Detection Free Tracking: Exploiting Motion and Topology for Segmenting and Tracking under Entanglement

Katerina Fragkiadaki
GRASP Laboratory, University of Pennsylvania
3330 Walnut St., Philadelphia, PA-19104
katef@seas.upenn.edu

Jianbo Shi
GRASP Laboratory, University of Pennsylvania
3330 Walnut St., Philadelphia, PA-19104
jshi@seas.upenn.edu

Abstract

We propose a detection-free system for segmenting multiple interacting and deforming people in a video. People detectors often fail under close agent interaction, limiting the performance of detection based tracking methods. Motion information often fails to separate similarly moving agents or to group distinctly moving articulated body parts. We formulate video segmentation as graph partitioning in the trajectory domain. We classify trajectories as foreground or background based on trajectory saliencies, and use foreground trajectories as graph nodes. We incorporate object connectedness constraints into our trajectory weight matrix based on topology of foreground: we set repulsive weights between trajectories that belong to different connected components in any frame of their time intersection. Attractive weights are set between similarly moving trajectories. Information from foreground topology complements motion information and our spatiotemporal segments can be interpreted as connected moving entities rather than just trajectory groups of similar motion. All our cues are computed on trajectories and naturally encode large temporal context, which is crucial for resolving local in time ambiguities. We present results of our approach on challenging datasets outperforming by far the state of the art.

1. Introduction

Our goal is to segment and track closely interacting and deforming agents in a video. Many recent tracking frameworks link and propagate detections over time with an appearance model learnt and updated on the fly. Frequent detections are needed for preventing drifting in tracking. We can group the tracking approaches into the following two categories:

- *No explicit pose representation* ([11], [10], [2]). These trackers use a bounding box for the object being tracked. Detectors that do not model pose explicitly fail under body deformation and articulation. Moreover, the use of bounding boxes has its own drawbacks: 1) It does not provide exact boundary of the object. Consequently, appearance features aggregated from the bounding box interior are often corrupted by background and thus are less discriminative. 2) When agents come close, the corresponding bounding boxes overlap. Then the features extracted leak from one agent to the other leading to drifting.
- *Explicit representation of agent's pose (articulated tracking) ([23]).* As the number of agents increases, there is combinatorial explosion of the number of models to test. Most importantly, the features used (edge map) are often too weak to pick the right model.

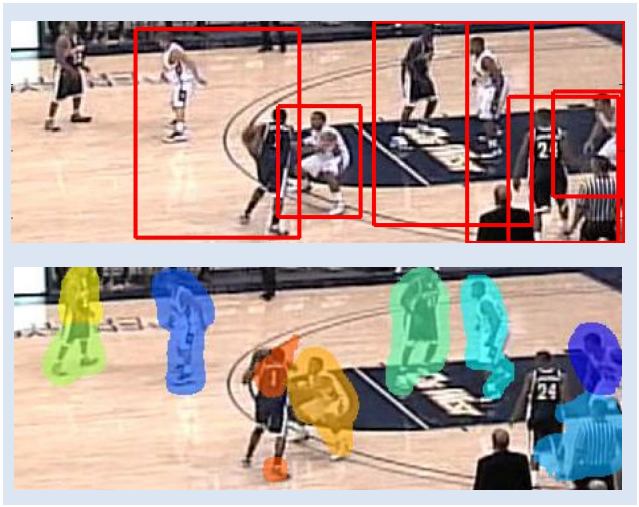

Figure 1. *Top:* Detection results from [7]. Detectors often fail under large body deformation or agent entanglement. *Bottom:* Our approach. Our bottom-up tracklets are semantically meaningful, exhibiting often one to one correspondence with the moving agents without any model guidance.

We propose a system that builds an explicit figure-ground representation of the video. Foreground topology and motion determine repulsive and attractive weights between foreground trajectories. The resulting spatiotemporal segments from partitioning the trajectory weight matrix can serve for later top-down reasoning for tracking or weakly supervised learning.

There is large body of work on spatio-temporal grouping, though dealing mostly with rigid motions or a few isolated moving objects. To tackle challenging real-life scenarios we employ novel cues and representations:

1. We use dense point trajectories as our basic units. Point trajectories aggregate information from large time windows and are more informative than static image cues or per frame optical flow. Under entanglement, static image boundaries are often faint and unreliable. Appearance information leads to over-fragmentation in cases of non uniform object appearance or leakage across object boundaries in cases of accidental across object appearance similarity. On the other hand, per frame optical flow measurements may not be informative for most of the frames. Important motion information for segmenting the video is not evenly distributed across the video frames. Computing motion similarity on trajectories naturally exploits maximum motion information available during their lifespans.

Previous work misses the importance of the imbalance of trajectory lifespans. In articulated motion some body parts deform less than others, for example the head deforms less than the legs. The resulting trajectories thus vary a lot in length. Our main insight is that this imbalance (asymmetry) can have a strong impact on the grouping result: affinities between long trajectories incorporate larger temporal context and are more reliable than those between short ones. If we cluster all trajectories at once, noisy affinities of short trajectpries can confuse the partitioning. Instead, we use a two step clustering, where in the first step we recover the basic scene skeleton by clustering the longest trajectories and in the second step we densify the representation (fill in the details).

2. We compute a novel figure-ground representation based on trajectory saliencies. For each frame, we compute center-surround saliency of the frame optical flow field. Then, pixel saliency values are aggregated along trajectories. Trajectory saliency computation incorporates *long range* rather than local motion contrast. Segmenting figure and ground using trajectory rather than video pixels saliencies provides time consistent figure-ground segmentation result without resorting to appearance models.

3. We untangle the moving agents by imposing object connectedness constraints. Moving agents that are closely interacting sooner or later will separate. We introduce repulsive forces between trajectories belonging to different connected components of the foreground map of any frame, and let these forces propagate in time through transitivity. Topology driven repulsion is an additional grouping cue, byproduct of our figure-ground reasoning.

We tested our framework on *moseg*, a recently released dataset for motion segmentation ([4]). We introduce a new challenging dataset comprised of basketball videos extracted from [19], with large body deformation and frequent occlusions. Our method outperforms by far the state of the art. We provide extensive evaluation of our different system components, quantifying the value of each one in isolation.

2. Related work

The observation that motion provides a strong cue for perceptual organization dates back to the Gestaltists ([22]) that suggested the grouping principle of "common fate". Various approaches have been proposed for exploiting this principle for spatio-temporal grouping. Part of video segmentation methods ([20, 17, 21]) are based on two frame optical flow. They rely on short time horizon and as such they are dependent on choosing the pair of frames with clear motion difference between the objects. If object motion is not consistent during the shot, the results are not expected to be time consistent either. Works of [14] and [24] combine flow estimation with an appearance model to propagate motion information across multiple frames for time consistent segmentation.

In order to take advantage of longer time horizon (multiple frames) many approaches use point trajectories. Multibody factorization methods ([5, 25, 6, 16]) can distinguish the motion of rigid objects relying on properties of an affine camera model. Apart from being restricted to rigid motions, these methods require all trajectories to have same length, which is not feasible under frequent occlusions or body deformation. Works of [6] and [16] have tried to recover from this requirement to a certain extent. However, deformable or articulated motion still poses challenges to the factorization framework. Authors of [26] obtain automatic background subtraction under a projective camera model by estimating the trajectory basis for background trajectories using RANSAC. Works of [4], [8] and [3] use clustering of trajectories and do not require them to be full length. The use of trajectories in contrast to per frame segmentation, provides time consistent clusters since grouping naturally propagates over time and does not depend on motion information extracted from a particular frame.

Apart from the gestaltic principle of common fate, psychophysics experiments of Nothdurft in [13] have shown that local motion contrast drives perception of simple visual concepts (bars) as a group when viewed in background of similar concepts that dier from them in motion or orientation. This suggests motion perception depends not only on motion similarity but also on local motion segregation. Numerous approaches have exploited center-surround motion saliency for automatic background subtraction ([12, 9]) or foreground object segmentation ([15]).

3. Graph setup

We cast video segmentation as graph partitioning in the trajectory domain. We define a trajectory tr^i to be a sequence of space time points:

$$\mathrm{tr}^i = \{(\mathbf{p}_k^i, t_k^i), \quad k = 1 \cdots |\mathrm{tr}^i|\}, \quad i = 1 \cdots |\mathcal{T}|$$

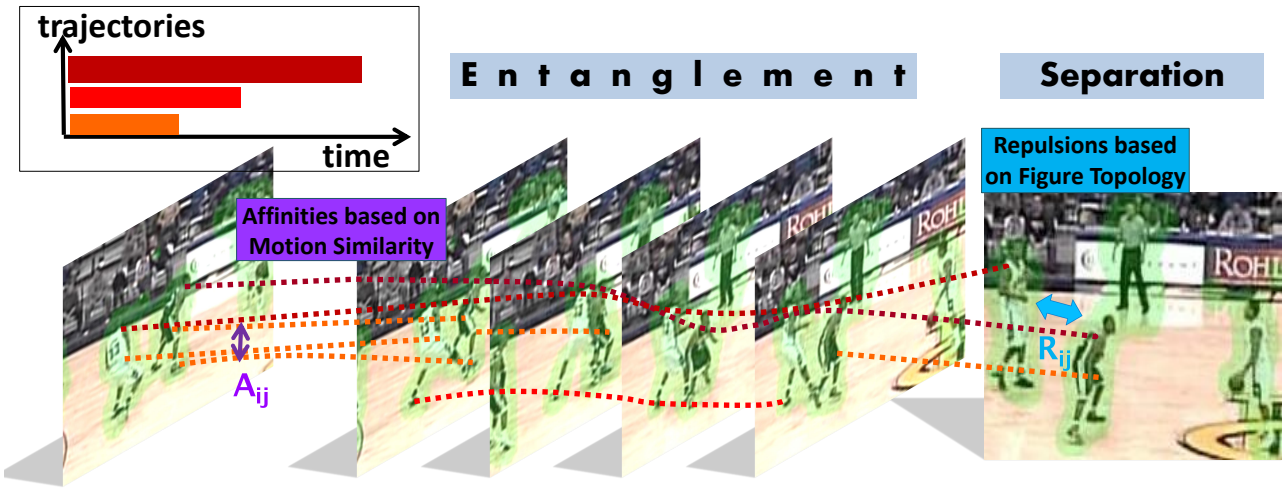

Figure 2. Graph setup. Affinities between long trajectories are more reliable since they reflect larger temporal context. In addition, they propagate further in time thanks to the large spatio-temporal neighborhoods of long trajectories.

where $\mathcal{T}$ is the set of trajectories in a video, $|\mathrm{tr}^i|$ is the length of tr^i, image pixel $\mathbf{p}_k^i = (x_k^i, y_k^i)$ is a vector of spatial coordinates corresponding to the kth point of tr^i and t_k^i is the frame index for the kth point of tr^i. We obtain our trajectories using a recently proposed approach that tracks densely using optical flow ([18]). Using dense rather than sparse corner trajectories results in denser coverage of the tracked objects.

We classify trajectories as foreground or background based on their saliency. For each trajectory, the trajectory saliency is the maximum of the saliency values of its points. Non salient trajectories are assigned to background without further consideration.

Our graph nodes are foreground trajectories. Our graph weights are attractive and repulsive forces between trajectories. Motion similarity (*common fate*) introduces attraction $\mathbf{A}_{ij}$ between trajectories having similar motion. Topology of foreground maps introduces repulsion $\mathbf{R}_{ij}$ between trajectories that belong to different connected components in any frame of their time intersection. We seek a graph partitioning that maximizes within-group attraction and between-group repulsion. We segment our trajectory graph using the normalized cut criterion:

$$\max \ \epsilon = \frac{\text{within-group } \mathbf{A}}{\text{total degree } \mathbf{A}} + \frac{\text{between-group } \mathbf{R}}{\text{total degree } \mathbf{R}}$$

We follow the solution proposed in [27] and calculate the K largest generalized eigenvectors of $(\mathbf{W}_{\text{eq}}, \mathbf{D}_{\text{eq}})$ where $\mathbf{W}_{\text{eq}} = \mathbf{A} - \mathbf{R} + \mathbf{D}_{\mathbf{R}}$ and $\mathbf{D}_{\text{eq}} = \mathbf{D}_{\mathbf{A}} + \mathbf{D}_{\mathbf{R}}$. $\mathbf{D}_{\mathbf{W}}$ stands for the degree matrix of $\mathbf{W}$, a diagonal matrix with $\mathbf{D}_{\mathbf{W}}(i,i) = \sum_j \mathbf{W}(i,j)$. See [27] for the derivation details. We obtain our final segmentation by clustering our trajectories in the embedding space. We use repulsion $\mathbf{R}$ to guide our discretization by discarding clusters with interior repulsion and merging clusters with no interior repulsion.

In section 3.1 we present our figure-ground representation. Sections 3.1.2 and 3.2 present our repulsive and attractive weights respectively. Section 3.3 presents priority clustering for exploiting trajectory asymmetry and section 3.3.1 presents our discretization procedure. We show results of our approach in section 4 and conclude in section 5.

3.1. Figure-ground spatio-temporal segmentation

We define foreground (figure) to be the ensemble of moving agents and background (ground) the static world scene that embraces them. We are not interested in segmenting the static scene into different layers as layer motion segmentation approaches. Rather, we concentrate on untangling the moving agents. Cameras can be moving freely.

Our system computes motion saliencies on trajectories and produces a figure-ground segmentation by assigning non-salient trajectories to background. Figure-ground representation is a central piece of our work:

1. It reflects the different kinematic nature of figure and ground: Trajectories covering a moving agent form a motion cluster or a set of clusters (torso and articulated limbs) depending on whether the agent moves rigidly or articulates. On the other hand, the ground, comprised of different planes and surfaces, usually has a motion that changes smoothly in space. As such, ground trajectories do not form compact motion clusters. Approaches clustering figure and ground trajectories simultaneously, often need to merge background clusters together in post processing. In contrast, our system exploits *motion contrast from large temporal context* encoded in motion saliencies of trajectories to produce a figure-ground segmentation for the video. Then clusters only what is found as foreground.

2. It naturally provides an idea about *object scale*: affinities between trajectories are often chosen proportional to euclidean distance. By clustering only foreground trajectories, different objects naturally pop out since free space between objects causes sudden drop of the affinity values, which becomes more prominent after normalization of the affinity matrix (normalized cut).

3. It provides us with additional cues for segmentation

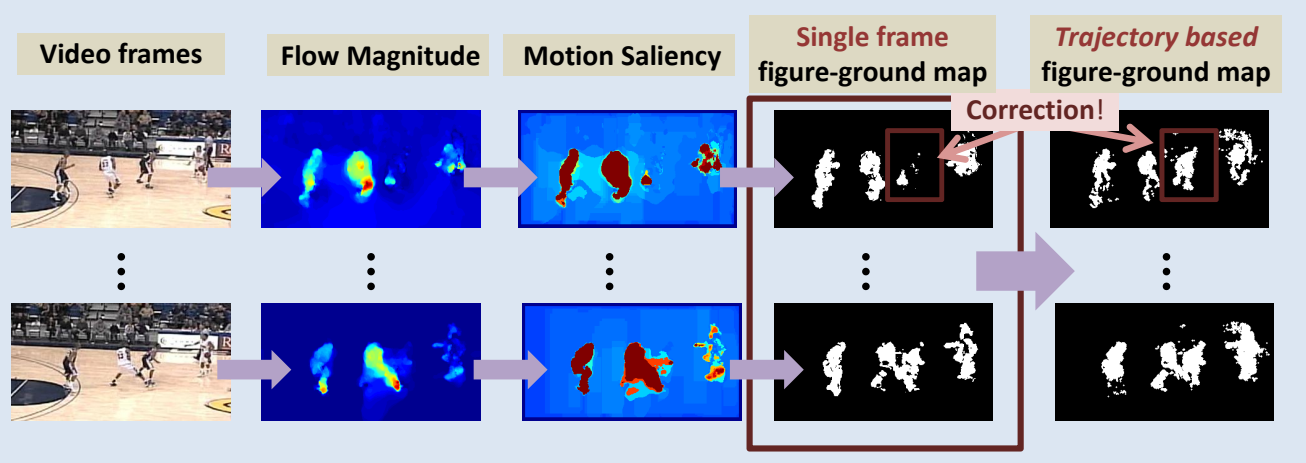

Figure 3. Trajectory based motion saliency. Notice the correction on the player in blue: information flows through trajectory point correspondences from informative frames of large motion difference between figure-ground to frames where agents are stationary, without using an appearance model. Our system makes no assumptions concerning camera motion.

by exploiting connectedness of objects as described in section 3.1.2.

4. It offers computational benefits: it suggests using motion not only as a grouping cue but also as a resource allocation mechanism.

3.1.1 Trajectory motion saliency

Most previous approaches on motion saliency use single frame optical flow measurements as features in a center-surround differencing operation. That is, they exploit per frame motion contrast to find what moves saliently in each video frame. We define *per frame* motion saliency of a video pixel as a function: $\mathrm{sal_p} : \mathcal{P} \rightarrow [0,1]$, where $\mathcal{P}$ is the set of video pixels $(\mathbf{p}, t) = (x, y, t)$.

However, taking advantage of large time horizon is crucial: an agent that appears stationary initially may move later on. By computing a motion saliency map for each frame in isolation, we end up erroneously assigning the partial stationary agent to the background in the initial frames. We circumvent this problem by computing motion saliency of *trajectories* rather than video pixels. In this way, our saliency maps are computed based on large time range motion contast. We define trajectory saliency as:

$$\mathrm{sal_{tr}} : \mathcal{T} \rightarrow [0,1], \quad \mathrm{sal_{tr}}(\mathrm{tr}^i) = \max_{k:1\cdots|\mathrm{tr}^i|} \mathrm{sal}((\mathbf{p}_k^i, t_k^i))$$

In practice, we set the saliency of a trajectory to be the 90th percentile of the saliencies of its points, since the maximum value may be noisy. Then, we define the foreground map f_l of frame l as the set of points of the salient trajectories:

$$\mathrm{f}_l = \{\mathbf{p}^i_{\mathrm{ind}^i(l)} : \ \forall\, i \text{ s. t. } \mathrm{sal_{tr}}(\mathrm{tr}^i) > s_{\min}\}$$

where $s_{\min}$ is a saliency threshold and $\mathrm{ind}^i(l)$ is the function that maps a frame index l to the trajectory point index of tr^i or to 0 if the trajectory is not on for that frame. By computing motion saliency over trajectories rather than pixels, the resulting figure-ground segmentation maps are consistent over time without resorting to appearance models (see also Fig. 3). For computing single frame motion saliency maps we used the publicly available code of [15] for center-surround differencing on optical flow magnitude field.

3.1.2 Topology driven repulsion R

For the resulting spatio-temporal segments to have a semantic interpretation, the grouping cues need to go beyond appearance or motion similarity. These cues alone often over-fragment an object into distinctly moving parts. We believe we can do better by exploiting *foreground topology* provided by our figure-ground representation.

Specifically, we compute the set of connected components $\mathcal{C}^l$ for each foreground map f_l: $\mathcal{C}^l = \{C^l_k, \ k = 1\cdots|\mathcal{C}^l|\}$. f^l_C assigns a foreground connected component to each trajectory point at frame l: $f^l_C : \mathcal{P}^l \rightarrow \mathcal{C}^l$ where $\mathcal{P}^l$ the pixels of frame l. We use the term foreground topology to describe the assignment of trajectories to connected components.

- Foreground topology *cannot* indicate when two trajectories should be grouped together: we cannot know, without additional information, whether a connected component in the foreground map is a single agent or a group of agents.
- Foreground topology *can* indicate when two trajectories *cannot* be grouped together if assigned to different connected components (assuming object connectedness).

Nevertheless, indicating separation is as useful as indicating attraction: an attractive force of value 0 can indicate either no attraction or lack of information. A strong repulsive force indicates dissimilarity and as such is useful in the grouping process ([27]).

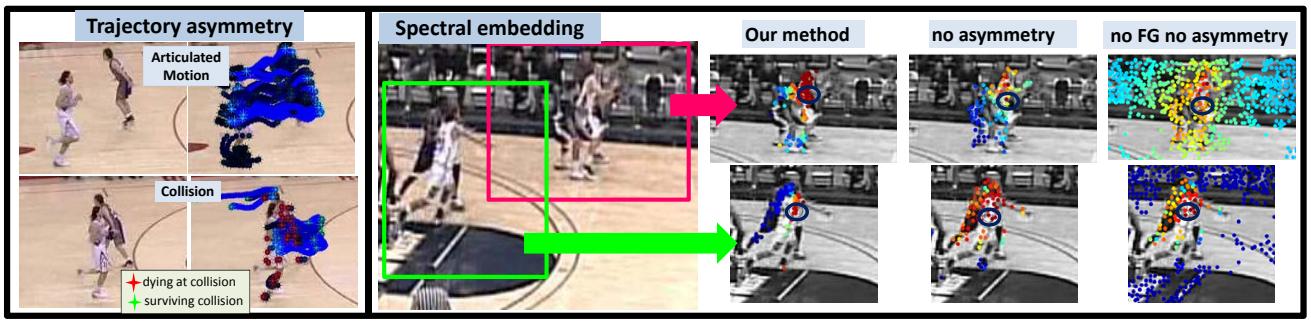

Figure 4. *Left:* Asymmetry of trajectories under body deformation and agent interaction. Long trajectories are shown in light blue and short ones in dark blue. Trajectories often die at occlusions or self-occlusions. This is the case even for occluding trajectories since descriptors get corrupted by sudden background change. *Right:* Spectral embedding. Circles denote the centering points and color denotes the embedding affinity given by the embedding affinity matrix $\hat{\mathbf{W}} = S \cdot \lambda \cdot S'$ (red denotes high and blue low affinity). Notice the discriminability of our embedding (column 1). Using no figure-ground (last column) or partitioning all foreground trajectories at once even with the use of repulsion (column 2) may result in severe leakages.

Moving agents have periods of entanglement and periods of separation during which there is free space between their masses (see fig 2). We define two trajectories tr^i and tr^j as disconnected if *in any frame* of their time overlap they belong to different connected components. We introduce repulsions between disconnected trajectories:

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{if } \exists\, t \text{ s.t. } f_C^t((\mathbf{p},t)^i_{\mathrm{ind}^i(t)}) \neq f_C^t((\mathbf{p},t)^j_{\mathrm{ind}^j(t)}) \\ 0 & \text{otherwise} \end{cases}$$

Large temporal context is crucial for effectiveness of repulsion: long trajectories will propagate the separation to the entangled frames (see also Fig. 2). Short trajectories that do not survive the entanglement period, will not see the separation.

3.2. Motion affinity A

Between each pair of trajectories tr^i and tr^j we compute an affinity score $\mathbf{A}_{ij}$ measuring motion similarity:

$$\mathbf{A}_{ij} = \exp\left(-\frac{D_{ij}}{g}\right), \quad D_{ij} = d \cdot \max_{t \in \mathrm{t}_s^{ij} \cdots \mathrm{t}_f^{ij}} ||\vec{u}_t^i - \vec{u}_t^j||^2 \quad (1)$$

where $\vec{u}_t^i = \mathbf{p}^i_{\mathrm{ind}^i(t+b)} - \mathbf{p}^i_{\mathrm{ind}^i(t)}$ is the velocity of tr^i at frame t, $d = \frac{\sum_{k=\mathrm{t}_s^{ij}}^{\mathrm{t}_f^{ij}} |\mathbf{p}^i_{\mathrm{ind}^i(k)} - \mathbf{p}^j_{\mathrm{ind}^j(k)}|}{(\mathrm{t}_f^{ij} - \mathrm{t}_s^{ij})}$ is the mean euclidean distance between tr^i and tr^j, $\mathrm{ind^i(t)}$ is the function that maps a frame index t to the trajectory point index of tr^i or to 0 if the trajectory is not on for that frame and $\mathrm{t}_s^{ij} \cdots \mathrm{t}_f^{ij}$ is the time overlap between tr^i and tr^j. We used $g = 330$ and $b = 3$. We set $A_{ij} = 0$ for trajectories that their time intersection is less that b frames.

When two objects are static or move similarly we cannot decide on their grouping. It is when they start moving differently with respect to each other that we gain certainty for their separation. The affinity model of equation 1 penalizes the maximum velocity difference between a pair trajectories, similar to [4]. In this way, we squeeze the most information available during the trajectories time intersection.

3.3. Trajectory asymmetry

Previous literature on trajectory clustering for video segmentation mostly focuses on rigid or nearly rigid motions and treats all trajectories equally. Multi-body factorization literature even requires all trajectories in a shot to have the same length. However, by the very nature of a deformable agent, some of the covering trajectories are expected to have longer lifespan than others. The imbalance of trajectory lifespans can have a large impact on the grouping result:

- Affinities between long trajectories reflect large time window. Long trajectories propagate their affinities further in time since they have large number of neighbors thanks to their longevity. They usually cover the rigid part of the torso or the head.
- Affinities between short trajectories reflect short time window and are less reliable due to accidental motion similarity. Accidentalness is way more frequent for short tracks as intuition suggests (see also Fig. 4). They usually cover the limbs since limb motion causes frequent self-occlusions.

Clustering all trajectories at once introduces noisy affinities computed from short trajectories that confuse the embedding since short trajectories often outnumber long ones. Limiting ourselves to long trajectories would create sparse clusters not covering the objects densely. Instead, we follow an intuitive two step priority clustering. We threshold trajectory length at a minimum value $\mathrm{len}_{\min}$ (we use $\mathrm{len}_{\min} = 15$ frames). We denote by $\mathcal{T}^L$ trajectories with length larger than $\mathrm{len}_{\min}$ and by $\mathcal{T}^S$ the rest of the trajectories. We have two steps:

1. In the first step we partition only trajectories in $\mathcal{T}^L$ using spectral clustering as described in section 3. The output is a set of trajectory clusters $\mathcal{K}^k$, $k = 1 \cdots K$ that ideally captures the skeleton of the scene, the basic moving entities. 2. In the second step we embed all trajectories in $\mathcal{T}$ and compute the approximate affinity matrix $\hat{\mathbf{W}} = S \cdot \lambda \cdot S'$ where S are the K largest generalized eigenvectors and λ the corresponding eigenvalues of

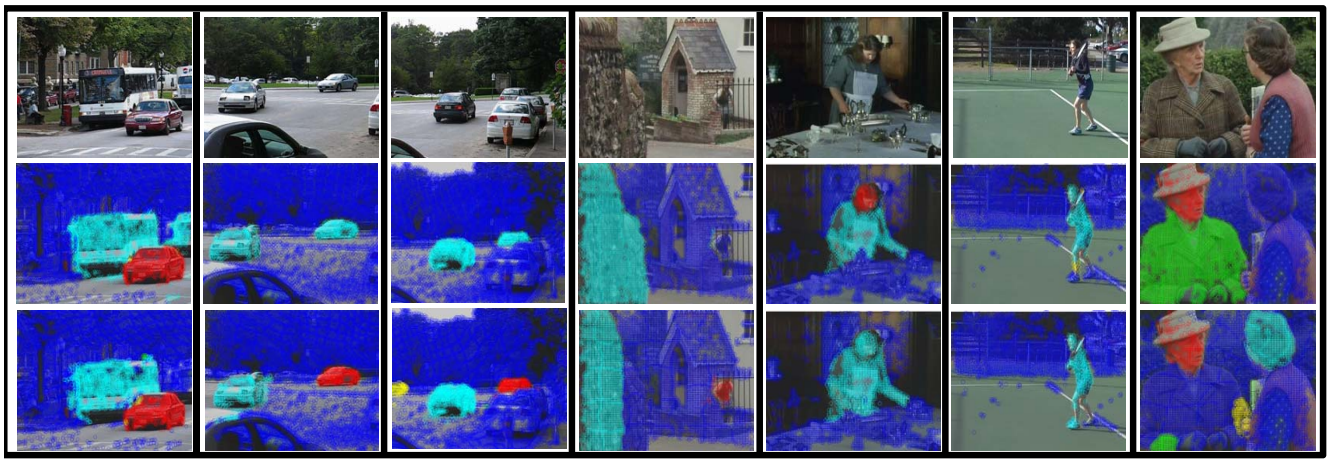

Figure 5. Segmentation in *moseg* dataset. *2d row:* Brox et al [4]. *3rd row:* Our method. Our system imposes object connectedness constraints: we successfully recover the cars in columns 2 and 3 and the bus in column 1 does not leak to the vehicle in the back. Our clusters can be interpreted as entities moving compactly in the scene rather than groups of similar motion: in column 5 the head is grouped with the torso despite the dissimilar motion and in cases of articulation (columns 4 and 6) we recover the whole body of the person. In last column, the body of the ladies was not found salient enough and was mistakenly assigned to background. Image best seen in color.

$(\mathbf{W}_{\text{eq}}, \mathbf{D}_{\text{eq}})$. For each trajectory $\text{tr}^i \in \mathcal{T}^S$ and for each cluster $\mathcal{K}^k$ computed in the previous step, we compute a mean affinity score: $\mathbf{A}_{\text{mean}}(i,k) = \frac{\sum_{j:\text{tr}^j \in \mathcal{K}^k} \hat{\mathbf{W}}(i,j)}{|\mathcal{K}^k|}$. Let $a_i^* = \max_k \mathbf{A}_{\text{mean}}(i,k)$ be the maximum affinity of trajectory tr^i to any of the clusters found in the previous step. If it is larger than twice the second best mean affinity, we assign tr^i to the corresponding cluster. In this way, we densify our output tracklets without loosing robustness.

3.3.1 Discretization

Discretizing using k-means on the embedding coordinates or our graph nodes would treat all trajectories equally since short trajectories would influence the positions of the cluster centers. Along with k-means clustering, we compute additional clusters centered on long trajectories: we consider the embedding affinity matrix $\hat{\mathbf{W}} = S \cdot \lambda \cdot S'$. For each trajectory tr^i we compute a corresponding cluster $\mathcal{N}(i)$: $\mathcal{N}(i) = \{\text{tr}^k \text{ such that } \hat{\mathbf{W}}(i,k) > \text{q}\}$ where q a threshold. We discard clusters that have interior repulsion among their trajectories as they do not respect object connectedness. We further discard clusters that do not obey a fist order motion model (clusters that do not exhibit time continuity). If necessary, we merge clusters whose concatenation forms cluster with space and time continuity and no interior repulsion. Finally, we greedily choose clusters from largest to smallest in size till most of the trajectories are covered. The robustness of the embedding allows a greedy approach like this one to give reasonable results.

4. Experiments

First we test our method in *moseg*, a recently released dataset for video segmentation ([4]). We used the trajectories and the evaluation software delivered with the dataset. We test on the first 50 frames for each sequence (when the sequence had less than 50 frames we used the whole sequence). We show results in the Table 4. Our results are comparable to state of the art albeit using way less trajectories (only what was found as foreground) and simpler discretization, thanks to the discriminability of our embedding. See also Fig. 5.

To test the performance of our system for tracking multiple interacting and deforming agents, we introduce *figment* (**fig**ure untangle**ment**), a challenging dataset of basketball clips. It contains 18 video sequences of 50-80 frames each, part of the basketball dataset of [19]. The groundtruth bounding box based player trajectories provided with the dataset were not sufficient for our purposes. Instead, we labelled player and background masks every 7-8 frames in each video sequence by marking superpixels to obtain segmentation masks for each player and the background respectively. We computed superpixels using the publicly available code of [1].

For evaluation, each trajectory cluster is optimally assigned to one labelled mask based on maximum intersection. Given this assignment, *clustering error* measures for each clip the percentage of wrongly labelled pixels (pixels carrying a label of a cluster not assigned to their mask) and averages across all clips. *Per region clustering error* measures percentage of wrongly labelled pixels per mask and averages across all labelled masks in the dataset. If in a clip most of the trajectories belong to background, a segmentation algorithm assigning everything to one cluster can get low clustering error. In contrast, per region clustering error is a stricter metric, since it balances the different labelled masks. Missing a region entails an error of 100% for that one. *Over-segmentation* measures the number of clusters assigned to each labelled mask on aver-

	density	clustering error	per region clustering error	over-segmentation	extracted objects
our method	3.22%	3.76%	22.06%	1.15	25
Brox et al. [4]	3.32%	3.43%	27.06%	0.4	26

Figure 6. Results on *moseg* dataset. *Extracted objects* counts labelled masks with less that 10% clustering error.

Figure 7. Segmentation and tracking in *figment* dataset. *2d row:* Detection results from [7]. *3rd row:* Segmentation results from Brox et al. [4]. *4rth row:* FG-r̃-asym. (our method without repulsion) *5th row:* Our method. Free space between agents causes sudden drop of normalized attraction affinities between foreground trajectories. Leakage to background or between agents is controlled. In hard cases, motion similarity and free space separation is still insufficient. Repulsion can still separate agents that are close and move similarly: notice the corrections in the boxes in the last row. White denotes the background cluster. Image best seen in color.

age. These metrics are also used for scoring performance in *moseg*. Due to the different nature of *figment* dataset, we introduce additional metrics to measure tracking quality: a frequent phenomenon tracking is for clusters to leak across multiple agents. *Leakage* measures the percentage of leaking clusters, clusters having large intersection over union score (more than $1/2$ of the one with their assigned mask) with more that one labelled masks for at least one frame. *Recall* measures the percentage of recalled pixels of a labelled mask by only the *single* best cluster (best is the one of the assigned to that mask clusters having the highest recall, where we set recall to 0 for the frames a cluster leaks) and averages across all masks. It measures how semantically meaningful the output tracklets are, how well they can track in isolation, without allowing merging. Finally, since *recall* does not have a space and time dimension (high recall may correspond to good space coverage but short in time tracking or the inverse), *tracking time* measures the number of frames the recall per frame for a labelled mask is above 20% and averages across all labelled masks. To calculate *recall* and *tracking time* we dilated trajectories by a radius of 8 pixels. For all metrics we averaged over our video sequences using trim mean : we discared the top and bottom 10% of the metric values and took the mean of the remaining ones.

To quantify the contribution of the different components of our system (figure-ground representation, topology driven repulsion and trajectory asymmetry) we evaluate the following versions of our framework (tilde denotes that component was not used, FG=figure-ground, r=repulsion, asym=asymmetry):
FG-r̃ - $\widetilde{\text{asym}}$: clustering all foreground trajectories at once with attraction $\mathbf{A}$ and discretization using k-means.
FG-r-$\widetilde{\text{asym}}$: clustering all foreground trajectories at once with $\mathbf{A}$ and $\mathbf{R}$ and discretization using k-means.
FG-r̃-asym : 2 stage priority clustering of foreground trajectories with only $\mathbf{A}$ and discretization using both k-means and asymmetric clusters.

The metrics *clustering error* and *recall* should be considered simultaneously, since an algorithm with 0 output does

	density	clustering error	per region clustering error	over-segmentation	recall	leakage	tracking time
our method	5.21%	4.73%	20.32%	1.57	31.07%	16.52%	75.13%
FG-r̃ - $\widetilde{\text{asym}}$	4.43%	11.13%	33.63%	1.29	20.41%	23.57%	50.77%
FG-r-$\widetilde{\text{asym}}$	3.28%	5.12 %	26.24%	2.07	18.89%	21.16%	46.63%
FG-r̃-asym	5.57%	12.91%	31.32%	1.36	26.95%	21.16%	65.79%
Brox et al. [4]	0.57%	20.74%	86.43%	0	0.46 %	81.55%	1.03%

Figure 8. Results on *figment* dataset.

not make any errors but also has 0 recall and tracking time. Our full method outperforms previous work as well as the partial versions of our system, verifying the importance of the various components of our framework. (see also fig 7 and 4).

5. Conclusion

We presented a method for video segmentation based on spectral clustering of trajectories. Assuming connectedness of the objects to be segmented, our system first computes a figure-ground segmentation of the video and then assigns repulsive forces between foreground trajectories that belong to different connected components in any frame. Information from foreground topology combined with motion information produces a robust trajectory embedding and guides the discretization procedure of the eigenvector solution. We showed segmentation results on challenging videos of multiple interacting agents. We introduced a new dataset for tracking under entanglement which we plan to enlarge with scenes from different entanglement scenarios (boxing, football, volleyball, crowded scenes).

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels. In *EPFL Technical Report no. 149300, June 2010*. 2078

[2] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, 2009. 2073

[3] G. J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *CVPR*, 2006. 2074

[4] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*. 2010. 2074, 2077, 2078, 2079, 2080

[5] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. *ICCV*, 1995. 2074

[6] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *CVPR*, 2009. 2074

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32, 2010. 2073, 2079

[8] M. Fradet, P. Robert, and P. Pérez. Clustering point trajectories with various life-spans. In *CVMP*, 2009. 2074

[9] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of vision*, 8, 2008. 2074

[10] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*. 2008. 2073

[11] B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool. Dynamic 3d scene analysis from a moving vehicle. In *CVPR*, 2007. 2073

[12] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *TPAMI*, 32, 2010. 2074

[13] H. C. Nothdurft. Feature analysis and the role of similarity in preattentive vision. *Perception and Psychophysics*, 52, 1992. 2074

[14] M. Pawan Kumar, P. H. Torr, and A. Zisserman. Learning layered motion segmentations of video. *IJCV*, 76, 2008. 2074

[15] E. Rahtu, J. Kannala, M. Salo, and J. Heikkil. Segmenting salient objects from images and videos. In *ECCV*. 2010. 2074, 2076

[16] S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *CVPR*, 2008. 2074

[17] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000. 2074

[18] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*. 2010. 2075

[19] C. Vondrick, D. Ramanan, and D. Patterson. Efficiently scaling up video annotation with crowdsourced marketplaces. In *ECCV*, 2010. 2074, 2078

[20] Y. A. Wang and H. Adelson. Representing moving images with layers. *TIP*, 1994. 2074

[21] Y. Weiss. Smoothness in layers: Motion segmentation using nonparametric mixture estimation. *CVPR*, 1997. 2074

[22] M. Wertheimer. Laws of organization in perceptual forms. *A Sourcebook of Gestalt Psycycholgy (Partial translation)*, 1938. 2074

[23] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *IJCV*, 2007. 2073

[24] J. Xiao and M. Shah. Motion layer extraction in the presence of occlusion using graph cuts. *TPAMI*, 2005. 2074

[25] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *ECCV*, 2006. 2074

[26] O. J. Yaser Sheikh and T. Kanade. Background subtraction for freely moving cameras. In *ICCV*, 2009. 2074

[27] S. X. Yu and J. Shi. Understanding popout through repulsion. In *CVPR*, 2001. 2075, 2076