Spatio-Temporal Moving Object Proposals

Katerina Fragkiadaki
University of California, Berkeley
katef@berkeley.edu

Pablo Arbeláez
Universidad de los Andes, Colombia
pa.arbelaez@uniandes.edu.co

Panna Felsen
University of California, Berkeley
panna@eecs.berkeley.edu

Jitendra Malik
University of California, Berkeley
malik@eecs.berkeley.edu

Abstract
We present a method that segments moving objects in monocular uncalibrated videos using a combination of moving segment proposal generation and moving objectness ranking. We compute segment proposals in each frame by multiple figure-ground segmentations on optical flow boundaries, we call them Moving Object Proposals (MOPs). MOPs increase the object detection rate by 7% over static segment proposals of [23]. MOPs are ranked by a Moving Objectness Detector (MOD) trained from image and motion fields using a multilayer convolutional network. The MOD discards over/under segmentations and focuses on the moving parts of the scene, but remains agnostic to the object categories contained in the video. Finally, the filtered 2D MOPs are extended into temporally coherent spatio-temporal tubes by label diffusion in a dense point trajectory embedding. Our results on the popular Moseg and VSB100 video benchmarks show the proposed method achieves much higher overlap with ground-truth objects using fewer tube proposals than previous works; further, it reaches levels of ground-truth coverage much beyond the maximum possible with existing supervoxel or trajectory clustering approaches.

1. Introduction
We want to compute spatio-temporal segments that span moving objects in monocular uncalibrated videos. Accurate object segments are useful for weakly supervised learning from video [32], video editing and montage [35], surveillance and tracking [5], activity recognition etc. Segmentation is challenging when the objects are deforming and/or highly articulating; then their motion is hard to track and long-range correspondences, that would much facilitate segmentation [8], are hard to establish. Further, clutter from objects moving in crowds and from the static background often causes segments to leak across faint contours of similar-looking objects [41].

Prior work [34, 38] has used optical flow to improve image boundary detection, as a means to better video segmentation [14]. Optical flow carries useful information for segmentation: motion is smooth across interior object boundaries, e.g., due to clothing or surface markings. However, combining flow and image boundaries has so far only shown moderate success [13, 38]. This is primarily due to optical flow misalignments with true image boundaries: flow “bleeds” across occluding contours to the background [39] because background pixels mimic the motion of the nearby foreground occluding boundaries, as Figure 2 shows. Approaches that condition on a static segmentation map to boost or suppress boundary fragments using optical flow, such as [28, 34, 38], are upper-bounded by the performance of the static boundary detector: for high thresholds, many boundaries are missed with no hope to be recovered; for low thresholds, overwhelming image clutter causes regions to be too small for the flow to be aggregated effectively to fight “bleeding” [38].

Our first insight in this work is to use optical flow boundaries for object segmentation directly, without considering improving image boundary detection. Recent benchmarks, such as VSB100 [14], which evaluate jointly object segmentation and boundary detection, assuming their direct interrelation: better image boundaries will produce better video segments. We challenge this view by providing boundaries misaligned with ground-truth as input to the multiple figure-ground segmentation algorithm of [23] in each video frame. The resulting Moving Object Proposals (MOPs) increase the detection rate by 7% over state-of-the-art static segment proposals, proving the usefulness of motion for object detection.

Our second contribution is a moving objectness detector for ranking MOPs in each frame in order to discard over/under segmentations and static parts of the scene. Such
filtering is necessary for computational reasons: we target good coverage of objects in each frame with a small number of MOPs. Our moving objectness detector detects instances of all object categories that exhibit independent motion, namely people, vehicles, birds, fish, etc. While in the past such a multi-modal detector would require a large number of templates to be evaluated, e.g., in a DPM framework [12], current multi-layer convolutional architectures share mixtures of different abstractions [24]. Our MOD is a dual-pathway multilayer Convolutional Neural Network (CNN) that operates on image and motion fields. It outperforms previous hand-designed center-surround saliency measures and other competitive multilayer objectness baselines [1, 20]. Previous works either assume known object categories to segment, e.g., car or pedestrian tracking-by-detection, or operate completely bottom-up, relying on color/motion. In this work we take an intermediate stand: we want to detect any moving object in the scene, yet we remain agnostic to the exact object categories.

The filtered per frame MOPs are extended in time into trajectory clusters using Random Walkers [18] on a dense trajectory motion embedding [8]. Each trajectory cluster is then projected to a dense pixel tube and accurate image boundaries are recovered.

In summary, our contributions are:

- Moving object proposals (MOPs) from multiple segmentations on optical flow boundaries that improve upon static image proposals, especially in cluttered scenes.
- A moving objectness detector (MOD) learnt from image and motion fields that outperforms center-surround saliency or static objectness measures on ranking static segments and tube proposals.
- Random walks in a trajectory motion embedding for extending 2D segments into spatio-temporal trajectory clusters.
- Multiple trajectory clusterings as a simple way of generating additional tube proposals. It handles segmentation ambiguities better than single level trajectory clustering, and remains tractable in combination with our MOD ranker.

We test our method on the two largest video segmentation benchmarks: Moseg [8] and VSB100 [14]. Our goal is to maximize Intersection over Union (IoU) of our tube proposals with the ground-truth objects using as few proposals as possible. This has recently become the standard performance metric for segment proposal generation in the static domain [3, 23], which our work extends to videos and shows improvements in spatio-temporal (rather than spatial) IoU. In each video, 55-65% of ground-truth objects are captured in the challenging VSB100 benchmark using 64-1000 tube proposals, much higher than competing approaches of [14, 27, 41]. We empirically show our method can handle both articulated objects as well as crowded video scenes, which are challenging cases for existing methods and baselines. Our code is available at www.xxxx.xxxxxxxxxxx.xxx.

2. Related work

Previous works are categorized according to their information regarding the object classes in the scene into: i) top-down tracking methods, and ii) bottom-up video segmentation methods. Tracking methods assume the object class known, e.g., pedestrian or car trackers [6, 15]. They take advantage of category specific detectors to focus on the relevant parts of the scene.

Video segmentation methods are oblivious to object categories. Our method belongs to this category. Methods of [19, 41] group pixels based on color and/or optical flow similarity to produce multiscale spatio-temporal segmentation maps. Each spatio-temporal superpixel is called supervoxel. Works of [8, 31] cluster dense point trajectories [36] using 2D motion similarities. They have shown excellent results on benchmarks of mostly rigid objects. Work of
Figure 2: Moving object proposals (MOPs). a) Optical flow field. b) Motion boundaries obtained by applying the structured forest boundary detector of [11] on optical flow magnitude. c) The MOP best overlapping with ground-truth. d) Static image boundaries detected by the structured forest boundary detector on the RGB image. e) The static proposal best overlapping with the ground-truth segment, output of the state-of-the-art static proposal generation algorithm of [23]. It fails to capture the dancer as a whole. Interior (fake) contours due to the clothing of the dancers cause segment fragmentations. g) The trajectory cluster best overlapping with ground-truth. It leaks across multiple people and over-fragments the dancer. The proposed MOPs capture moving objects missed by static segments due to interior contours, or by trajectory clusters due to articulated or faint motion.

Many of the aforementioned approaches do not show results on standard benchmarks and comparison with them is difficult. In our experimental section, we compare with the popular supervoxel methods of [13,41] and the state-of-the-art method of [27] which are scalable and whose code is publicly available.

3. Moving Object Proposals (MOPs)

We generate segment proposals for the moving objects in each frame by providing motion boundaries as input to a multiple figure-ground segmentation algorithm. The resulting object proposals, which we call Moving Object Proposals (MOPs), capture moving objects that are missed by static segments due to fake interior contours, and/or by point trajectory clusters due to articulated or faint motion.

Given a video sequence, we compute the optical flow field in each frame using large displacement optical flow of Brox and Malik [7]. Then, we compute motion boundaries by applying the state-of-the-art structured forest boundary detector of Dollar et al. [11] on the magnitude of the optical flow field, which we replicate into a three channel image. Though the boundary detector has been trained on static image statistics of the BSDS boundary benchmark, it effectively detects boundaries of the flow field, as shown in Figure 2b. We did not consider re-training the detector using optical flow input because misalignments of the flow boundaries with the true image boundaries (due to flow “bleeding”) widely vary depending on the background texturedness, which would confuse the detector.

We use motion boundaries as input to the geodesic object proposal method of [23]. Given a boundary map as in-
4. Multilayer Moving Objectness Detector

Many MOPs under-segment or over-segment the objects, or capture background (static) parts of the scene. We train a moving objectness detector (or else ranker) to discard wrong MOPs and cast attention to the moving objects. A ranker is an essential part of an object proposal method because it allows to capture most of the objects in the scene with a small number of proposals. Previous proposal generation methods in the static domain [3, 23, 40] do not employ such a ranker, or have a rather basic one; they postpone the problem of choosing the “good” proposals to a later object detection step. In videos, due to the high data throughput, such a ranker becomes an essential bit for scalable video segmentation.

We train a Moving Objectness Detector (MOD) using a multilayer convolutional neural net (CNN) with a dual-pathway architecture operating on both image and flow fields, shown in Figure 3. The architecture of each network stack is as follows: \( C(7, 96, 2) - RL - P(3, 2) - N - C(5, 384, 2) - RL - P(3, 2) - N - C(3, 512, 1) - RL - C(3, 512, 1) - RL - C(3, 384, 1) - RL - P(3, 2) - FC(4096) - RL - D(0.5) - FC(4096) - RL \). For the optical flow channel, we supply a 3 channel image containing scaled \( x \) and \( y \) displacement fields and the optical flow magnitude. We did not consider background motion stabilization. The relu7 features of the image and flow stacks are joined and a final layer is trained to regress to intersection over union score of the input bounding box, as shown in Figure 3.

We initialize the weights in each stack using the network of [17] trained on the ImageNet detection task of 200 object categories from RGB images. Most of the moving object categories are well represented in the ImageNet training set. We also expect a network trained for object detection, as opposed to image classification, e.g., [22], to have incorporated some notion of objectness. We finetune the network weights using bounding boxes of image and flow collected from the training sets of the VSB100 and Moseg video benchmarks. For each bounding box, we regress to its intersection over union score with a ground-truth object. We train our MOD using standard stochastic gradient descent with momentum in Caffe [21], a publicly available deep learning package.

5. Tube proposal generation

Given a MOP detected at frame \( t \), we want to produce a corresponding temporally coherent tube proposal that respects the MOP segmentation at frame \( t \). Ideally, the tube proposal should track the corresponding object in time until its final occlusion. Point trajectory clustering has been shown to provide clusters with framesspans longer than supervoxel methods [14] that do not exploit point trajectories. Trajectory clusters are robust to object shape variations in time because they do not rely on static image boundaries but rather on long range motion trajectory similarities. We extend 2D MOPs to trajectory clusters by diffusing per frame trajectory labels through a trajectory motion embedding using Random Walkers [18], as depicted in Figure 4.
Given a video sequence, dense point trajectories are computed by linking optical flow fields [37]. A trajectory terminates when the forward-backward consistency check fails, indicating ambiguity in correspondence. This is usually the case under occlusion or dis-occlusions of the reference pixel, or under low image texturedness. Let $T$ denote the set of trajectories in the video and let $n$ denote the number of trajectories, $n = |T|$. We compute pairwise trajectory affinities $A \in [0, 1]^{n \times n}$ where motion similarity between two trajectories is a function of their maximum velocity difference, as proposed in [8]. We compute affinities between each pair of trajectories that overlap in time and are within a spatial distance of 60 pixels. Trajectory affinities are visualized in Figure 4b).

Let $t_i$ denote the frame that $\text{MOP}_i$ is detected. Point trajectories that intersect with $t_i$ are labelled as foreground or background. They are shown in Figure 4d in blue and light blue, respectively. Trajectories that terminate before or are initialized after $t_i$ are unlabelled. They are shown in white in Figure 4e. Let $x \in \{0, 1\}^n$ denote trajectory labels, 1 stands for foreground and 0 for background. Let $F$ denote the foreground and $B$ the background trajectory sets, respectively, and let $M = F \cup B$ denote the set of labelled (marked) trajectories and $U = T \setminus M$ the set of unlabelled trajectories. Let $L$ denote the trajectory unnormalized Laplacian matrix: $L = \text{Diag}(A1_n) - A$, where $\text{Diag}(y)$ stands for a diagonal matrix with vector $y$ in the diagonal. We minimize the random walker cost function proposed in [18]:

$$
\min_{x} \frac{1}{2}x^T L x \quad \text{subject to} \quad x_B = 0, \quad x_F = 1.
$$

(1)

It is easy to show that minimizing $x^T L x$ is exactly equivalent to minimizing $\sum_{i,j} A_{ij} (x_i - x_j)^2$. We relax $x$ to take real values, $x \in [0, 1]^n$. Then Eq. 1 has a closed form solution given by: $L_U x_U = -L_M x_M$, where $x_U$ are the labels of the unlabelled trajectories we are seeking, and $x_M$ are the labels of the marked trajectories. We approximate computationally this close form solution by performing a sequence of label diffusions using the normalized affinity matrix:

$$
x' = \text{Diag}(A1_n)^{-1} A x.
$$

(2)

We have found 50 diffusions to be adequate for our radius of affinities of around 60 pixels in each frame. We show in Figure 4f the diffused labels. The deer has been fully segmented from its background.

**Multiscale trajectory clustering** We compute additional tube proposals by multiscale trajectory clustering. Previous state-of-the-art video segmentation methods that use point trajectories focus on obtaining a single trajectory clustering [8]. We have found that multiple trajectory clusterings is a simple way to deal with segmentation ambiguities caused by motion variations of the objects in the scene. It handles over and under fragmentations of single trajectory clustering.

Let $V \in \mathbb{R}^{n \times K}$ and $\lambda \in [0, 1]^{K \times 1}$ denote the top $K$ eigenvectors and eigenvalues of the normalized affinity ma-
trix \( \text{Diag}(A_1^n)^{-1} A \), respectively. We compute multiple trajectory clusters by discretizing varying number of eigenvectors using the method of \cite{42}; it provides a deterministic discretization as compared to \( K \)-means in embedding coordinates, which depends on cluster center initialization. We used 50 as the maximum number of eigenvectors in all our experiments.

**Trajectory clusters to pixel tubes** We map trajectory clusters to pixels using a weighted average over supervoxels obtained by greedily smoothing superpixels labels in time, similar to \cite{14}. The weight of each supervoxel is its Intersection over Union (IoU) score with the trajectory cluster. We threshold the weighted average to obtain a binary spatio-temporal segmentation for each trajectory cluster, shown at Figure 4g. Notice that sharp boundaries have been recovered despite the misaligned boundaries of the generating MOP in Figure 4c. Also, image parts sparsely populated by trajectories due to low image texturedness, such as the deer body, have been correctly labelled.

**6. Experiments**

We test our method on the two largest publicly available video segmentation benchmarks: VSB100 \cite{14} and Moseg \cite{8}. VSB100 contains 100 video sequences, 40 training and 60 testing, they are high resolution videos collected from Youtube. Object motion can be very subtle (chameleon sequence) or extremely articulated (capoeira dance). Many crowded scenes are included, such as video scenes from a parade, a cycling race, beach volley, ballet, salsa dancing etc. We focus on “rigid and non-rigid motion subtasks” of the VSB100 benchmark that concern moving object segmentation (as opposed to segmenting static background).

The Moseg dataset contains 59 videos that depict scenes from the Hollywood movie “Miss Marple”, as well as cars and animals, such as cats, rabbits, bears, camels, horses, etc. The moving objects have distinct motion to surroundings and the scenes are relatively uncluttered, with few (one or two on average) objects per video.

First, we benchmark our full method on tube proposal generation and compare against state-of-the-art single level point trajectory clustering of \cite{27}, as well as the supervoxel methods of \cite{14,41}. Second, we benchmark the proposed 2D MOPs on single image segmentation metrics. We show that when combined with static segment proposals of \cite{23} the average best overlap, coverage and detection rates increase to levels beyond the saturation point of the static proposal algorithms. Last, we benchmark our moving objectness detector on ranking 2D segments as well as spatio-temporal tube proposals, and compare it with alternative multilayer architectures, hand-designed center-surround saliency and static image objectness measures.

**Benchmarking our tube proposals on video segmentation** We compare our method with popular supervoxel methods of \cite{14,41} as well as the state-of-the-art point trajectory clustering and pixelization method of \cite{27}. Non-streaming versions of hierarchical spatio-temporal segmentation \cite{19}, distributed with the code of \cite{41}, were not scalable enough to use in our benchmarks. Work of \cite{14} showed that smoothing multiscale state-of-the-art segmentation maps of \cite{2} in time using optical flow, provides a state-of-the-art segmentation method in VSB100; we use it here as one of our baselines. Our supervoxel computation in the last step of our pipeline is also based on their approach. For \cite{27,41} we use the publicly available codes. The method of \cite{41} segments a video sequence into temporally coherent superpixels and provides a multi-level segmentation map as output. In contrast, motion segmentation method of \cite{27} outputs a single segmentation map. For both our method and the baselines we use our moving objectness detector to rank the resulting spatio-temporal segments. Score diversification has been used as in \cite{10} for a soft non-maxima suppression step.

We show tube proposal segmentation results on VSB100 and Moseg benchmarks in Figure 5 columns 1 and 2, respectively. We score separately MOP tubes, multiscale trajectory clusters (ms tr. cluster) as well as their union, which is our full method. Our method outperforms the baselines reaching well beyond previous levels of IoU with a moderate number of tube proposals per video. Multiscale trajectory clustering overall does not offer additional boost over MOP tubes. Notice also the big difference in performance of our method and baselines across the two datasets, indicative of the challenging nature of VSB100 over Moseg.

**Benchmarking 2D Moving Object Proposals on 2D object segmentation** We consider the following four widely used static image segmentation metrics: a) **Average best overlap**: the average (across all 2D ground-truth segments in our dataset) of the best IoU score of a ground-truth segment with all segment proposals. b) **Coverage**: the weighted average of IoU scores, weighted by the area of the ground-truth segments (larger segments matter more). c) **Detection rate at 50%**: the percentage of ground-truth segments that have IoU above 50 % with a segment proposal. d) **Detection rate at 70%**. It has been shown in \cite{23} that a threshold of 70 % better indicates perceptual similarity between objects, and is a better metric for object detection. We further present *anytime best* (ab) versions of the a, c and d metrics, where for each ground-truth tube (rather than 2D segment) we consider the best overlap with a segment proposal throughout its timespan; this metric upper bounds the performance of our MOP tubes, described in Section 5.

We show results of the proposed MOPs, static geodesic object proposals of \cite{23} (GOPs) and combined segment
Figure 5: Video segmentation results in VSB100 (col. 1) and Moseg (col. 2) datasets. Our method combines proposals from MOP tubes (Sections 3, 5) and multiscale spectral clustering (Section 5) and outperforms by a large margin previous approaches: we reach much higher levels of segmentation Intersection over Union (IoU) scores with smaller number of tubes, and further reach well beyond the saturation level of previous approaches. Combining multiscale spectral trajectory clustering with MOP tubes gives only a small boost at VSB100, suggesting MOP tubes are sufficient for good performance. Notice the lower performance of all methods in VSB100 in comparison to Moseg, due to its increased difficulty. Columns 3, 4: Ranking space-time tube proposals (col. 3) and 2D segments (col. 4). Our dual-pathway CNN regressor (piCNN-regress) outperforms other CNN alternatives and hand-designed center-surround saliency.

Table 1: Image segmentation results in VSB100 and Moseg. (ab) stands for anytime best, as explained in the text. We compare static geodesic object proposals (GOPS) of [23], MOPs proposed in this work, and a joint method that combines both sources of segment proposals (GOP+MOP). Next to each method in parentheses we show the number of proposals used. We highlight in bold font the most significant metric improvements. The performance boost by combining static and moving object proposals, though significant for both datasets, is larger in the VSB100 benchmark due to the large amount of static boundary clutter, in comparison to the relatively uncluttered Moseg scenes. MOPs provide a boost of almost 8% and 6% for 2D detection rates at 50% and 70% overlap thresholds, respectively. They further boost by 4% 6% the corresponding anytime best detection rates.

| Dataset | GOP (2715) | GOP (873) | GOP+MOP (2659=1786+873) | GOP (2500) | GOP (839) | GOP+MOP (2512=1673+839) | avg best ol | coverage | det 50% | det 70% | avg best ol | det 50% | det 70% |
|---------|------------|-----------|-------------------------|------------|-----------|-------------------------|--------------|----------|---------|---------|--------------|---------|--------|
| VSB100  | 53.74      | 66.84     | 60.34                   | 26.12      | 65.08     | 82.6                    | 53.74        | 66.84    | 60.34   | 26.12   | 65.08        | 82.6    | 53.74  |
| MOP     | 46.47      | 61.3      | 47.25                   | 13.85      | 57.92     | 73.75                   | 46.47        | 61.3     | 47.25   | 13.85   | 57.92        | 73.75   | 46.47  |
| GOP+MOP | 56.17      | 69.85     | 66.48                   | 31.50      | 67.15     | 86.14                   | 56.17        | 69.85    | 66.48   | 31.50   | 67.15        | 86.14   | 56.17  |
| MOSEG   | 68.47      | 76.56     | 87.59                   | 64.54      | 74.72     | 79.03                   | 68.47        | 76.56    | 87.59   | 64.54   | 74.72        | 79.03   | 68.47  |
| MOP     | 57.74      | 68.49     | 70.57                   | 37.94      | 66.42     | 59.68                   | 57.74        | 68.49    | 70.57   | 37.94   | 66.42        | 59.68   | 57.74  |
| GOP+MOP | 69.65      | 78.29     | 87.59                   | 70.21      | 75.38     | 83.87                   | 69.65        | 78.29    | 87.59   | 70.21   | 75.38        | 83.87   | 69.65  |

Proposal importance. A key aspect of our system is the ability to rank proposals efficiently. Table 1 shows the result of ranking 2D static object proposals from the VSB100 and Moseg datasets on a per-image basis. Following the same protocol as section 5, we use the dual-pathway CNN as regressor (piCNN-regress) against our implementation of the center-surround saliency function. The ranking performance suggests that piCNN-regress is able to efficiently rank proposals in a way that maximizes the IoU with the ground-truth, outperforming the hand-designed saliency model used in [23]. Importantly, objects in videos do not move continuously. At frames when they are static, there are no motion boundaries and MOPs miss them. However, MOPs segment them correctly at frames at which they move, and label diffusion propagates the segmentation from those “lucky”, large motion frames, to the rest of the video.

Benchmarking Moving Objectness Detector on segment and tube ranking We test our moving objectness detector on ranking 2D MOPs and spatio-temporal tubes produced by multiscale trajectory clustering in the VSB100 dataset. We define the score of each tube as the sum of the scores of the bounding boxes throughout its lifespan. We use sum instead of average because we want to bias towards longer tube proposals. We show the corresponding ranking curves produced by averaging across images in the first case and across video sequences in the second in Figure 5 at columns 3 and 4, respectively. The curves indicate how many segments/tubes are needed to reach a specific level of Intersection over Union score with the ground-truth segments/tubes. We compare our dual-pathway (pi) multilayer CNN regressor from image and flow fields (piCNN-regress) against a pi classification CNN (piCNN-class), an image only CNN (imgCNN), a flow only CNN (flowCNN), our implementation...
tation of a standard center-surround saliency measure from optical flow magnitude [16], and an objectness detector using the 7000 category detector from the Large Scale Domain Adaptation (LSDA) work of [20]. Our pi classification CNN is trained to classify boxes as positive or negatives using a threshold of 50% of IoU, instead of regressing to their IoU score. For our LSDA baseline, we consider for each 2D segment bounding box \( b \) a weighted average of the confidences of the detection boxes of [20], where weights correspond to their intersection over union with box \( b \). We have found this objectness baseline to provide a competitive static objectness detector. Notice that our CNN networks operate on the bounding box of a segment rather than its segmentation mask. While masking the background is possible, context is important for judging over and under-segmentations. Our pi CNN regressor performs best among the alternatives considered, though has close performance with the pi classification CNN.

**Discussion - Failure cases** In the VSB100 dataset, our failure cases often concern temporal object fragmentations. They are caused by large motion or full object occlusions, where the point trajectories terminate and the label diffusion cannot proceed any further. Our method, as well as our baselines, would benefit from an additional temporal linking step, where similarly looking tubes are linked across frames to form longer ones; we did not consider such a step to keep the method clean. In Moseg, most failure cases are due to inaccurate mapping of trajectory clusters to pixels: we often slightly leak to the background or miss thin animal limbs. A more elaborate trajectory cluster to pixel mapping, that considers background stabilization and low-rank background appearance [9] would potentially provide more accurate tube pixel masks.

**Complexity** The following numbers are for a single cpu. Large displacement optical flow takes on average 16 secs per image. Given an optical flow field, computing MOPs takes 4 seconds on a 700X1000 image. The projection of each MOP to the trajectory embedding takes 2 seconds for 70000 trajectories, all MOPs can be projected simultaneously using matrix diffusion operations. Supervoxel computation is causal and takes 7 seconds in each frame. Computing motion affinities for 70000 trajectories takes 15 seconds in each video. Our supervoxel computation, optical flow computation, MOP computation and projection are completely parallelizable.

7. Conclusion

We have presented a simple method for computing tube proposals in monocular videos. 2D segment proposals from optical flow boundaries are pruned by a multilayer CNN moving object detector, and are diffused into a trajectory embedding in order to provide temporally coherent tubes. Our moving object proposals complement static ones, and boost by a large margin their performance on capturing moving objects, especially in cluttered, challenging scenes. The proposed moving object detector effectively discards over and under object fragmentations or background parts of the scene, and provides a ranking that allows to capture the ground-truth objects with few tube proposals per video. The proposed method takes a step towards bridging the gap between video segmentation and tracking research, by going beyond a single object category detector for video parsing. Using motion saliency through motion boundaries, motion-based ranking and trajectory embedding diffusion, our tube generation method is more economic than static object proposal generation methods. Static image segmentation ambiguities are diminished under the strong effect of motion perception.

**References**

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In CVPR, pages 73–80, 2010. 2
[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 2294–2301, June 2009. 6
[3] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In Computer Vision and Pattern Recognition, 2014. 2, 4
[4] D. Banica, A. Agape, A. Ion, and C. Sminchisescu. Video object segmentation by salient segment chain composition. In The IEEE International Conference on Computer Vision (ICCV) Workshops, June 2013. 3
[5] C. Bibby and I. Reid. Real-time tracking of multiple occluding objects using level sets. In CVPR, 2010. 1
[6] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In ICCV, 2009. 2
[7] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. TPAMI, 2010. 3
[8] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In ECCV. 2010. 1, 2, 5, 6
[9] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? J. ACM, 58(3):11:1–11:37, June 2011. 8
[10] J. Carreira and et al. Constrained parametric min-cuts for automatic object segmentation, 2010. 6
[11] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In ICCV. International Conference on Computer Vision, December 2013. 2, 3
[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. TPAMI, 32, 2010. 2
