exits the scene and another person with a very different appearance enters afterwards. These two tracklets will be linked in the solution even though M_{ij} is big because every tracklet must be linked to a "next" tracklet unless conflict exists. One solution is the link-cut strategy, in which a threshold is set and when M_{ij} exceeds it, the link between these two tracklets is cut. However, the final result (after cutting) for this strategy is based on the initial result that may include very bad linking, as it was forced to link every tracklet to a next tracklet. In this work, we use the cut-while-linking strategy similar to [9] in a simple way and augment the M_{ij} matrix as

$$M^{new} = \left(\begin{array}{c|c} M & C \\ \hline B & B \end{array}\right). \quad (14)$$

C is a diagonal matrix (infinity elsewhere) with values c indicating the "finishing" threshold, and B is a matrix of infinities. The constant c can be varied if scene structure (such as exit positions) is known. The augmented columns (and rows) are virtual finishing tracklets. Linking to them indicates the end of a track. Once termination is determined, initialization of tracks come along naturally.

4.4. Augmentation of K-means clustering

In this section we describe some details of the implementation of our K-means clustering, specifically for lines 9–12 in Alg. 1. We view each tracklet and grouping mean trajectory as a $2D$ time series. $D(\tau_i, G_k)$ is the Euclidean distance integrated over time between one tracklet and the grouping mean trajectory. Using positions on image plane works, but for more accurate distance measurement, in this work, positions on the image plane are projected to the $2D$ world coordinates of the ground plane.

The grouping mean trajectories (centers) must exist for the entire time, since for the observation assignment step, each tracklet needs to be compared with all groups. So given a set of tracklets for which $\psi_{ik} = 1$, we generate the kth grouping mean trajectory as follows: For each frame number in the time window, we use the mean position of corresponding tracklets positions occurring at certain time. For time steps with no assigned tracklets, we use linear interpolation or extrapolation for the mean trajectories.

4.5. Duality gap

A duality gap exists between Eq. 3 and Eq. 6, meaning the constraints in Eq. 3 may not all be satisfied when our algorithm converges. Though in practice it does not evidently affect the performance, we propose the following strategy to address such concerns: When each run converges, we still output the tracking result ϕ, but for the evaluation of the objective cost, we force the constraints of Eq. 3 to be true by generating the centers of the final tracks one last time and evaluating the objective cost again. This helps to get a more accurate cost for comparison of different optimization runs.

5. Experiments

We evaluate how modeling social grouping behavior helps to improve multi-person tracking on two public datasets: CAVIAR and TownCentre. We use the popular evaluation metrics of [15]: the number of ground truth trajectories (GT), mostly tracked trajectories (MT), mostly lost trajectories (ML), fragments (Frag) and ID switches (IDS).

Our way of modeling social grouping behavior is independent of the choice of tracklet extraction methods and affinity model. However, it is clear that the tracking performance is related to detector performance and we think it unfair to compare methods with different tracklet extraction frameworks. So, for the CAVIAR dataset, we compare our SGB algorithm with the basic affinity model, a particle filter[10], and the published result of [23] using the same set of tracklets provided by the authors. For the TownCentre dataset, we show improvement of our SGB over the basic affinity model using our proposed detection-based tracklet extraction framework with the same set of detections.

5.1. CAVIAR dataset

The CAVIAR Test Case Scenarios dataset[1] captures people moving in a shopping center with frequent occlusions and interactions. We use the videos selected by [23], which are the relatively challenging parts of the dataset. A comparison is shown in Tbl. 1. Our basic affinity model achieves reasonable results and better results than [23] can be achieved by employing our social grouping model.

Fig. 2 shows representative cases of the strong grouping information that allows us to improve tracking performance. Fig. 2(a) and 2(b) show results under challenging conditions

Frame 890 Frame 950 Frame 980 Frame 1010

(a) Even under heavy occlusions and interactions, the usual identity switch of 2 and 10 is avoided.

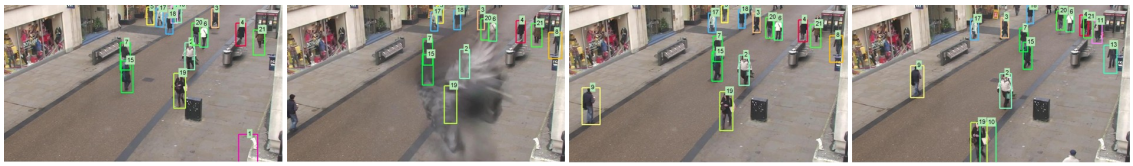
Frame 510 Frame 1245 Frame 1650 Frame 1682

(b) Long-term tracking of the couple (20,21) is possible under challenging conditions: small target, illumination change (frame 510), and false detection (frame 1245)

Frame 1127 Frame 1147 Frame 1200 Frame 1320

(c) The baseline model labels 12 and 6 as the same person. Our model identifies a new track.

Figure 2: Some representative tracking results for CAVIAR dataset.



Frame 2105 Frame 2133 Frame 2147 Frame 2199

Robust tracking result in crowded scene under occlusion.

Figure 3: A representative tracking result for TownCentre dataset.

(noisy velocity estimation, interaction, appearance change and false detection). Fig. 2(c) shows robustness to the selection of C in Eq. 14. Though in the baseline, the last tracklet of person 12 and first tracklet of person 6 are linked since the corresponding M value is less than C, our model finds it inappropriate to do so since person 6 deviates seriously from the mean trajectory of person 12 and 13.

5.2. TownCentre dataset

We use the TownCentre dataset[3] to further test our proposed method, with the detection-based tracklet extraction framework. This video is a high-resolution video of a busy street, with an average of 16 people visible at each frame.

Table 1: Comparison of the tracking result on the CAVIAR dataset: 75 ground truth (GT) tracks.

Method	MT	ML	Frag	IDS
Basic particle filter	53.3%	10.7%	15	19
Basic affinity model	77.3%	6.7%	9	12
MCMC prediction	84.0%	4.0%	6	8
Our SGB model	88.0%	2.6%	5	6

We use the raw HOG pedestrian detections for the first 4500 frames (3 minutes) provided by the authors. The scene is very crowded with 220 people (provided by the annotated ground truth and excluding those who only appear briefly at

Table 2: Comparison of the tracking result on the TownCentre dataset: 220 ground truth (GT) tracks.

Method	MT	ML	Frag	IDS
Basic affinity model	76.8%	7.7%	37	60
Our SGB model	83.2%	5.9%	28	39

the image border) in just 3 minutes, but there are many social grouping behaviors. A quantitative comparison of our SGB model and basic affinity model is shown in Tbl. 2. We see that our SGB model improves tracking performance in all aspects, especially for ID switches (reducing the number by 35%). Some sample frames are shown in Fig. 3, in which robust tracking is achieved under a crowded scenario and unexpected full occlusion.

6. Conclusion

In this paper, we propose a principled way of incorporating social grouping behavior information into multi-target tracking as a high-level reasoning tool for better performance. Our model is independent of the detection method and affinity model, mitigating the need for heavy training or sophisticated detection in DAT. Our optimization results in a simple form which can be achieved by off-the-shelf algorithms. In experiments we show our method performs better than recent work, using a simple affinity model.

Acknowledgements We thank the UC Riverside Video Computing Group for providing data and useful discussion.

References

[1] Caviar dataset. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. 5

[2] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*, 2008. 2

[3] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, 2011. 6

[4] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *CVPR*, 2006. 2

[5] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Trans. PAMI*, 2011. 2

[6] W. Brendel, M. Amer, and S. Todorovic. Multiobject tracking as maximum weight independent set. In *CVPR*, 2011. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 4

[8] W. Ge, R. Collins, and C. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Trans. PAMI*, 2011. 2

[9] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008. 2, 5

[10] C. Hue, J.-P. Le Cadre, and P. Perez. Sequential Monte Carlo methods for multiple target tracking and data fusion. *Signal Processing, IEEE Trans.*, 2002. 2, 5

[11] H. Jiang, S. Fels, and J. J. Little. A linear programming approach for multiple object tracking. In *CVPR*, 2007. 2

[12] C.-H. Kuo, C. Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *CVPR*, 2010. 2

[13] C.-H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? In *CVPR*, 2011. 4

[14] B. Leibe, K. Schindler, and L. V. Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, 2007. 2

[15] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *CVPR*, 2009. 2, 4, 5

[16] M. Moussaid, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behavior of pedestrian social groups and its impact on crowd dynamics. *PLoS ONE*, 5, 2010. 2

[17] K. Okuma, A. Taleghani, N. d. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, 2004. 2

[18] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 1, 2

[19] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*, 2010. 2

[20] A. G. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In *CVPR*, 2006. 2

[21] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011. 2

[22] V. Rabaud and S. Belongie. Counting crowded moving objects. In *CVPR*, 2006. 2

[23] B. Song, T.-Y. Jeng, E. Staudt, and A. K. Roy-Chowdhury. A stochastic graph evolution framework for robust multi-target tracking. In *ECCV*, 2010. 2, 4, 5

[24] X. Song, X. Shao, H. Zhao, J. Cui, R. Shibasaki, and H. Zha. An online approach: Learning-semantic-scene-by-tracking and tracking-by-learning-semantic-scene. In *CVPR*, 2010. 2

[25] Z. Wu, T. H.Kunz, and M. Betke. Efficient track linking methods for track graphs using network-flow and set-cover techniques. In *CVPR*, 2011. 2

[26] J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *CVPR*, 2009. 2

[27] K. Yamaguchi, A. C. Berg, T. Berg, and L. Ortiz. Who are you with and where are you going? In *CVPR*, 2011. 2

[28] B. Yang, C. Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a CRF model. In *CVPR*, 2011. 2

[29] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38, December 2006. 2