LEARNING TO GENERATE VIDEO OBJECT SEGMENT PROPOSALS

Jianwu Li, Tianfei Zhou, Yao Lu

Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology

ABSTRACT

This paper proposes a fully automatic pipeline to generate accurate object segment proposals in realistic videos. Our approach first detects generic object proposals for all video frames and then learns to rank them using a Convolutional Neural Networks (CNN) descriptor built on appearance and motion cues. The ambiguity of the proposal set can be reduced while the quality can be retained as highly as possible Next, high-scoring proposals are greedily tracked over the entire sequence into distinct tracklets. Observing that the proposal tracklet set at this stage is noisy and redundant, we perform a tracklet selection scheme to suppress the highly overlapped tracklets, and detect occlusions based on appearance and location information. Finally, we exploit holistic appearance cues for refinement of video segment proposals to obtain pixel-accurate segmentation. Our method is evaluated on two video segmentation datasets *i.e.* SegTrack v1 and FBMS-59 and achieves competitive results in comparison with other state-of-the-art methods.

1. INTRODUCTION

Object segmentation proposal has recently attracted substantial attention in the computer vision community. Given an image, the problem provides a manageable number of regions such that each object can be approximately delineated by at least one region. The last two years have witnessed great performance improvement on this task: recent benchmark [1] shows that in PASCAL Visual Object Classes Challenge [2], the most promising current approaches [3] can cover more than 60% object instances whose overlap with the ground-truth masks is above 0.7 with respect to 1000 proposals in each image. The impressive progress has been demonstrated to promote various vision applications, including object detection, semantic segmentation, etc.

Following the successes of single image object proposals, we aim to explore how we can automatically and effectively generate object segment proposals in video domain. For many video analytics tasks, *e.g.* action localization or recognition, spatio-temporal object proposals are expected to largely reduce the computational cost as well as improve the performance. However, the problem is rather challenging, which has the following requirements:

- It should be able to capture the accurate boundaries of object instances. This is important to capture finer information for objects of interest, especially for easily deformable objects such as humans and animals. Actually, the latest results [4, 5] have shown that precise human silhouette boundaries can benefit human action localization and recognition.
- It should achieve a high detection recall with as few proposals as possible. Like object proposal in static images, spatio-temporal proposals can be used as a preprocessing to improve video tasks like moving object detection. Thus, a reduced set of proposals, keeping the true objects as much as possible, will not only speedup more advanced tasks but boost their performance.

Although there has been some recent work [6, 7, 8, 9] proposed to resolve related tasks, the above challenges are not fully addressed. For instance, [6, 7] can only extract a single dominant object for each video, while [8] requires that objects have to appear in the first frame. [9] addresses these restrictions, however, we observe that they achieve a high recall by means of producing a large number of proposals, many of which are false positives.

In this study, we develop a novel method to fulfill the aforementioned requirements when proposing object segments in videos. Given a video, our approach starts from frame-level object segment proposals using the state-of-the-art method MCG [3]. For a 640×480 video, more than 2000 object segments are generated in each frame. Directly inferring over all frames in such a large number of proposals is computationally expensive, and often results in too many non-object proposals. To determine a reduced proposal set, we adopt two steps: 1) Inspired by [10], we score the proposals with a combination of appearance and motion CNN features, and discard those with low objectness scores. 2) We augment the proposal set by propagating the highest-scoring proposals in each frame to neighboring frames. Such a propagation can not only provide more accurate object segments, but facilitates latter steps to find coherent regions across frames. Next, each proposal is tracked forward and backward across frames to build a graph of tracklets. To pick reliable space-time proposals, we take a greedy scheme to select the tracklets with high confidence and independence, and suppress those that

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

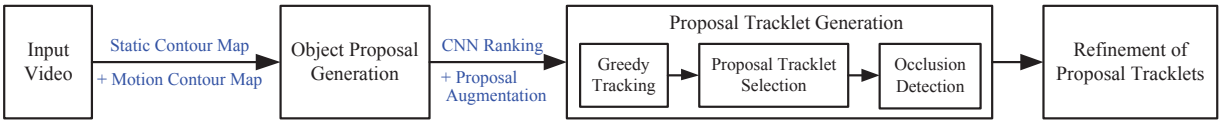

Fig. 1. Pipeline of our video object segment proposal method.

are highly overlapped with them. Finally, in order to achieve pixel-accurate segmentations, we learn a holistic appearance model for each tracklet, and estimate the label of each pixel in a Markov Random Field.

2. RELATED WORK

In this section, we discuss the most relevant work on video segment proposal, unsupervised tracking and video segmentation.

Object segment proposal techniques [3, 11] have been actively studied in the past few years. However, there is only a limited amount of work on spatio-temporal proposal in video domain. Work of [6, 12, 7] firstly consider this problem in the task of video object segmentation. All these methods apply object proposal techniques in individual frames to obtain a pool of object segments. They assume that the primary object appears in each video frame, and propose various methods to obtain a tracklet of proposals over the entire sequence. For example, in [12], the authors cast the selection of object segments as finding maximum weight cliques in a region graph, while [7] finds the optimal track of the dominant object by solving a longest path problem in a layered directed acyclic graph. One major limitation of the above methods is that they can only capture a single primary object. [8] overcomes this by simultaneously tracking all segments generated in each frame. However, this method restricts that each segment tracks must start from the first frame. Along this thread, work of [9] lifts this restriction by combining forward tracking and backtracking to track the segments through occlusions. Note that [8] and [9] track all the segments in each video frame, this is computationally expensive and also produce many tracks with low spatio-temporal coherency. Our work is also related with [10] and [13]. [10] combines long-term point trajectories with spatio-temporal segment regions for segmenting objects in videos, while in [13], the author utilizes the supervoxel cues directly for proposals.

The problem we address is also related to unsupervised object tracking. Different from conventional tracking methods which require annotation in the first frame, unsupervised tracking methods [14, 15, 16, 17] discover object candidates in each frame, and then link them across frames to form tracklets. [14] proposes a method for joint object discovery and tracking in a large video collection. Work of [15, 16, 17] attempt to address the problem of action localization. They firstly propose object-likely regions using [18, 19], and then link the proposals to locate human actions. All the methods mentioned above represent objects using bounding box, while this representation might be good enough for certain types of roughly-rectangular objects (*e.g.* cars), it is certainly imprecise to represent non-rigid objects (*e.g.* humans and animals). Thus, in this paper, we focus on the segmentation representation, and enforce spatio-temporal constraints between the segments to track them.

3. OBJECT PROPOSAL GENERATION

There has been a large number of methods that focus on object proposal generation based on color, edge and texture cues [20, 3, 21]. We use mutilscale combinatorial grouping (MCG) [3] in this paper due to its high detection recall and segmentation accuracy as validated in PASCAL VOC2012 Challenge. For an image, MCG starts by computing a hierarchy of regions at multiple scales, which are represented by an ultrametric contour map (UCM), a weighted contour image with a remarkable property of yielding closed contours (*i.e.* regions) for any threshold. Then, object candidates are produced by combinatorially grouping these regions.

In our method, given a video, we compute two complementary ultrametric contour maps for object proposal generation in each frame: 1) *static contour map* computed on the original RGB frame; 2) *motion contour map* calculated on the magnitude map of optical flow [22]. Directed by boundary strength of contour maps, we can compute a ranked list of object proposals by combinatorial grouping. For a 640×480 frame I_t, we produce a proposal set $\mathcal{P}_t$ with more than 2000 object segments, including true positives as well as a number of false positives. Next, we take two steps to reduce the number of object hypotheses and then augment the proposal set by introducing more temporally coherent proposals.

We learn to rank the object segment proposals using a two-stream CNN regressor [10] which combines spatial and motion networks, and discard the low-scoring proposals. Given a region resized to 227×227 pixels, the spatial-CNN operating on image field aims to capture the appearance of objects, while the motion-CNN operates on flow field and captures the movement of objects. For the flow channel, we stack scaled x and y components and the magnitude of flow to form a 3-dimensional image. For each region, the CNN features we use are the concatenation of the relu7 layer (4096 dimensions) from both networks. Finally, for a proposal $\mathcal{P}_t^i$,

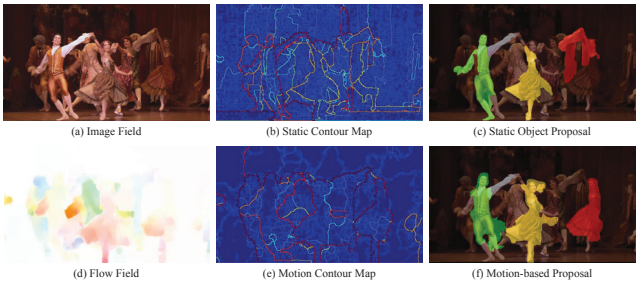

Fig. 2. **Object Proposal.** We exploit static and motion cues for object proposal generation. Clearly, we see that the static contour map (b) and motion contour map (e) complement to each other. That is, (b) provides more accurate closed boundary for objects labelled with green and yellow colors in (c)(f), while (e) captures the contour of the dancer in red. In (c) and (f), we show three proposals obtained by applying MCG to the two contour maps respectively.

a regressor takes its CNN features as input to compute an objectness score S_t^i, which is the probability of $\mathcal{P}_t^i$ containing an object. Once the scores are calculated for all proposals, we simply discard the proposals with scores less than 0.5.

4. PROPOSAL TRACKLET GENERATION

We track all proposals in $\{\mathcal{P}_t, t = 1 \dots T\}$ forward and backward over the entire sequence to obtain a redundant set of tracklet candidates. Then, we formulate the selection of temporally coherent tracklets as finding maximum-weight independent set over a graph of tracklets. Note that we in the present section assume an object exists in all frames, and initialize each tracklet as an entire video. This will be violated in cases of full occlusions or objects leaving the scene. We will address this problem in the next section.

4.1. Greedy Tracking

Take a proposal $\mathcal{P}_t^i$ at frame I_t as an example, and assume it will be tracked forward to frame I_{t+1}. For a proposal $\mathcal{P}_{t+1}^j$ at time $t+1$, we define the tracking score between the two proposals to be

$$s(\mathcal{P}_t^i, \mathcal{P}_{t+1}^j) = S_{t+1}^j + \lambda C(\mathcal{P}_t^i, \mathcal{P}_{t+1}^j) \tag{1}$$

where $C(\mathcal{P}_t^i, \mathcal{P}_{t+1}^j)$ measures the temporal consistency between the two proposals and λ is a scalar. In other words, to track $\mathcal{P}_t^i$, we aim to select a proposal in I_{t+1} with high objectness score as well as strong consistency with $\mathcal{P}_t^i$.

The pairwise term $C(\mathcal{P}_t^i, \mathcal{P}_{t+1}^j)$ is used to enforce temporal consistency of appearance and location between proposals in successive frames.

$$C(\mathcal{P}_t^i, \mathcal{P}_{t+1}^j) = C_c(\mathcal{P}_t^i, \mathcal{P}_{t+1}^j) * C_l(\mathcal{P}_t^i, \mathcal{P}_{t+1}^j) \tag{2}$$

Algorithm 1: Proposal tracklect selection in a graph

Input: A graph $G = (V, E, \omega)$
Output: A subset S

```
S ← ∅;
Sort V = {v'_1, v'_2, ..., v'_n} such that
  ω(v'_1) ≥ ω(v'_2) ... ≥ ω(v'_n);
while G is not empty do
    Let v'_i be the vertex satisfying
      v'_i = arg max_{v'_i ∈ V} ω_i / d_i in G;
    S ← S ∪ v_i;
    Remove v_i from G;
    Remove the neighbors of v_i whose overlap rate
      with v_i is above a threshold τ.
Return S
```

where C_c is the color similarity computed by color histogram intersection distance and C_l denotes the overlap rate between $\mathcal{P}_t^i$ and $\mathcal{P}_{t+1}^j$.

Once the scores between $\mathcal{P}_t^i$ and the proposals in I_{t+1} are computed, we employ a greedy scheme to track it by selecting the proposals with the largest scores in adjacent frames.

4.2. Proposal Tracklet Selection

This section presents our tracklet selection method to avoid generating tracklets that are highly overlapped.

Let $Tr = \{\tau_1, \tau_2, \dots, \tau_n\}$ be the set of tracklets computed in the previous subsection, where n is the number of tracklets. For every $\tau_i \in Tr$, we associate it with a score

$$\omega_i = S_T^k + \sum_{t=1}^{T-1} s(\mathcal{P}_t^i, \mathcal{P}_{t+1}^j) \tag{3}$$

which encodes the similarity of the segments in the track.

We then formalize our problem by constructing a graph $G = (V, E, \omega)$, in which $V = \{v_1, v_2 \dots, v_n\}$ is the set of vertices, each one corresponding to a tracklet tr_i, and E is the edge set connecting two vertices if and only if their tracklets share a common segment proposal. Besides, each vertex v_i is associated with a weight $\omega_i \in \omega$ that is defined to be equal to the tracklet score. Given the attributed graph, a greedy algorithm is proposed to address the selection of proposal tracklets. Iteratively, we find a vertex v_i (*i.e.* a tracklet) according to $v_i = \arg\max_{v_i \in V} \frac{\omega_i}{d_i}$ and put it into the subset S, where d_i is the degree of vertex v_i. Then, we remove v_i and its neighbors whose overlapping rate with v_i is too large from G. The greedy algorithm is described in Algo. 1.

4.3. Occlusion Handling

Above we assume the length of each tracklet be equal to the video length. This will not be satisfied once objects are occluded, which will result in invalid association. Our idea for

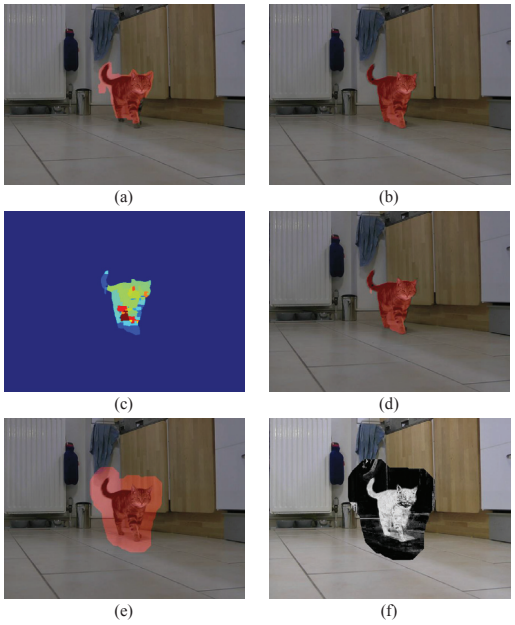

Fig. 3. **Tracklet Refinement.** (a) A segment proposal $\mathcal{P}_t^i$ at frame #42 in *cats01* sequence. The proposal is clearly not accurate in object boundary, especially in the tail part. (b) The segmentation result after refinement. (c) Weighted average map after supervoxel projection. (d) Binary segmentation mask of (c). (e) The mask in (d) provides us the positive pixels that belong to the cat. To obtain negative pixels for learning linear regression appearance model, we dilate the mask in (d). (f) The probability map for foreground computed by the linear regression appearance model.

occlusion detection is to model the variation of object appearance and location in each tracklet. We consider that when an object encounters occlusion, its appearance may vary faster than that with no occlusion. Besides, in our task, a tracklet may drift from one object to distant ones, leading to large location change. Particularly, eack proposal $\mathcal{P}_t^i$ in a tracklet is associated with a score $C(\mathcal{P}_t^i, \mathcal{P}_{t+1}^j)$ (Eq. (1)). We use the mean value ϕ of the scores of all proposals in a tracklet as a threshold to determine occlusion. In a tracklet tr_i, if all scores in K consecutive frames are below ϕ, we split tr_i into two parts from the first consecutive response. We repeat the above process several times until there is no occlusion detected.

5. TRACKLET REFINEMENT

In order to achieve pixel-accurate segmentations, tracklet refinement is considered here. To refine each tracklet, we map the segment proposals into pixels using an average score over supervoxels. The supervoxels are generated in a similar way with [7, 9]. The score of each supervoxel is the average of its intersection-over-union (IOU) score with the tracklets segment in each frame. Then, an objectness prior for each pixel in a frame is defined as the score of the supervoxel to which the pixel belongs. The refinement is finally carried out by solving a problem of pixel-level segmentation in which the objectness prior is integrated with a learnt holistic appearance likelihood of the tracklet.

The holistic appearance likelihood is learnt via linear regression [23]. A codebook of the RGB intensities is built for the tracklet. And the regression is carried out on the feature of codebook histogram, with positive and negative pixel samples obtained according to the objectness prior (see Figure 3).

	Training Set (29 sequences)			Testing Set (30 sequences)		
	Precision	Recall	F-measure	Precision	Recall	F-measure
[24]	79.17	47.55	59.42	77.11	42.99	55.20
[25]	81.50	63.23	71.21	74.91	60.14	66.72
[10]	70.12	55.70	61.42	69.03	54.11	54.23
Ours	**83.75**	**66.18**	**73.94**	**78.65**	**62.37**	**69.57**

Table 1. Comparison of our approach to [10], [24] and [25] on FBMS-59 dataset.

	birdfall	cheetah	girl	monkey-dog	parachute
[7]	155	633	1488	365	220
[8]	242	1156	1573	483	328
[6]	288	905	1785	521	201
[12]	189	806	1698	472	221
Ours	162	798	1251	455	218

Table 2. Results on the SegTrack v1 dataset. We did not consider the *penguin* sequence for evaluation because annotations provided in the ground truth are not reliable.

6. EXPERIMENTS

In this section we evaluate our approach on two datasets: SegTrack v1 [26] and FBMS-59 [25]. SegTrack v1 is a small benchmark for video object segmentation with six videos (*birdfall*, *cheetah*, *girl*, *monkey-dog*, *parachute*, *penguin*) with pixel-level foreground object annotations for each frame. The database offers various challenges, including camera motion, object deformation, motion blur, etc. The FBMS-59 dataset contains 59 image sequences which depict scenes from the Hollywood movie "Miss Marple", as well as cars and animals (*e.g.* cats, rabbits, bears, camels, horses, etc). Each sequence in FBMS-59 has pixel-wise ground-truth annotation for a sparse subset of frames (3-41).

To fully evaluate our method, we report results on two tasks: first, we evaluate our object proposal methods on FBMS-59 and compare it with the method in [10]; second, with our segmentation output, we evaluate the segmentation accuracy of our method on the two benchmarks.

	ABO	covering	$J = 0.5$	$J = 0.7$	ABO ab	$J = 0.5$ ab	$J = 0.7$ ab
GOP+MOP (2512=1673+839)	69.65	78.29	**87.59**	70.21	75.38	91.94	83.87
Spatial-MCG (1946)	72.03	78.31	85.14	66.91	75.35	92.07	78.31
Motion-MCG (1718)	71.54	76.85	82.17	50.2	74.11	85.81	60.29
Combined (1832 = 973+859)	**72.69**	**80.02**	86.10	**72.32**	**78.90**	**92.40**	**84.31**

Table 3. **Object Segment Proposal Evaluation.** We compare the method in [10], spatial-MCG [3], motion-MCG in this paper, and a method considers spatial- and motion-MCG. The number of proposals in each method is shown in parentheses. We can see that the results are significantly improved by combining motion-MCG and spatial-MCG. Furthermore, our method outperforms [10] by a large margin.

Object Segment Proposal We firstly evaluate the performance of object segment proposal algorithm on FBMS-59 dataset. As a measure of quality of a specific candidate with respect to an annotated object, we use the Jaccard index J, also known as overlap, defined as the intersection over the union of the two sets. When computing the overall quality for the whole database, we consider the following four widely used metrics: a) *average best overlap (ABO):* the mean best overlap for all the ground-truth instances in the database; b) *covering:* the weighted average of IOU scores, weighted by the size of the ground-truth segments; c) *recall at* $J = 0.5$: the percentage of ground-truth segments that have at least one segment proposal with IOU score above 0.5; d) *recall at* $J = 0.7$: a threshold of 0.7 asks for more perceptual similarity between objects and thus is more suitable for object detection. Besides, in order to evaluate our spatio-temporal proposals, we follow [10] to report *anytime best* (ab) versions of a, c and d metrics, where we compute the best overlap for a spatio-temporal proposal.

As shown in Table 3, we list the results of spatial-MCG, motion-MCG and the combination of the two methods. Furthermore, we compare these methods with the proposal method in [10]. We can clearly see that by combination of static and motion contour maps, we boost the accuracy of object proposal by a large margin: 7% performance improvement in terms of recall at $J = 0.7$.

Video Object Segmentation In the second part of our experiments, we evaluate our method on SegTrack v1 for video object segmentation task. We compare our method with those based on ranking object segment proposals [6, 12, 8, 7]. Note that for *birdfall* sequence, the object moves slowly. Thus, we obtain an object tracklet using background subtraction method. The segments are then refined using the proposed appearance model. For other sequences, we run our method to get a set of segment tracks. Among all the tracklets return, we report the performance on the best track. To quantify segmentation accuracy, we use the average per-frame pixel error rate compared to the ground-truth, which is defined as: $\mathrm{e} = \frac{\mathbf{XOR}(f, GT)}{F}$ where f is the segmentation results of the method, GT is the ground-truth labelling of a video sequence, and F denotes the number of frames in the video. As Table 6 shows, we obtain intermediate-level overall results in comparison with other 4 algorithms.

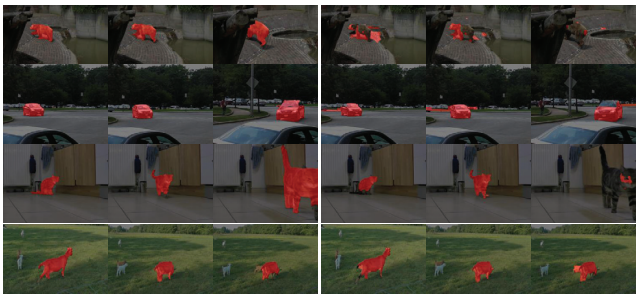

Fig. 4. Sample results of (a) the proposed method and (b) [10] on FBMS-59. From top to bottom: *bear01*, *cars04*, *cats01*, *goats01*.

Motion Segmentation We further assess our method on FBMS-59 for motion segmentation task. In this task, we compare our method with moving object segmentation method [10], multi-label dense motion segmentation method [25] and video over-segmentation method [24]. As in [25], we use *precision*, *recall* and *F-measure* to measure the performance of the methods. The results are listed in Table 1. Clearly, the proposed method outperforms other ones in terms of all metrics. We see that the spatio-temporal proposals extracted by [10] are often not accurate, especially in the situation of occlusion and scale variation. Benefiting from the proposed object proposal method, we can capture objects even during heavy occlusion (*e.g. bear01* in Figure 4). Besides, the tracking scheme ensures us to obtain spatial-temporal proposals with high coherency.

7. CONCLUSION

This paper presents an approach for automatically generating video segment proposals. We evaluate it on two video segmentation benchmarks: SegTrack v1 and FBMS-59, and the results demonstrate the effectiveness of the proposed method. We attribute its success to 1) the new frame-level segment proposal method incorporating both static and motion cues;

2) the unsupervised tracking method that links the segments into temporally coherent tracklets; 3) the occlusion detection scheme and finally 4) the refinement of tracklets using non-local appearance information. In future work, we will evaluate the performance of our method on other high-level tasks, such as human action localization or recognition.

Acknowledgements: This work was supported in part by the National Natural Science Foundation of China (No. 61271374 and No. 61273273)

8. REFERENCES

[1] J. Pont-Tuset and L. Van Gool, "Boosting object proposals: From Pascal to COCO," in *ICCV*, 2015.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[3] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *CVPR*, 2014.

[4] Jiasen Lu, ran Xu, and Jason J. Corso, "Human action segmentation with hierarchical supervoxel consistency," in *CVPR*, 2015.

[5] Hueihan Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *ICCV*, 2013.

[6] Yong Jae Lee, Jaechul Kim, and Kristen Grauman, "Key-segments for video object segmentation," in *ICCV*, 2011.

[7] Dong Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *CVPR*, 2013.

[8] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg, "Video segmentation by tracking many figure-ground segments," in *ICCV*, 2013.

[9] Zhengyang Wu, Fuxin Li, Rahul Sukthankar, and James M. Rehg, "Robust video segment proposals with painless occlusion handling," in *CVPR*, 2015.

[10] Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik, "Learning to segment moving objects in videos," in *CVPR*, 2015.

[11] Joao Carreira and Cristian Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation.," *IEEE Transactions on Software Engineering*, vol. 23, no. 3, pp. 3241–3248, 2010.

[12] Tianyang Ma and L. J. Latecki, "Maximum weight cliques with mutex constraints for video object segmentation," in *CVPR*, 2012.

[13] Oneata Dan, Jerome Revaud, Jakob Verbeek, and Cordelia Schmid, "Spatio-temporal object detection proposals," in *ECCV*, 2014.

[14] Suha Kwak, Minsu Cho, Ivan Laptev, Jean Ponce, and Cordelia Schmid, "Unsupervised object discovery and tracking in video collections," in *ICCV*, 2015.

[15] G. Gkioxari and J. Malik, "Finding action tubes," in *CVPR*, 2015.

[16] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid, "Learning to track for spatio-temporal action localization," in *ICCV*, 2015.

[17] Gang Yu and Junsong Yuan, "Fast action proposals for human action detection and search," 2015.

[18] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective searchfor object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[19] C. Lawrence Zitnick and Piotr Dollr, "Edge boxes: Locating object proposals from edges," *ECCV*, vol. 8693, pp. 391–405, 2014.

[20] B Alexe, T Deselaers, and V Ferrari, "Measuring the objectness of image windows," *TPAMI*, vol. 34, no. 11, pp. 2189–2202, 2012.

[21] Philipp Krhenbhl and Vladlen Koltun, "Geodesic object proposals," in *ECCV*, 2014.

[22] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *TPAMI*, vol. 33, no. 3, pp. 500–513, 2011.

[23] Tianfei Zhou, Yao Lu, Huijun Di, and Jian Zhang, "Video object segmentation aggregation," in *ICME*. IEEE, 2016, pp. 1–6.

[24] V. Kwatra, Mei Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," *CVPR*, 2010.

[25] P Ochs, J Malik, and T Brox, "Segmentation of moving objects by long term video analysis," *TPAMI*, vol. 36, no. 6, pp. 1–1, 2014.

[26] David Tsai, Matthew Flagg, and James M.Rehg, "Motion coherent tracking with multi-label mrf optimization," *BMVC*, 2010.