Two-Granularity Tracking: Mediating Trajectory and Detection Graphs for Tracking under Occlusions

Katerina Fragkiadaki[1], Weiyu Zhang[1], Geng Zhang[2] and Jianbo Shi[1]

[1]Department of Computer and Information Science, University of Pennsylvania
{katef,zhweiyu}@seas.upenn.edu,jshi@cis.upenn.edu
[2]Institue of Articial Intelligence and Robotics, Xi'an Jiaotong University
zhangtetsu@gmail.com

Abstract. We propose a tracking framework that mediates grouping cues from two levels of tracking granularities, detection tracklets and point trajectories, for segmenting objects in crowded scenes. Detection tracklets capture objects when they are mostly visible. They may be sparse in time, may miss partially occluded or deformed objects, or contain false positives. Point trajectories are dense in space and time. Their affinities integrate long range motion and 3D disparity information, useful for segmentation. Affinities may leak though across similarly moving objects, since they lack model knowledge. We establish one trajectory and one detection tracklet graph, encoding grouping affinities in each space and associations across. Two-granularity tracking is cast as simultaneous detection tracklet **cl**assification and **cl**ustering (cl^2) in the joint space of tracklets and trajectories. We solve cl^2 by explicitly mediating contradictory affinities in the two graphs: Detection tracklet classification *modifies* trajectory affinities to reflect object specific dis-associations. Non-accidental grouping alignment between detection tracklets and trajectory clusters boosts or rejects corresponding detection tracklets, changing accordingly their classification. We show our model can track objects through sparse, inaccurate detections and persistent partial occlusions. It adapts to the changing visibility masks of the targets, in contrast to detection based bounding box trackers, by effectively switching between the two granularities according to object occlusions, deformations and background clutter.

1 Introduction

We address the problem of object segmentation and tracking in crowded scenes. We propose a framework for combining top-down, model driven information and bottom-up, grouping driven information for tracking through persistent partial occlusions while maintaining accurate spatial support for the objects.

Frameworks combining bottom-up and top-down information have a long history in segmenting and recognizing static images [1,2,3], leading to popular multiple segmentation approaches [4] and recently to competitive detection and pose estimation results [5]. In the video domain, most previous approaches can be categorized into two orthogonal lines of work namely top-down tracking-by-detection, mostly oblivious to grouping information, and bottom-up video segmentation, oblivious to model knowledge.

Video segmentation approaches segment objects following the Gestalt principle of "common fate", often enhanced by large temporal context of point trajectories [6,7,8,9].

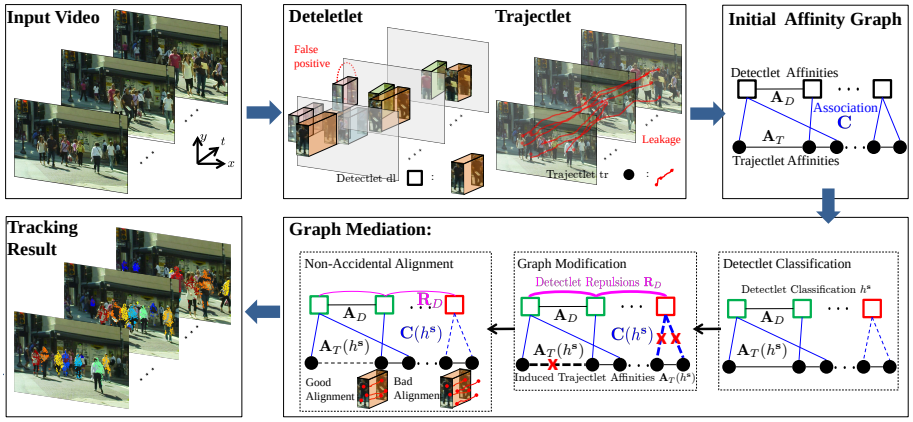

Fig. 1. Two-granularity tracking. We jointly optimize over detectlet classification $h^{\mathbf{s}}$ and clustering in the joint space via multiple model aware segmentations: Selected repelling detectlets induce dis-associations between their corresponding trajectlets. Clustering in the modified graph $\mathbf{A}(h^{\mathbf{s}})$ verifies or rejects detectlet hypotheses depending on their alignment with trajectlet clusters, changing accordingly their classification $h^{\mathbf{s}}$. Here, the green detectlets are accepted while the red one is misaligned and thus rejected.

Such approaches employ dense spatio-temporal representations of trajectories or superpixel clusters that adapt to motion discontinuities across object boundaries. However, they often have difficulty handling deformable or articulated motion or close object interactions, resulting in under-segmentation of similarly moving objects or over-segmentation of articulated objects into rigid parts.

Current state-of-the-art object tracking algorithms [10,11,12] link detections over time. Detection under persistent partial occlusions is challenging since features extracted from a window around an object may be corrupted by surrounding occluders. A box representation cannot adapt to the changing visibility mask of a partial occluded object. As a result, detection responses come as loose-fit / under-fit boxes around a target, or as hallucinated detections spanning over or in the gap of two objects. Apart from occlusions, object deformation poses additional challenges, resulting in a difficult trade-off between precision and recall for deformable object detection.

Our main insight is that the granularity of the tracking representation needs to vary opportunistically between a whole object tracker during full visibility of the template, and fine-grained point trackers during partial occlusions. We propose a tracking framework that exploits cues in two levels of tracking granularity: 1) tracking-by-detection and 2) dense point trajectories. We establish one *detectlet* (*detect*ion track*let*) and one *trajectlet* (*traject*ory track*let*) graph, and encode information of our tracking units as affinities and repulsions (incompatibilities) in each space, and associations across the two (Sections 3.1 - 3.3). Two-granularity tracking is cast as the simultaneous detectlet **cl**assification and **cl**ustering (cl^2) in the joint space of trajectlets and selected detectlets.

We introduce a graph mediation process that solves cl^2 by integrating complementary information in detectlets and trajectlets via multiple model aware segmentations (Section 4). Classification and grouping are tightly coupled as shown also in Fig. 1: Detectlets classified as true positive modify trajectlet affinities by inducing object specific dis-associations. In this way, they resolve affinity contradictions and correct trajectlet affinities leaking across similarly moving objects. We prove clustering in the modified affinity graph is resistant to noisy dis-associations of false detectlets (Section 4.1). Further, non-accidental trajectlet clusters of the *modified affinity graph* provide feedback to detectlet classification: 1) they reject detectlets misaligned with them and 2) they boost isolated detectlets that might have been missed due to low detection score in each frame by linking them across partial occlusions. Our framework is robust against both false or drifting detectlet hypotheses as well as leaking model unaware affinities, in contrast to traditional co-clustering formulations that consider the initial, unmodified affinity graph.

To evaluate the performance of our framework we introduce a new tracking dataset, we call UrbanStreet, captured from a stereo rig mounted on a car driving in the streets of Philadelphia, USA. In contrast to most previous tracking datasets, UrbanStreet provides segmentation masks rather than bounding boxes as pedestrian labels, since often times the targets are partially occluded while navigating in traffic. We compare against a traditional detection based tracker and show our method can segment and track the targets better under persistent occlusions. We further compare against alternative versions of our two-granularity framework to illustrate the contribution of each component in isolation.

2 Related work

Researchers have explored ways of linking sparse detections in time using region information [13,14], body part tracking [15], walking pose cycles [16] or motion smoothness priors [17]. Cluttered environments and persistent partial target occlusions can pose challenges to such linking approaches. Region based tracking can drift to surroundings under accidental appearance similarity. Part tracking may be ineffective due to the large number of false alarms in cluttered scenes. Human dynamics modelling lacks the necessary information to infer the right body pose during persistent occlusions. Smoothness motion priors are not always useful in complex urban environments, where pedestrians change their motion in a complicated fashion (e.g., wait for the green light, stop for a car or bicycle to cross, avoid collisions with surrounding pedestrians, etc.).

Level set trackers [18,19,14,20] have been recently proposed for propagating detection information to no-detection frames and thus tolerate sparse detection responses. Under the assumption that objects do not interlock, authors of [18] represent object shape using level sets and compute pixel wise object posteriors based on depth, motion and shape information. In [14] a level set is initialized by a detection and segments the object tracked in frames with no detections. However, level set optimization can easily get stuck in local minima, being unable to distinguish when a target gets occluded [14]. As a result, additional checks are required to recover from such wrong segmentations and indicate termination of a track. Trajectory units are more powerful than level set

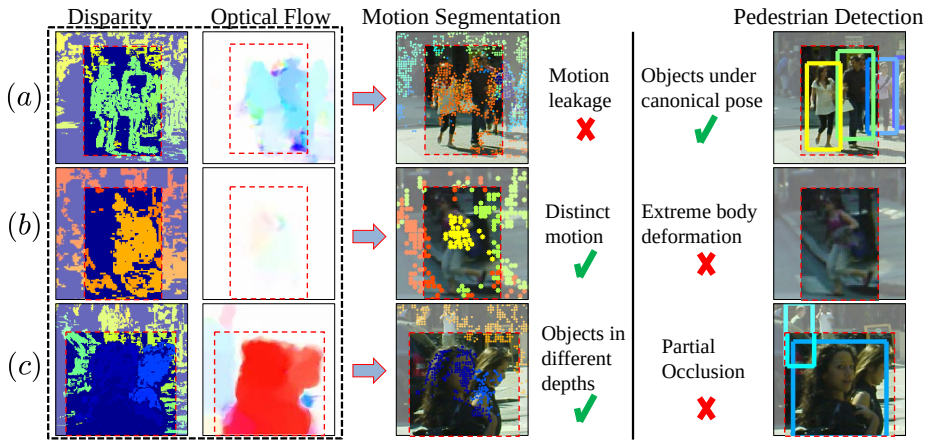

Fig. 2. Complementarity of detectlets and trajectlets. (a) Similarly moving objects. (b) Extreme body deformation. (c) Partial occlusion.

segmentors thanks to their large temporal support: detection and bottom-up separation are not needed in each frame: Single frame separation propagates to entangled frames and correctly segments even interlocked objects assuming they unlock at least in one frame during their time overlap.

3 Tracking Units

Detectlets and trajectlets provide complementary information for tracking in different points in space and time: 1) Detectlets are sparse in time, they often miss objects under severe occlusions or extreme deformations. On the contrary, trajectlets are dense in space and time and extend to frames with no detections. Their long range motion or disparity affinities can separate targets under partial occlusions. 2) The bounding box representation of detectlets is often spatially inaccurate. On the contrary, trajectlets have small spatial support, hence adapt to the changing visibility mask of occluded pedestrians (Fig. 2c). 3) Detectlets can separate objects under canonical pose despite their motion or disparity being similar to surroundings. In contrast, trajectory affinities leak across objects with (persistently) similar motion and disparity (Fig. 2a), or fail to delineate stationary objects from the background scene.

In Sections 3.1 and 3.2 we present our trajectlet and detectlet units and their pairwise affinities $\mathbf{A}_T$, $\mathbf{A}_D$, respectively. In Section 3.3 we present cross-space associations $\mathbf{C}$. These relationships can be summarized in the extended $n \times n$ affinity matrix $\mathbf{A}$:

$$\mathbf{A} \quad = \quad \begin{bmatrix} \mathbf{A}_T & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{A}_D \end{bmatrix} \begin{matrix} \} \ n_T \\ \} \ n_D, \end{matrix} \tag{1}$$

where $n = n_T + n_D$, n_T is the number of trajectlets and n_D the number of detectlets.

3.1 Fine-Grained Trajectlets

We define a trajectlet tr_i to be a sequence of space-time points: $\mathrm{tr}_i = \{(x_i^t, y_i^t), t \in T_i\}$ where T_i is the frame span of tr_i. We obtain trajectlets by tracking pixels across frames following the frame optical flow field [21,22]. Trajectlets are dense in space and can have various lengths depending on the occlusion frequency of the scene part they capture. Further, they are oblivious to any model information.

Trajectlet Affinities $\mathbf{A}_T$: Trajectlets encode rich grouping information in their motion differences. We set affinities $\mathbf{A}_T(\mathrm{tr}_i, \mathrm{tr}_j)$ between trajectlets tr_i and tr_j according to the maximum velocity difference $\mathrm{vel}_{i,j}$ and the maximum disparity difference $\mathrm{dsp}_{i,j}$ computed during their time overlap:

$$\mathbf{A}_T(\mathrm{tr}_i, \mathrm{tr}_j) = \exp\left[-\mathrm{dst}_{i,j}\left(\frac{\mathrm{vel}_{i,j}^2}{\sigma_v^2} + \frac{\mathrm{dsp}_{i,j}^2}{\sigma_d^2}\right)\right], \tag{2}$$

where $\mathrm{dst}_{i,j}$ denotes the maximum Euclidean distance between $\mathrm{tr}_i, \mathrm{tr}_j$. Penalizing maximum velocity and disparity difference takes advantage of the most informative frames in the time overlap between $\mathrm{tr}_i, \mathrm{tr}_j$ [6]. The longer a trajectory the more informative the corresponding affinities [7]. We set affinities between trajectlets that do not overlap in time to zero, indicating lack of information regarding their association.

3.2 Coarse-Grained Detectlets

We define a detectlet dl_p to be a sequence of detector responses $\mathrm{dl}_p = \{(\mathrm{box}_p^t, f_p^t), t \in T_p\}$, where box_p^t is the detection bounding box at frame t, f_p^t is the corresponding detection score and T_p is the frame span of the detectlet. We obtain detectlets by conservatively linking detections between consecutive frames. We define the confidence of detectlet dl_p to be the sum of confidences of its detection responses: $\mathbf{f}_p = \sum_{t \in T_p} f_p^t$.

Detectlet Affinities $\mathbf{A}_D$: We set affinities $\mathbf{A}_D(\mathrm{dl}_p, \mathrm{dl}_q)$ between detectlets dl_p and dl_q that do not overlap in time according to the anchoring score (Fig. 3 *Left*) between their closest in time detections:

$$\mathbf{A}_D(\mathrm{dl}_p, \mathrm{dl}_q) = \mathrm{anchor}(\mathrm{box}_p^{t_1}, \mathrm{box}_q^{t_2}, \mathrm{Tr}(\mathrm{dl}_p, \mathrm{dl}_q)), \tag{3}$$

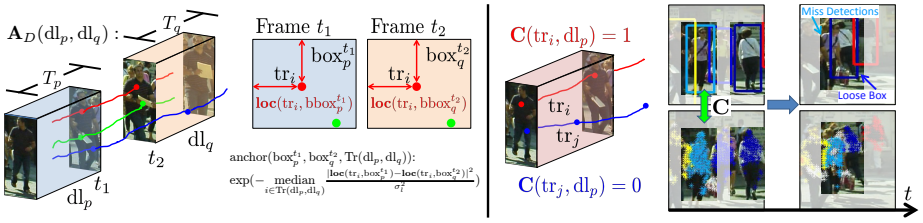

Fig. 3. *Left:* Detectlet affinities $\mathbf{A}_D$. *Right:* Trajectlet to detectlet associations $\mathbf{C}$. Overlaps of close-by detectlet boxes do not confuse associations in $\mathbf{C}$. Trajectlets have the color of their associated detectlet. Associations persist to frames without detections.

where (t_1, t_2) are the indices of frames closest in time, and $\mathrm{Tr}(\mathrm{dl}_p, \mathrm{dl}_q)$ are the set of trajectlets overlapping with both detectlets. The anchoring score of a pair of bounding boxes is high when their common trajectlets have similar relative positions inside the two boxes, as depicted in Fig. 3 *Left*. We set affinities to zero between detectlets whose closest in time detections have no common trajectlets, indicating lack of information regarding their association. We do not encode any motion smoothness priors or color similarity in detectlet affinities. Instead, our mediation framework computes trajectlet clusters that reach further in time than any trajectlet in isolation, and informatively suggest links between isolated detectlets. In contrast, color or motion similarity scores attempt to interpolate blindly information across detection gaps [17].

3.3 Trajectlet to Detectlet Associations C

We set associations $\mathbf{C}(\mathrm{tr}_i, \mathrm{dl}_p)$ between trajectlet tr_i and detectlet dl_p according to spatio-temporal overlap:

$$\mathbf{C}(\mathrm{tr}_i, \mathrm{dl}_p) = 1 \text{ if } \ \forall t \in T_i \cap T_p, \ \ (x_i^t, y_i^t) \in \mathrm{box}_p^t. \tag{4}$$

Computing associations between trajectlets and detectlets rather than between pixels and detections benefits from large time horizon: It saves from erroneous associations between a detectlet and background trajectlets or trajectlets of nearby targets due to accidental per frame overlaps, as shown also in Fig. 3 *Right*.

4 Mediation

We seek to mediate complementary information encoded in the combined affinity matrix $\mathbf{A}$ of Eq. 1. We want to discard false or drifting detectlet hypotheses, link true positive ones through occlusions via fine grain trajectlet clusters, block leaking trajectlet affinities and boost detection recall by trajectlets proposing objects with distinct motion or disparity.

Previous approaches that consider co-clustering (or co-embedding) in a joint space use the combined affinity matrix $\mathbf{A}$ of Eq. 1. For example, clustering in the joint space of detections and image pixels has been considered in [23] for simultaneous detection and segmentation in static images. Co-clustering approaches bypass explicit object hypotheses classification by assigning false alarms to a background cluster. We identify two problems with such standard co-clustering formulations: 1) *False associations.* Assigning a false detectlet to the background cluster needs to cut association edges between the false detectlet and its associated in $\mathbf{C}$ trajectlets (overlapping with it). Such cut cost may be prohibitively large and can confuse the co-embedding solution, as shown in Fig. 4 *Left*. 2) *Affinity contradictions*. In places detectlet and trajectlet graphs disagree, incorrect affinities, namely, trajectlet affinities leaking across the two correctly detected individuals confuse the co-embedding solution, as shown in Fig. 4 *Right*.

In the light of the above observations we propose a mediation framework that optimizes jointly and explicitly over detectlet classification $h^{\mathbf{s}} \in \{0,1\}^{n_D \times 1}$ into true or false positives and detectlets-trajectlet co-clustering, a problem which we call cl^2.

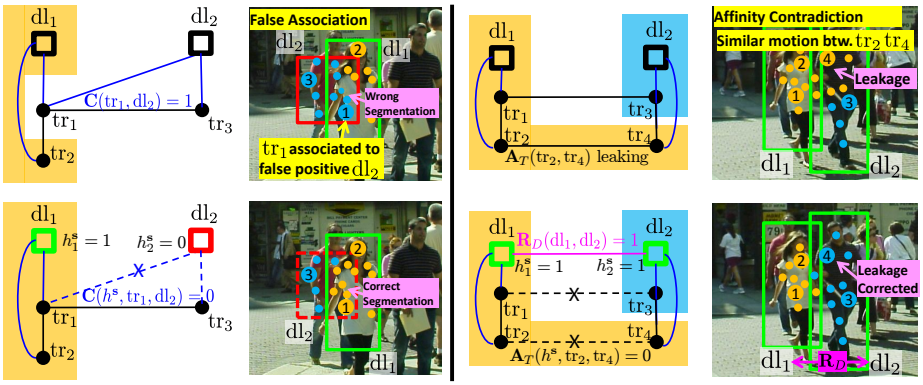

Fig. 4. Model aware affinity graph $\mathbf{A}(h^{\mathbf{s}})$. *Left:* Correcting false associations. Only selected detectlets, here dl_1, can claim trajectories. In this way, we avoid persistently contaminating the spectral partitioning solution with false detectlets (here dl_2). *Right:* Resolving affinity contradictions. We cancel trajectory affinities $(\mathbf{A}_T(\mathrm{tr}_2, \mathrm{tr}_4), \mathbf{A}_T(\mathrm{tr}_1, \mathrm{tr}_3))$ between incompatible detectlets $(\mathrm{dl}_1, \mathrm{dl}_2)$. Spectral partitioning in the modified graph does not leak across similarly moving individuals. Standard co-clustering results are shown in top row.

Instead of working with the initial affinity graph of Eq. 1, we actively modify it according to detectlet classification $h^{\mathbf{s}}$, to reflect corresponding model aware dis-associations between selected incompatible detectlets. This alleviates from the false association or affinity contradiction problems of previous co-clustering frameworks, while correcting possibly leaking, model unaware trajectlet affinities. We present our graph modification in Section 4.1. We further prove that such graph modification can tolerate noisy false alarm detectlets and boost true positive ones, that align better with respect to the underlying grouping links. We present our cl^2 cost function in Section 4.2 and describe our mediation process that solves a relaxed version of it via multiple model aware segmentations in Section 4.3.

4.1 Model Aware Affinity Graph $\mathbf{A}(h^{\mathbf{s}})$

Detectlet to Trajectlet Associations $\mathbf{C}(h^{\mathbf{s}})$: Only selected detectlets dl_p, $h^{\mathbf{s}}(p) = 1$ can claim trajectlets through associations in $\mathbf{C}(h^{\mathbf{s}})$:

$$\mathbf{C}(h^{\mathbf{s}}, \mathrm{tr}_i, \mathrm{dl}_p) = h^{\mathbf{s}}(p) \cdot \mathbf{C}(\mathrm{tr}_i, \mathrm{dl}_p). \tag{5}$$

Induced Trajectlet Affinites $\mathbf{A}_T(h^{\mathbf{s}})$: Active selection of detectlets as true or false positives in $h^{\mathbf{s}}$ changes accordingly the trajectlet graph affinities $\mathbf{A}_T$ by inducing dis-associations between trajectlets associated with incompatible detectlets, as shown in Fig.4 *Right*. Detectlets overlapping in time are defined as incompatible or else repulsive, expressing their inability to span the same object:

$$\mathbf{R}_D(\mathrm{dl}_p, \mathrm{dl}_q) = 1 \text{ if } \; |T_p \cap T_q| > 0. \tag{6}$$

Such incompatibilities are implicit in most previous approaches which never link detectlets overlapping in time. Repulsions $\mathbf{R}_D$ function as *hard constraints* for cross detectlet dis-associations and induce trajectlet repulsions:

$$\mathbf{R}_T(h^{\mathbf{s}}, \mathrm{tr}_i, \mathrm{tr}_j) = \max_{\{p,q|h^{\mathbf{s}}(p)=1, h^{\mathbf{s}}(q)=1\}} \mathbf{C}_{i,p}(1-\mathbf{C}_{i,q})\mathbf{R}_D(\mathrm{dl}_p, \mathrm{dl}_q)\mathbf{C}_{j,q}(1-\mathbf{C}_{i,p}), \tag{7}$$

where we use $C_{i,p}$ to denote $C(\mathrm{tr}_i, \mathrm{dl}_p)$. We induce repulsions only between trajectlets associated exclusively with one or the other of the incompatible detectlets. Induced trajectlet affinities $\mathbf{A}_T(h^{\mathbf{s}})$ take the final form:

$$\mathbf{A}_T(h^{\mathbf{s}}, \mathrm{tr}_i, \mathrm{tr}_j) = (\mathbf{1}_{|\mathcal{T}|}\mathbf{1}_{|\mathcal{T}|}^{\top} - \mathbf{R}_T(h^{\mathbf{s}})) \bullet \mathbf{A}_T(\mathrm{tr}_i, \mathrm{tr}_j), \tag{8}$$

where $\bullet$ denotes point-wise multiplication. We summarize the selection aware relationships in the the following combined affinity graph $\mathbf{A}(h^{\mathbf{s}})$:

$$\mathbf{A}(h^{\mathbf{s}}) \quad = \quad \begin{bmatrix} \mathbf{A}_T(h^{\mathbf{s}}) & \mathbf{C}(h^{\mathbf{s}}) \\ \mathbf{C}(h^{\mathbf{s}})^{\top} & \mathbf{A}_D \end{bmatrix} \begin{matrix} \} \ n_T \\ \} \ n_D. \end{matrix} \tag{9}$$

Graph Modification Analysis During graph modification, affinities between trajectlets associated to incompatible detectlets are canceled. If the corresponding detectlets are true positives, then graph modification results in cancellation of leaking trajectlet affinities across the objects captured by the detectlets. If the corresponding detectlets are false positives, then such modification involves cancellation of affinity links within object interiors, that can potentially cause object over-segmentation.

For simplicity we assume that the grouping cues $\mathbf{A}_T$ are binary. Let ϵ_{miss} be the rate an affinity link is missing between two trajectories belonging to the same object

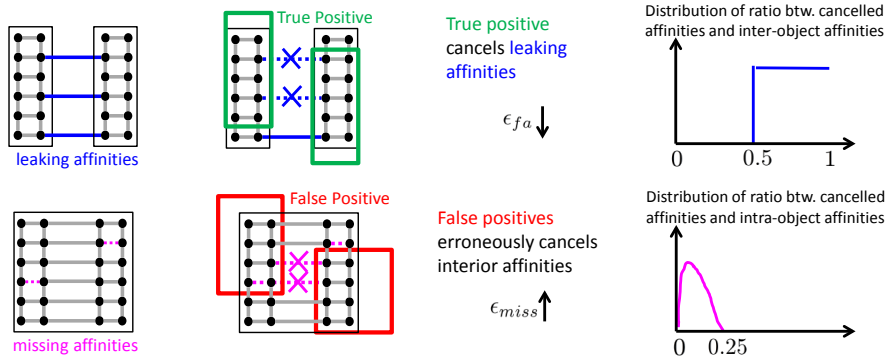

Fig. 5. Resistance of graph modification.True positives (*Top*) and false positives (*Bottom*) modifies the graph $\mathbf{A}(h^{\mathbf{s}})$. We compute the distribution of canceled affinities in the rightmost column by integrating over all detection configurations.

(missing affinity rate). Let ϵ_{fa} be the rate in which an affinity link is present while the corresponding trajectlets belong to distinct objects (leaking affinity rate). Perfect segmentation would require both ϵ_{miss} and ϵ_{fa} to be small.

As shown in Fig. 5, in the case of true positive detectlets, graph modification blocks leaking affinities and essentially reduces error rate ϵ_{fa}. In the case of false positive detectlets, graph modification may dis-associate object interior links and increase ϵ_{miss}. By definition, true positive detectlets overlap at least 50% with the object they capture. Thus, two true positive detectlets tend to align well with respect to the underlying objects and cancel the majority (50% at least) of the inter-object affinities. On the other hand, false alarm detectlets tend to randomly distribute over object interiors. As such two false positives can cancel at most 25% of the intra-object links. Transitivity on the remaining untouched links can still group the object as a whole! Since the first case happens more often than the second, we conclude that our graph modification improves bottom-up segmentation overall.

4.2 Classification-clustering cl^2

We formulate two-granularity tracking as the simultaneous detectlet classification and clustering in the joint (selected) detectlet and trajectlet space. Let $X_k \in \{0,1\}^{n_T \times 1}$, $Y_k \in \{0,1\}^{n_D \times 1}$ denote indicator vectors for trajectlets and detectlet clusters respectively, where $k = 1 \cdots K$, and K being the total number of clusters. We have the following joint optimization over detectlet classification $h^{\mathbf{s}}$ and co-clustering (X, Y):

$$\begin{aligned} \max_{h^{\mathbf{s}}, X, Y, K} \quad & \sum_{k=1}^{K} \mathrm{ncut}(\mathbf{A}(h^{\mathbf{s}}), X_k, Y_k) \cdot \mathrm{confidence}(Y_k) \\ \text{s.t.} \quad & \sum_{k=1}^{K} X_k = \mathbf{1}_{n_T}, \quad \sum_{k=1}^{K} Y_k = h^{\mathbf{s}}, \quad \mathrm{align}(X_k, Y_k) > \mathrm{th}, \end{aligned} \tag{10}$$

where $\mathrm{ncut}(\mathbf{A}(h^{\mathbf{s}}), X_k, Y_k)$ denotes the normalized cut score of cluster (X_k, Y_k):

$$\mathrm{ncut}(\mathbf{A}(h^{\mathbf{s}}), X_k, Y_k) = \frac{[X_k^\top, Y_k^\top]\mathbf{A}(h^{\mathbf{s}})[X_k^\top, Y_k^\top]^\top}{[X_k^\top, Y_k^\top]\mathrm{Diag}(\mathbf{A}(h^{\mathbf{s}})\mathbf{1}_n)[X_k^\top, Y_k^\top]^\top}, \tag{11}$$

$\mathrm{confidence}(Y_k)$ is the total confidence of detectlets in Y_k:

$$\mathrm{confidence}(Y_k) = \mathbf{f}^\top Y_k, \tag{12}$$

and $\mathrm{align}(X_k, Y_k)$ is the alignment score between trajectlets in X_k and trajectlets associated to detectlets in Y_k, measured as the intersection over union:

$$\mathrm{align}(X_k, Y_k) = \frac{|\{i|\ X_k(i) = 1 \wedge \exists j, \mathbf{C}(\mathrm{tr}_i, \mathrm{dl}_j) = 1\}|}{|\{i|\ X_k(i) = 1 \vee \exists j, \mathbf{C}(\mathrm{tr}_i, \mathrm{dl}_j) = 1\}|}. \tag{13}$$

The first term in the summation of Eq. 10 requires the detectlet/trajectlet cluster (X_k, Y_k) to be a salient (stable) group under the normalized cut criterion [24], with

high normalized intra-cluster affinities and associations and low inter-cluster affinities and associations. Notice the dependence of the combined affinity matrix on the detectlet classification $\mathbf{A}(h^{\mathbf{s}})$. The second term biases towards selecting highly confident detectlet hypotheses. The first constraint ensures each trajectlet is assigned to exactly one cluster while the second one allows only detectlets classified as true positive to participate in the clustering.

Importantly, the last constraint in Eq. 10 ensures good alignment between the detectlets in Y_k and trajectlet cluster X_k in each cluster (X_k, Y_k). This constraint acts as a feedback loop from trajectlet clustering to detectlet classification, by boosting or rejecting detectlets that align well (or not) with trajectlet clusters. Alignment is *non-accidental* since it signifies that the two sources of information, or two views [25] of the data, trajectory affinities and detectlets, independently decide the same grouping for the video scene. Our proof in Section 4.1 guarantees the non-accidentalness since it shows that clustering in $\mathbf{A}(h^{\mathbf{s}})$ is not dominated by modifications of selected in $h^{\mathbf{s}}$ detectlets.

4.3 Mediation via Multiple Model Aware Segmentations

We optimize the cost function of Eq. 10 via multiple model aware segmentations: We sample $h^{\mathbf{s}}$ according to detectlet confidence $\mathbf{f}$. Detectlets in $h^{\mathbf{s}}$ induce repulsions $\mathbf{R}_T(h^{\mathbf{s}})$ between trajectlets associated with them. We compute multiple segmentations in the induced affinity graph $\mathbf{A}(h^{\mathbf{s}})$ by varying the number of segments K and the minimum trajectory length L. This results in a pool of detectlet-trajectlet clusters. For each cluster (X_k, Y_k) we measure alignment score $\mathrm{align}(X_k, Y_k)$ and confidence score $\mathrm{confidence}(Y_k)$. We prune clusters whose alignment score is below a threshold $\rho = 0.8$. We obtain the tracking solution by sequentially choosing the best scoring cluster from the remaining ones, that does not overlap with already chosen ones. Detectlets participating in the final tracking solution update accordingly $h^{\mathbf{s}}$ as true positives.

In contrast to multiple segmentation approaches [4,26] that generate a number of bottom-up segmentation proposals to be verified with an object model, we incorporate model information *earlier*, in the segmentation graph $\mathbf{A}(h^{\mathbf{s}})$. This allows recovery from mistakes of model unaware affinities. Furthermore, in contrast to segmentation verification approaches [27] that accept or reject object hypotheses by comparing to the local induced segmentation, we use grouping not only to classify detectlet hypotheses, but to propose detectlets by linking them through large time gaps or partial occlusions.

5 Experiments

To evaluate our tracking framework we introduce UrbanStreet, a pedestrian tracking dataset containing 18 sequences taken from a stereo rig mounted on a car driving in the streets of Philadelphia during rush hours. Part of this dataset was used in [28]. Ground-truth is provided in the form of segmentation masks for all visible targets every four frames (0.6 seconds) in each sequence, with a total of about 2500 pedestrian masks labelled. See also Fig. 6. We further evaluate our framework in TUD crossing dataset [10]. Ground-truth is provided as a set of pedestrian boxes which we link manually into

ground-truth tracklets. We label extra bounding boxes missing from the ground-truth (pedestrians partially occluded more than 50%), resulting in a total of about 1095 labelled boxes. We use the pre-trained pedestrian detector of [29], not specifically tailored to the datasets in hand.

Fig. 6. *Results in UrbanStreet.* Trajectlet clusters adapt to the changing target visibility masks during partial occlusions, in contrast to bounding boxes. We do not track through full occlusions (see red circle): the girl switches cluster ids during its full occlusion. Last row shows ground-truth labelling for few frames in UrbanStreet.

We perform an ablative analysis of our system, by comparing to the following baselines: 1) mediation by co-embedding, which clusters in the unmodified affinity matrix $\mathbf{A}$ of Eq. 1, 2) trajectlet classification, which classifies trajectlets to detectlets according to associations in $\mathbf{C}$, discarding grouping information in $\mathbf{A}_T$ and 3) bottom-up trajectlet clustering in $\mathbf{A}_T$. Each of the above baselines produces a set of clusters which we prune according to alignment scores and let them compete with their detection confidence to populate the tracking result, same as in our full method. We additionally evaluate our input detectlets as well as the detection based tracker of [28] whose results are available for the first seven sequences in UrbanStreet. For these last comparison we fit bounding boxes to our ground-truth segmentation masks.

We measure performance using CLEAR MOT metrics [30]. We compute an one-to-one assignment of hypotheses to ground-truth objects in each frame, measuring intersection over union of segmentation masks in UrbanStreet and of boxes in TUD crossing. We report numbers of miss detections (objects not assigned to any hypothesis), false alarms (segment hypotheses not assigned to any objects), id-switches (number of times a ground-truth track changed his assigned segment identity) and tracking accuracy. We show quantitative results in Table 1 and qualitative in Fig. 6 and 7. Our model does not track across full occlusions (Fig. 6): Trajectlets across full occlusions do not have any time overlap and thus have zero affinities. As such, in both datasets we terminate an object track when the object becomes fully (100%) occluded.

	UrbanStreet				TUD crossing			
	MD(%)	FA(%)	ID-sw.	Acc.(%)	MD(%)	FA (%)	ID-sw.	Acc. (%)
Our Method	50.3	15.6	73	**30.1**	12.3	4.5	0	**82.9**
Co-embedding	57.7	47.0	72	-8.2	14.6	25.1	27	57.8
Trajectory classification	61.0	23.7	71	11.6	18.7	14.5	17	65.2
Bottom-up clustering	78.7	11.5	19	8.5	32.5	8.8	12	57.6
Detectlets	82.6	19.4	49	-4.7	42.6	7.3	81	42.6
Our Method*	44.7	12.0	28	**37.5**				
Gong et al.* [28]	76.5	24.3	36	-6.8				

Table 1. Tracking results. Last two rows in UrbanStreet concern only the first seven sequences for which results of Gong et al. [28] are available.

UrbanStreet contains challenging scenes of complex pedestrian motion with frequent persistent occlusions. Detection based trackers typically interpolate bounding boxes across detection gaps according to motion smoothness. Interpolated boxes, shown dashed in Fig. 6, are often spatially inaccurate. In contrast, two-granularity tracking provides accurate spatial grounding for the targets, switching to fine grain trajectlet trackers during detection gaps, rather than blind interpolation.

The numerous miss detections of bottom-up trajectory clustering are due to stationary pedestrians as well as to pedestrian groups with similar motion. Detectlet sparsity is verified in the large number of detectlet miss detections in both datasets. However, miss detections of detectlets and bottom-up clustering do not coincide, as shown in Fig. 2. As such, our mediation framework outperforms both, being robust to false alarm detectlets and leaking trajectory affinities. Trajectlet classification in Table 1 has large number of miss detections which suggests that isolated trajectories cannot "jump" large time gaps. In contrast, spectral trajectory partitioning propagates information through transitivity further than any single trajectlet. Id switches in trajectory classification are due to drifting point trajectories. Similar drifting trajectlet problem has our framework in the UrbanStreet dataset due to the low frame rate (6.5 fps).

Implementation details When stereo information is available disparity maps are computed with the code of [31]. We evaluate our method against ground-truth bounding boxes by interpolating for each resulting cluster detections to no detection frames using trajectlet anchoring. We evaluate against ground-truth segmentation masks by computing a dense pixel segmentation for each frame using graph cuts on superpixels.

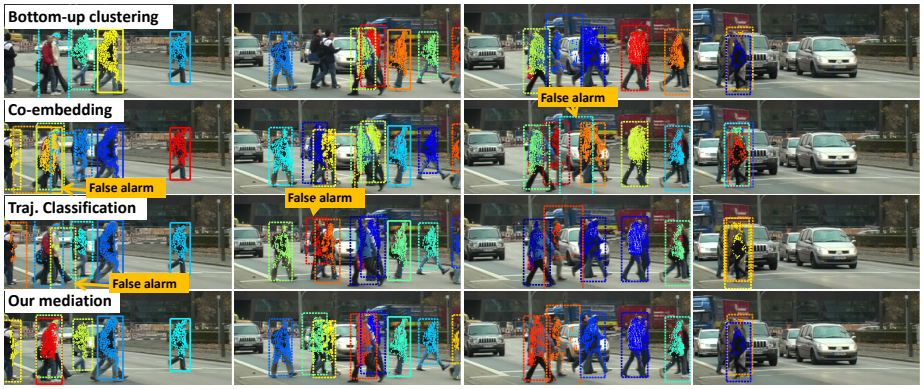

Fig. 7. *Results in TUD crossing.* Our method tracks pedestrians under heavy partial occlusions. Co-embedding and trajectory classification are sensitive to noisy, false alarm or drifting detectlets. Interpolated boxes are shown dashed.

6 Conclusion

We presented a two-granularity tracking framework for segmenting objects in crowded scenes by mediating grouping information of trajectlets and object specific information of detectlets. We cast two-granularity tracking as the simultaneous classification and clustering problem in the joint space of detectlets and trajectlets. Our mediation process optimizes jointly over detectlet classification and co-clustering in the space of selected detectlet and trajectlets and explicitly resolves affinity contradictions via multiple model aware segmentations, alleviating from the problems of standard co-clustering formulations. We believe two-granularity tracking representation can greatly benefit tracking-by-detection approaches, for better handling detection gaps and tolerating detection sparsity while providing a target accurate representation.

Acknowledgement. This work was conducted through collaborative participation in the Robotics Consortium sponsored by the U.S Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement W911NF-10-2-0016 and funcded by grant ONR MURI N000141010934.

References

1. Borenstein, E., Ullman, S.: Combined top-down/bottom-up segmentation. TPAMI (**30**)
2. Levin, A., Weiss, Y.: Learning to combine bottom-up and top-down segmentation. In: ECCV. (2006)
3. Zhang, W., Srinivasan, P., Shi, J.: Discriminative image warping with attribute flow. In: CVPR. (2011)
4. Pantofaru, C., Schmid, C., Hebert, M.: Object recognition by integrating multiple image segmentations. In: ECCV. (2008)
5. Ionescu, C., Li, F., Sminchisescu, C.: Latent structured models for human pose estimation. In: ICCV. (2011)

6. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV. (2010)
7. Fragkiadaki, K., Shi, J.: Exploiting motion and topology for segmenting and tracking under entanglement. In: CVPR. (2011)
8. Lezama, J., Alahari, K., Sivic, J., Laptev, I.: Track to the future: Spatio-temporal video segmentation with long-range motion cues. In: CVPR. (2011)
9. Wang, S., Lu, H., Yang, F., Yang, M.H.: Superpixel tracking. In: ICCV. (2011)
10. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV. (2009)
11. Leibe, B., Cornelis, N., Cornelis, K., Gool, L.V.: Dynamic 3D scene analysis from a moving vehicle. In: CVPR. (2007)
12. William Brendel, M.A.: Multiobject tracking as maximum-weight independent set. In: CVPR. (2011)
13. Ren, X., Malik, J.: Tracking as repeated figure/ground segmentation. In: CVPR. (2007)
14. Mitzel, D., Horbert, E., Ess, A., Leibe, B.: Multi-person tracking with sparse detection and continuous segmentation. In: ECCV. (2010)
15. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. IJCV (2007)
16. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR. (2008)
17. Huang, C., Wu, B., Nevatia, R.: Robust object tracking by hierarchical association of detection responses. In: ECCV. (2008)
18. Bibby, C., Reid, I.: Robust real-time visual tracking using pixel-wise posteriors. In: ECCV. (2008)
19. Bibby, C., Reid, I.: Real-time tracking of multiple occluding objects using level sets. In: CVPR. (2010)
20. Mitzel, D., Horbert, E., Ess, A., Leibe, B.: Level-set person segmentation and tracking with multi-region appearemnce models and top-down shape information. In: ICCV. (2011)
21. Brox, T., Malik, J.: Large displacement optical flow: Descriptor matching in variational motion estimation. TPAMI (2010)
22. Sundaram, N., Brox, T., Keutzer, K.: Dense point trajectories by GPU-accelerated large displacement optical flow. In: ECCV. (2010)
23. Yu, S.X., Gross, R., Shi, J.: Concurrent object recognition and segmentation by graph partitioning. In: NIPS. (2002)
24. Shi, J., Malik, J.: Normalized cuts and image segmentation. TPAMI (2000)
25. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. (COLT' 98)
26. Malisiewicz, T., Efros, A.A.: Improving spatial support for objects via multiple segmentations. In: BMVC. (2007)
27. Ramanan, D.: Using segmentation to verify object hypotheses. In: CVPR. (2007)
28. Gong, H., Simy, J., Likhachev, M., Shi, J.: Multi-hypothesis motion planning for visual object tracking. In: ICCV. (2011)
29. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: ECCV. (2010)
30. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. J. Image Video Process. (2008)
31. Cech, J., Sára, R.: Efficient sampling of disparity space for fast and accurate matching. In: CVPR. (2007)