IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

Towards Segmenting Stereo videos in foreground Propagation in Cinematographic Optimization

Mr. P. Thangaraju, M.Sc., MCA., M.Phil., PGDCA.[1], Mrs. R. Nanthini., M.Sc., B.Ed.,[2]

Computer Science, Bishop Heber College, Bharathidasan University[1,2]

Abstract: A main encounter in video segmentation is that the foreground object may move rapidly in the scene at the same time its presence and shape changes over time. While pairwise potentials used in graph-based algorithms support smooth labels between neighboring (super) pixels in space and time, they proposal only a myopic view of consistency and can be misled by inter-frame optical flow errors. We propose a higher order supervoxel label consistency potential for semi-supervised foreground segmentation. Given an initial frame with manual annotation for the foreground object, our approach propagates the foreground region through time, leveraging bottom-up supervoxels to guide its evaluations towards long-range coherent regions. We endorse our approach on three challenging datasets and complete state-of-the-art results.

Keywords: video segmentation, foreground region, semi-supervised foreground segmentation, coherent regions.

I. INTRODUCTION

The process of Image mining is searching and discovering valuable data and knowledge in large volumes of data. The image mining process have some of the methods used to gather knowledge are, Image Retrieval, Data Mining, Image Processing and Artificial Intelligence. In video, the foreground object segmentation problematic consists of classifying those pixels that belong to the primary object(s) in every frame. A resulting foreground object segment is a space-time "tube" whose shape may deform as the object moves over time. The difficult has an array of potential applications, including doings recognition, object recognition, video summarization, and postproduction video editing.

Current algorithms for video segmentation can be organized by the amount of manual annotation they assume. At one exciting, there are only unsupervised methods that produce coherent space-time regions from the bottom up, without any video-specific labels [8, 12, 14, 17, 19, 21, 36, 38, and 39]. At the other extreme, there are strongly supervised communicating methods, which require a human in the loop to correct the system's errors [4, 10, 20, 25, 34, and 35]. Between either extreme, there are semi-supervised methods that need a partial amount of direct supervision—an outline of the foreground in the first frame—which is then propagated spontaneously to the rest of the video [2, 3, 10, 27, 31, 33]. We are concerned in the final semi-supervised task: the aim is to take the foreground object segmentation drawn on an initial frame and accurately propagate it to the remains of the frames. The propagation paradigm is a convincing middle ground. First, it removes ambiguity about what object is of interest, which, despite impressive advances [17, 19, 21, 39], remains an inherent drawback for unsupervised methods. Accordingly, the propagation setting can accommodate a broader class of videos, e.g., those in which the object does not move much,or shares appearance with the background. Second, propagation from just one human-labeled frame can be considerably less burdensome than human-in-theloop systems that require constant user interaction, making it a promising tool for gathering object tubes at a large scale. While heavier supervision is warranted in some domains (e.g., perfect rot scoping for graphics), in many applications it is worthwhile to trade pixel-perfection for data volume (e.g., for learning object models from video, or assisting biologists with data collection). We propose a foreground propagation method using supervoxel higher order potentials. Supervoxels—the space-time analog of spatial superpixels—provide a bottom-up volumetric segmentation that tends to preserve object boundaries [8, 12, 14, 36, 38]. To leverage their broader structure in a graph-based propagation algorithm, we supplement the usual adjacency-based collections with potentials for supervoxel-based cliques. These original cliques specify soft preferences to assign the same label (fg or bg) to superpixel nodes that occupy the same supervoxel. Whereas existing models are restricted to adjacency or flow-based links, supervoxels proposal valuable longer-term temporal constraints. We authenticate our approach on three challenging datasets, Seg Track [31], YouTube Objects [23], and Weizmann [13], and compare to state-of-the-art propagation methods. Our approach outperforms current techniques overall, with particular benefit when foreground and background appearance similar, inter-frame motion is high, or the aim changes shape between frames.

II. RELATED WORK

Unsupervised video segmentation Unsupervised video segmentation methods proficiently extract coherent collections of voxels. Hierarchical graph-based approaches use arrival and flow to group voxels [14, 38], while others group super

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

pixels using spectral clustering [12] or novel tracking techniques [5, 32]. Distinct from the region-based methods, tracking methods use point trajectories to detect cohesive moving object parts [7, 18]. Any such bottom-up method tends to reservation object boundaries, but "over segment" them into numerous parts. As such, they are notintended as object segmentations; rather, they make available a mid-level space-time grouping valuable for downstream tasks.

2.1. Interactive Video Segmentation

At the other end of the spectrum are interactive methods that assume a human annotator is in the loop to correct the algorithm's mistakes [4, 20, 25, 35], any by monitoring the results closely, or by responding to active queries by the system [10, 33, 34]. While such intensive supervision is warranted for some submissions, particularly in graphics [4, 20, 25, 35], it may be overkill for others. We focus on the foreground propagation problem, which assumes supervision in the form of a single labeled frame. Regardless, developments due to our super voxel idea could also benefit the interactive methods, some of which start with a similar MRF graph structure [10, 20, 25, 33] (but lack the proposed higher order potentials).

2.2. Semi-supervised Foreground Propagation

Most relevant to our work are methods that accept a frame labeled physically with the foreground region and propagate it to the remaining clip [3, 10, 27, 31, and 33]. While different in their optimization strategies, greatest prior approaches use the core MRF structure designated above, with i) unary potentials determined by the labeled foreground's appearance/motion and ii) pairwise potentials determined by nodes' temporal or spatial adjacency. Pixel-based graphs can maintain actual fine boundaries, but suffer from high computational cost and noisy temporal links due to unreliable flow [3, 33]. Superpixelbased graphs form nodes by segmenting each frame independently [10, 27, 31]. Associated to their pixel counterparts, they are much more efficient, less disposed to to optical flow drift, and can estimate neighbors' similarities additional robustly due to their better spatial extent. Nonetheless, their use of per-frame segments and frame-to-frame flow links limits them to short range communications. In contrast, our key idea is to impose a super voxel potential to encourage consistent labels across broad spatial-temporal regions.

III. METHODOLOGY

The input to our approach is a video clip and one labeled frame in which an annotator has outlined the foreground object of interest. The output is a spacetime segmentation that propagates the foreground (fg) or background (bg) label to every pixel in every frame. While the foreground object must be present in the labeled frame, it may leave and re-enter the scene at other times.

3.1 Motivation and Approach Overview

Our highest objective is to define a space-time graph and energy function that respect the "big picture" of how objects move and evolve throughout the clip. Key to our idea is the use of super voxels. Super voxels are space-time regions computed with a bottom-up unsupervised video segmentation algorithm [14, 36, and 38]. They typically over segment—meaning that objects may be parceled into many super voxels—but the object boundaries remain visible among the super voxel boundaries. They vary in shape and size, and will typically be larger and longer for content more uniform in its color or motion. Though a given object part's super voxel is unlikely to remain stable through the entire length of a video, thanks to temporal endurance, it will often persist for a series of frames. For example, in Figure 1, we see a number of larger super voxels remain steady in early frames, then some split/merge as the dog's pose changes, then a revised set again stabilizes for the latter chunk of frames. As we will see below, our approach exploits the partial stability of the super voxels but also recognizes their noisy imperfections.

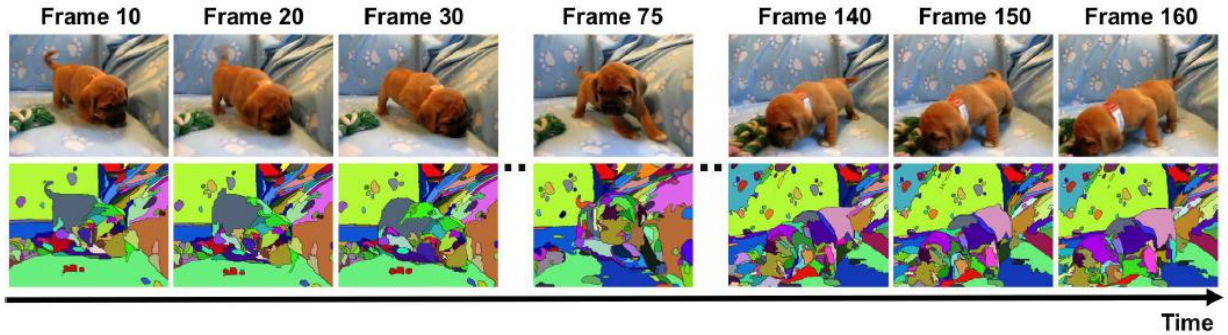

Fig. 1: Example super voxels, using [14]. Unique colors are unique super voxels, and repeated colors in adjacent frames refer to the same super voxel. Best viewed in color.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

A naive simplification to video would build a graph with super voxels as nodes, connecting adjacent super voxels in space and time. The problematic is the irregular shape of super voxels—and their widely varying temporal extents—lead to brittle graphs. As we will see in the results, the pairwise potentials in such amethod lead to frequent bleeding across object boundaries.

Instead, we propose to leverage super voxels in two ways. First, for each super voxel, we project it into all of its child frames to obtain spatial super pixel nodes. These nodes have sufficient spatial extent to calculate rich visual features. Plus, compared to standard super pixel nodes computed independently per frame [3, 8, 10, 12, 25, 27, 31], they benefit from the broader perspective provided by the hierarchical space-time segment that generates the super voxels. For example, optical flow similarity of voxels on the dog's textured collar may preserve it as one node, whereas per-frame segments may break it into many. Secondly, we leverage super voxels as a higher-order potential. Augmenting the usual unary and pairwise terms, we enforce a soft label consistency constraint among nodes originating from the same super voxel. Again, this offers broader context to the propagation engine.

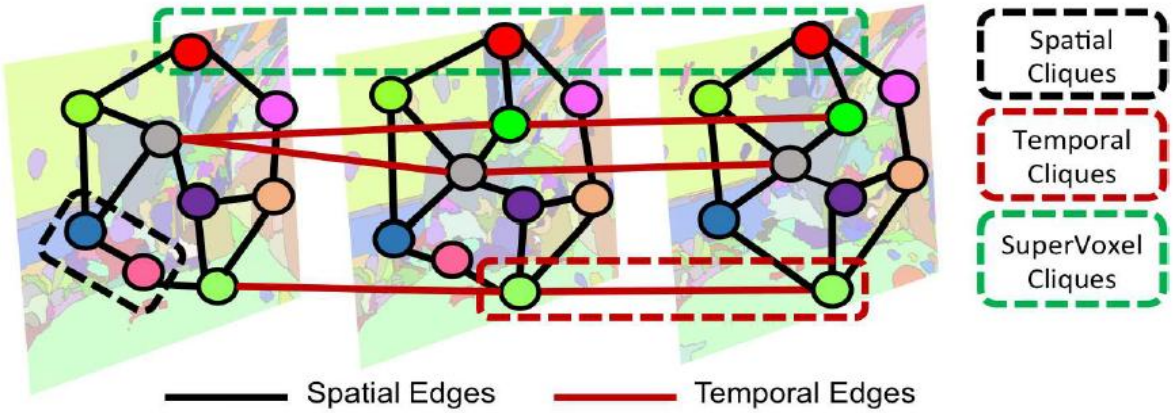

Fig. 2: Proposed spatio -temporal graph. Nodes are super pixels (projected from super voxels) in every frame. Spatial edges exist if the super pixels have boundary overlap (black); temporal edges are computed using optical flow (red). Higher order cliques are defined by super voxel membership (dotted green). For legibility, only a small subset of nodes and connections are depicted. Best viewed in color.

3.2 Space-time MRF Graph Structure

We first formally define the proposed spatio-temporal Markov Random Field (MRF) graph structure G consisting of nodes X and edges E. Let X = {Xt} T t=1 be the set of superpixels2 over the entire video volume, where T refers to the number of frames in the video. Xt is a subset of X and contains super pixels belonging only to the t-th frame. Therefore each Xt is a collection of super pixel nodes {x i t} Kti=1, where Kt is the number of super pixels in the t-th frame.

We subordinate a random variable y i t ∈ {+1, −1} with every node to represent the label it may take, which can be either object (+1) or background (-1). Our goal is to obtain a labeling Y = {Yt} T t=1 over the entire video. Here, Yt = {y i t} Kti=1 represents the labels of super pixels belonging only to the t-th frame. Below, (t, i) indexes a super pixel node at position i and time t.

We define an edge set E = {Es, Et} for the video. Es is the set of spatial edges between super pixel nodes. A spatial edge exists between a pair of super pixel nodes (x it , x j t) in a given frame if their boundaries overlap (black lines in Figure 2). Et is the set of temporal edges. A temporal edge exists between a pair of super pixels (x it , x j t+1) in adjacent frames if any pixel from x i t tracks into x j t+1 using optical flow (red lines in Figure 2). We use the algorithm of [6] to calculate dense flow between following frames. Let [(t, i),(t 0 , j)] index an edge between two nodes. For spatial edges, t 0 = t; for temporal edges, t 0 = t + 1.

Finally we use S to denote the set of super voxels. Each element v ∈ S represents a higher order clique (one is shown with a green dashed box in Fig. 2)over all the super pixel nodes which are a part of that super voxel. Let yv denote the set of labels assigned to the super pixel nodes belonging to the super voxel v. For each super pixel node x it , we compute two image features using all its pixels: 1) an RGB color histogram with 33 bins (11 bins per channel), and 2) a histogram of optical flow, which bins the flow orientations into 9 uniform bins. We concatenate the two descriptors and compute the visual dissimilarity between two super pixels D(x it , x j t 0) as the Euclidean distance in this feature space.

3.3. Energy Minimization and Parameters

The energy function defined in Eqn. 1 can be capablyreduced using the α- expansion algorithm [16]. The optimal labeling corresponding to the minimum energy yields our initial fg-bg estimate. We iteratively refine that output by

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

reestimating the appearance model—using only the most self-assured samples based on the current unary potentials—then answering the energy function again. We perform three such iterations to obtain the final output.
The only three parameters that must be set are λapp and λloc, the weights in the appearance potential, and the truncation parameter Q. We determined reasonable values (λapp = 100, λloc = 40, Q = 0.2 |yv|) by visual inspection of a couple outputs, then fixed them for all videos and datasets. (This is minimal effort for a user of the system. It could also be done with cross-validation, when adequate pixel-level ground truth is available for training.) The remainingparameters βu, βp, and βh, which scale the visual dissimilarity for the unary, pairwise, and higher order potentials, respectively, are all set automatically as the inverse of the mean of all individual distance terms.

Fig. 3: Example results on SegTrack. Best viewed in color.

IV. RESULTS

Datasets and metrics: We evaluate on 3 publicly available datasets: SegTrack [31], YouTube-Objects [24], and Weizmann [13]. For SegTrack and YouTube, the true object region in the first frame is supplied to all methods. We use standard evaluation metrics: average pixel label error and intersection-over-union overlap.
Methods compared: We compare to five state-of-the-art methods: four for semi-supervised foreground label propagation [9, 10, 31, 33], plus the state-ofthe-art higher order potential method of [8]. Note that unsupervised multiplehypothesis methods [17, 19, 21, 39] are not comparable in this semi-supervised single-hypothesis setting. We also test the following baselines:
– **SVX-MRF:** an MRF comprised of super voxel nodes. The unary potentials are initialized through the labeled frame, and the smoothness terms are defined using spatio-temporal adjacency between super voxels. It highlights the importance of the design choices in the proposed graph structure.
– **SVX-Prop:** a simple propagation scheme using super voxels. Starting from the labeled frame, the propagation of foreground labels progresses through temporally linked (using optical flow) super voxels. It illustrates that it's non-trivial to directly extract foreground from super voxels.
– **PF-MRF:** the existing algorithm of [33], which uses a pixel-flow (PF) MRF for propagation. This is the only video segmentation propagation algorithm with publicly available code.4 Note that the authors also propose a method to actively select frames for labeling, which we do not employ here.
– **Ours w/o HOP:** a simplified version of our method that lacks higher order potentials (Eqn. 7), to isolate the impact of super voxel label consistency.

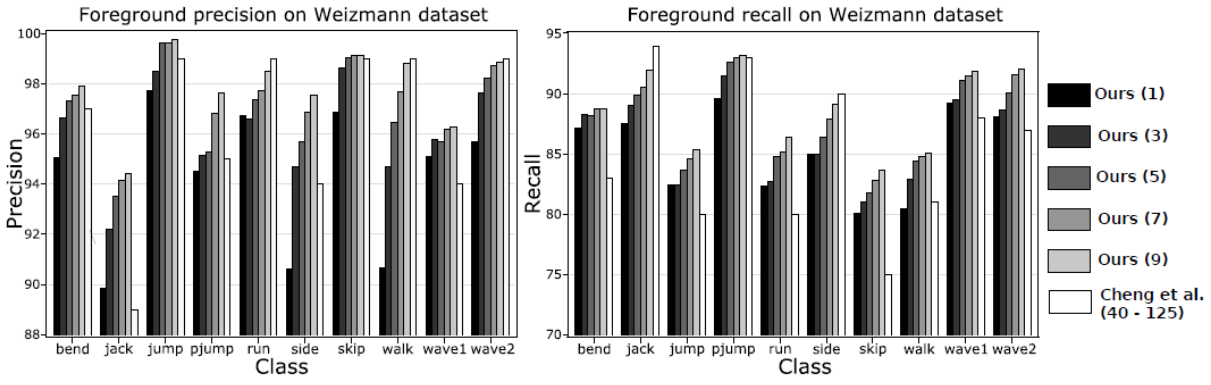

Fig. 4: Foreground precision (left) and recall (right) on Weizmann. Legend shows number of labeled frames used per result (1 to 9 for our method, 40-125 for [8]).

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

Figure 4 shows the results in terms of foreground precision and recall, following [8]. Whereas we output a single fg-bg estimate (2 segments), the method of [8] outputs an over segmentation with about 25 segments per video. Thus, the authors use the ground truth on each frame to map their outputs to fg and bg labels, based on mainstream overlap; this is equivalent to obtaining on the order of 25 manual clicks per frame to label the output. In contrast, our propagation method uses just 1 labeled frame to generate a complete fg-bg segmentation. Therefore, we show our results for increasing numbers of labeled frames, spread uniformly through the sequence. This requires a multi-frame extension of our method—namely, we take the appearance model Gyt from the labeled frame nearest to t, and re-initialize the spatial prior L i t (y i t) at every labeled frame.

V. CONCLUSIONS

We introduced a new semi-supervised method to propagate object regions in video. Outstanding to its higher order super voxel potential, it outperforms the state-of the-art on over 200 classifications from 3 distinct datasets. In future work, we plan to extend the idea to accommodate numerous and/or hierarchical super voxel inputs, and to explore shape descriptors to augment the foreground representations.

REFERENCES

1. Ahuja, N., Todorovic, S.: Connected segmentation tree: a joint representation of region layout and hierarchy. In: CVPR (2008)
2. Ali, K., Hasler, D., Fleuret, F.: Flowboost: Appearance learning from sparsely annotated video. In: CVPR (2011)
3. Badrinarayanan, V., Galasso, F., Cipolla, R.: Label propagation in video sequences. In: CVPR (2010)
4. Bai, X., Wang, J., Simons, D., Sapiro, G.: Video snapcut: Robust video object cutout using localized classifiers. In: SIGGRAPH (2009)
5. Brendel, W., Todorovic, S.: Video object segmentation by tracking regions. In: ICCV (2009)
6. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. PAMI 33(3), 500–513 (2011)
7. Brox, T., Malik, J.: Object Segmentation by Long Term Analysis of Point Trajectories. In: ECCV (2010)
8. Cheng, H.T., Ahuja, N.: Exploiting nonlocal spatiotemporal structure for video segmentation. In: CVPR (2012)
9. Chockalingam, P., Pradeep, S.N., Birchfield, S.: Adaptive fragments-based tracking of non-rigid objects using level sets. In: ICCV (2009)
10. Fathi, A., Balcan, M., Ren, X., Rehg, J.: Combining self training and active learning for video segmentation. In: BMVC (2011)
11. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. IJCV 59(2) (2004)
12. Galasso, F., Cipolla, R., Schiele, B.: Video segmentation with superpixels. In: ACCV (2012)
13. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. PAMI 29(12), 2247–2253 (2007)
14. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph based video segmentation. In: CVPR (2010)
15. Hartmann, G., Grundmann, M., Hoffman, J., Tsai, D., Kwatra, V., Madani, O., Vijayanarasimhan, S., Essa, I., Rehg, J., Sukthankar, R.: Weakly supervised learning of object segmentations from web-scale video. In: ECCV Workshop on Vision in Web-Scale Media (2012)
16. Kohli, P., Ladicky, L., Torr, P.H.S.: Robust higher order potentials for enforcing label consistency. In: CVPR (2008)
17. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV (2011)
18. Lezama, J., Alahari, K., Sivic, J., Laptev, I.: Track to the future: Spatio-temporal video segmentation with long-range motion cues. In: CVPR (2011)
19. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video Segmentation by Tracking Many Figure-Ground Segments. In: ICCV (2013)
20. Li, Y., Sun, J., Shum, H.Y.: Video object cut and paste. ACM Trans. Graph. 24(3), 595–600 (2005)
21. Ma, T., Latecki, L.: Maximum weight cliques with mutex constraints for video object segmentation. In: CVPR (2012)
22. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV (2013)
23. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: CVPR (2012).
24. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3282–3289. Ieee (Jun 2012), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6248065
25. Price, B.L., Morse, B.S., Cohen, S.: Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues. In: ICCV (2009)
26. Ren, X., Malik, J.: Learning a classification model for segmentation. In: ICCV (2003)
27. Ren, X., Malik, J.: Tracking as repeated figure/ground segmentation. In: CVPR (2007)
28. Rubio, J.C., Serrat, J., Lopez, A.M.: Video co-segmentation. In: ACCV (2012)
29. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV (2006)
30. Tang, K., Sukthankar, R., Yagnik, J., Fei-Fei, L.: Discriminative segment annotation in weakly labeled video. In: CVPR (2013)
31. Tsai, D., Flagg, M., Rehg, J.: Motion coherent tracking with multi-label mrf optimization. In: BMVC (2010)
32. Vazquez-Reina, A., Avidan, S., Pfister, H., Miller, E.: Multiple hypothesis video segmentation from superpixel flows. In: ECCV (2010)
33. Vijayanarasimhan, S., Grauman, K.: Active frame selection for label propagation in videos. In: ECCV (2012)
34. Vondrick, C., Ramanan, D.: Video annotation and tracking with active learning. In: NIPS (2011)
35. Wang, J., Bhat, P., Colburn, A., Agrawala, M., Cohen, M.F.: Interactive video cutout. ACM Trans. Graph. 24(3), 585–594 (2005)
36. Xu, C., Corso, J.: Evaluation of super-voxel methods for early video processing. In: CVPR (2012)
37. Xu, C., Whitt, S., Corso, J.: Flattening supervoxel hierarchies by the uniform entropy slice. In: ICCV (2013)
38. Xu, C., Xiong, C., Corso, J.J.: Streaming Hierarchical Video Segmentation. In: ECCV (2012)
39. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: CVPR (2013)