Learning to Segment Moving Objects in Videos

Katerina Fragkiadaki[1], Pablo Arbeláez [2], Panna Felsen[1] Jitendra Malik[1]
[1]University of California, Berkeley [2]Universidad de los Andes, Colombia

We present a method that segments moving objects in videos by ranking spatio-temporal region proposals according to "moving objectness"; how likely they are to contain a moving object. Region proposal generation and ranking using an object detector is currently the dominant paradigm for object detection in the static image domain [7]. It has shown excellent performance against sliding window classifiers or Markov Random Field based pixel classification that cannot distinguish closeby instances of the same object class [4]. In this paper, we propose a similar paradigm for detecting moving objects in videos and present large quantitative advances over previous multiscale segmentation and trajectory clustering methods.

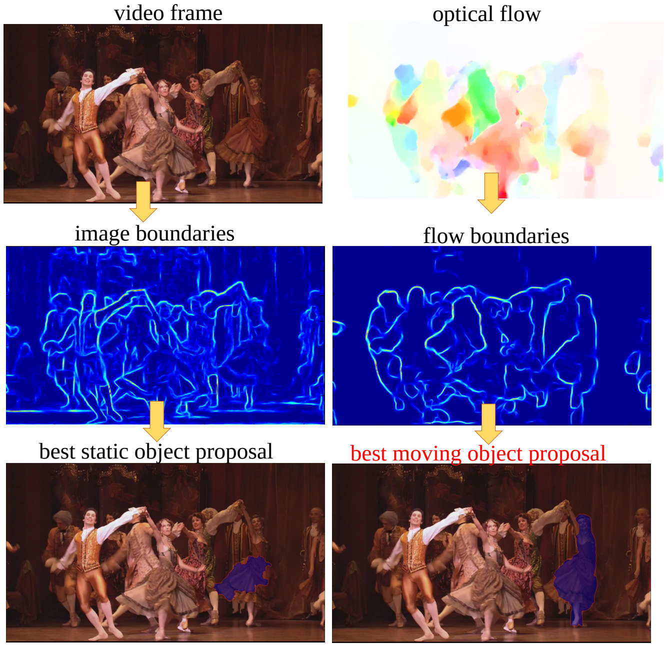

Figure 1: **Per frame moving object proposals.** Static segment proposals fail to capture the dancer as a whole due to internal clothing contours. Flow boundaries suffer less from albedo or shading edges in object interiors. Segmentation on them correctly delineates the dancer.

In each video frame, we compute segment proposals using multiple figure-ground segmentations on per frame motion boundaries. We extract motion boundaries by applying the learning based boundary detector of [3] on the magnitude of optical flow. The extracted boundaries establish pixel affinities for multiple figure-ground segmentations [9] that generate a pool of segment proposals; we call them per frame Moving Object Proposals (MOPs). Our per frame MOPs increase the object detection rate up to 7% over previous state-of-the-art static proposals and demonstrate the value of motion for object detection in videos. Objects, however, are not constantly in motion. At frames when they are static, there are no optical flow boundaries and MOPs miss them. Thus, we extend MOPs to spatio-temporal tubes using random walkers on dense point trajectory motion affinities.

We rank per frame segment and spatio-temporal tube proposals with a Moving Objectness Detector (MOD) that learns to detect moving objects from a set of training examples. Our MOD has a dual-pathway CNN architecture that operates on both RGB and flow fields. It outperforms hand-coded center-surround saliency and other competitive multilayer objectness baselines [8].

Our method bridges the gap between motion segmentation and tracking methods. Previous motion segmenters [11, 12] operate "bottom-up", they exploit color or motion cues without using a training set of objects. Previous trackers [1, 6] use object detectors (e.g., car or pedestrian detector) to cast attention to the relevant parts of the scene. We do use a training set for learning the concept of a moving object, yet remain agnostic to the exact object classes present in the video.

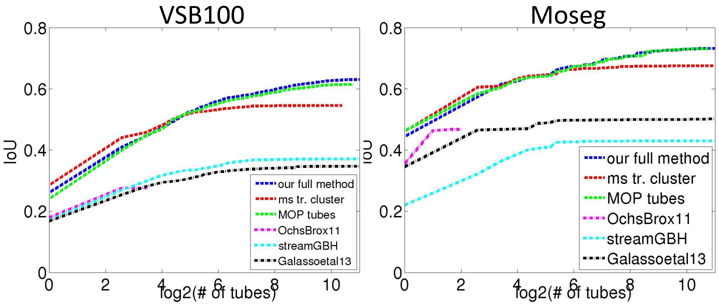

Figure 2: Cols 1,2: **Motion segmentation results in VSB100** (col. 1) **and Moseg** (col. 2). Our method outperforms previous supervoxel and trajectory clustering approaches.

We test our method on the two largest video segmentation benchmarks currently available, Moseg [2] and VSB100 [5], and outperform competing approaches of [5, 10, 12]. 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.eecs.berkeley.edu/~katef/`.

[1] M.D. 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] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*. 2010.

[3] Piotr Dollár and C. Lawrence Zitnick. Structured forests for fast edge detection. In *ICCV*, 2013.

[4] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 35, 2013.

[5] Fabio Galasso, Naveen Shankar Nagaraja, Tatiana Jimenez Cardenas, Thomas Brox, and Bernt Schiele. A unified video segmentation benchmark: Annotation, metrics and analysis. In *ICCV*, 2013.

[6] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *TPAMI*, 2011.

[7] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[8] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *NIPS*. 2014.

[9] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *ECCV*. 2014.

[10] Peter Ochs and Thomas Brox. Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In *ICCV*, 2011.

[11] P.Ochs and T.Brox. Higher order motion models and spectral clustering. In *CVPR*, 2012.

[12] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *CVPR*, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.