Robust Video Segment Proposals with Painless Occlusion Handling

Zhengyang Wu,[1] Fuxin Li,[1] Rahul Sukthankar,[2] James M. Rehg,[1]
[1]Georgia Institute of Technology [2]Google Research

In this paper, we propose a video segment proposal framework with a minimal set of assumptions. Our approach generates a pool of video segment proposals starting from any frame, identifies both moving and still objects as well as parts of an object, handles both partial and complete occlusions, and is free of any specific motion model (e.g., linear, smooth, etc.). It is our belief that such a proposal method could provide the necessary preprocessing for many subsequent algorithms.

Our approach trains long-term holistic appearance models on image segment proposals based on least squares, similar to [5]. Thousands of appearance models are efficiently trained on a pool of image segments, and tracked segments are gradually filtered in subsequent frames using appearance and motion constraints. There are two major differences: One is that we do not require the restrictive assumption as in [5] that all segments must start from the first frame. This is implemented via a series of forward and backtracking moves within the algorithm, without greatly increasing the time and space complexities. The second difference is that we handle complete occlusion, by automatically detecting the onset of complete occlusions, maintaining persistent identities of the occluded segments, and detecting them when they re-enter the scene. Importantly, occlusion handling is implemented within the same least squares framework and occluded tracks receive extra negative training examples on each frame without needing to perform any computation (dubbed as **free addition** moves).

1 Methodology

Our system is built on the flexible least squares tracker utilized in [5], which adopts on the following regularized least squares formulation:

$$\min_{\mathbf{W}} \|\mathbf{XW}-\mathbf{V}\|_F^2 + \lambda\|\mathbf{W}\|_{\mathrm{F}}^2, \tag{1}$$

where the goal is to recover the $d \times t$ weight matrix $\mathbf{W}$, given $n \times d$ input matrix $\mathbf{X}$ and $n \times t$ output matrix $\mathbf{V}$. n is the number of examples, d the dimensionality t the number of distinct targets and $\|\mathbf{W}\|_{\mathrm{F}}$ a Frobenius norm regularization. The solution of the least squares is given by the linear system:

$$(\mathbf{H}+\lambda\mathbf{I})\mathbf{W} = \mathbf{C}. \tag{2}$$

The pair $(\mathbf{H},\mathbf{C})$ is the sufficient statistics of the model. We denote $L = (\mathbf{H},\mathbf{C})$ a **least squares object** (LSO). Note each column in $\mathbf{C}$ corresponds to a distinct target (with the output a column in $\mathbf{V}$).

Starting from image segment proposals, we use appearance features as $\mathbf{X}$, the pairwise overlap matrix as $\mathbf{V}$ to start the training of an online tracker with each proposal as a target. In subsequent frames, we find the matching segment to each track, and update the model accordingly with appearance features from the new frame, and newly computed overlaps between each segment proposal and the matching segment of each track.

Since the computation and inversion of $\mathbf{H}$ do not involve operations on the targets, one can use the same $\mathbf{H}$ for many targets without incurring much additional computation. Suppose all segment tracks use the same set of training examples and only differ in their target scores, one can use the same $\mathbf{H}$ for lots of tracks $(1,000+)$ which can include both visible and occluded ones. That is the basis for **free addition**, since when the visible tracks are updated, $\mathbf{H}$ is updated and occluded tracks receive the updates for free.

Some segment tracks do not start on the first frame, hence their appearance models cannot be trained use the same set of training examples. We generate an LSO starting at each frame and merge LSOs after they have lasted several frames so that we maintain a low count of LSOs while spanning segments that start/end at all frames.

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

Table 1: VSB-100 results for different algorithms. S_v denotes the overlap score averaged per video, S_o denotes the overlap averaged per object, see the paper for details of the metric. RIGOR upp. bnd. represents the accuracy of the raw proposals, and is the theoretical best for this framework.

	S_v	S_o	# Segs
Backtrack & Occlusion	50.12	45.81	324
+Post-processing	**56.13**	**51.92**	324
Original Li et al. [5]	44.81	41.25	46
Grundmann et al. [3]	45.28	42.94	737
Galasso et al. [2]	45.32	42.54	2850
RIGOR upp. bnd [4]	69.63	65.89	1420/frame

2 Occlusion Detection and Cross Video Testing

In each frame we detect potential occluded tracks, which are tracked using free addition in the subsequent frames until they return in a future frame. For occlusion detection, we build a linear regression model on the segment size of each track and mark tracks as occluded if their predicted segment size in next frame falls below a threshold. After a track becomes occluded, we track it using free addition. At each later frame, those occluded tracks are tested to check whether they have returned to the video.

Once we have gained a reliable appearance model, we can test the model either on the same video to recover lost frames, or on another candidate video to retrieve object in the same class across video.

3 Results

We test our video segment proposal framework on the challenging VSB-100 dataset from [1]. This dataset includes pixel-level ground truth and several videos with extreme occlusions. Each labeled frame in this dataset contains annotations by four different annotators. To address the disagreement among annotators, we propose a novel metric using maximal clique on a similarity graph among ground truth segments. Segments from different annotators are in the same clique if their pairwise similarity (overlap) is above a threshold. Then we test our proposal against these cliques by taking the maximal overlap between our proposal and segments in the same clique (object). Our algorithm significantly outperform competitors in Table 1.

The algorithm takes about 55 seconds per frame, but half of the time is spent generating the features which can be greatly optimized. It can also be sped-up significantly with reduced feature dimensionality and approximate solvers for least squares. The cross-video segmentation results in the paper also shows interesting trends of the tracking capability of colorSIFT features as compared with convolutional neural network features.

References

[1] 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.

[2] Fabio Galasso, Margret Keuper, Thomas Brox, and Bernt Schiele. Spectral graph reduction for efficient image and streaming video segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.

[3] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. Efficient hierarchical graph based video segmentation. In *CVPR*, 2010.

[4] Ahmad Humayun, Fuxin Li, and James M. Rehg. RIGOR: Reusing Inference in Graph Cuts for generating Object Regions. In *CVPR*, 2014.

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