Supplementary Material: Semantic Co-segmentation in Videos

Yi-Hsuan Tsai[1⋆], Guangyu Zhong[12⋆], Ming-Hsuan Yang[1]

[1]UC Merced, [2]Dalian University of Technology
{ytsai2,gzhong,mhyang}@ucmerced.edu

1 Analysis of Tracklet Co-selection

We analyze the proposed tracklet co-selection method based on the setting without knowing any prior knowledge on the Youtube-Objects dataset. We first evaluate the importance of facility location $\mathcal{F}(\mathcal{A})$ and unary terms $\mathcal{U}(\mathcal{A})$ in the submodular function. We show both the intersection-over-union (overlap) ratio for semantic segmentation and the average precision (AP) for classification in Table 1 under the same threshold (i.e., 0.85 as used in the manuscript). With only the facility location term that measures the object similarity, the results are less accurate caused by noisy tracklets, while the unary term can ensure the quality of selected tracklets, and hence produce better results by combining two terms.

In Table 2, we show the average overlap ratio over all categories for semantic segmentation with different thresholds applying on re-ranked tracklets. Since a low threshold may result in selecting more tracklets and including more noisy ones, we also report the average F-measure for object classification. Note that we achieve the best result for both segmentation and classification with the threshold 0.75.

Table 1. Segmentation and classification results on the Youtube-Objects dataset with an without the unary term in the submodular function.

	Overlap ratio		Average precision	
Category	$\mathcal{F}(\mathcal{A})$	$\mathcal{F}(\mathcal{A})+\mathcal{U}(\mathcal{A})$	$\mathcal{F}(\mathcal{A})$	$\mathcal{F}(\mathcal{A})+\mathcal{U}(\mathcal{A})$
aeroplane	**69.3**	**69.3**	**95.8**	**95.8**
bird	53.3	**76.0**	78.7	**97.6**
boat	**55.0**	53.5	**100**	**100**
car	**70.4**	**70.4**	82.4	**85.7**
cat	60.3	**66.8**	72.4	**76.9**
cow	**53.9**	49.0	86.4	**93.1**
dog	**50.3**	47.5	81.2	**84.5**
horse	45.8	**55.7**	**42.5**	**42.5**
motorbike	**43.2**	39.5	**89.7**	**89.7**
train	53.3	**53.4**	82.9	**86.6**
Mean	55.5	**58.1**	81.2	**85.3**

⋆ Both authors contribute equally to this work.

Table 2. Segmentation and classification results on the Youtube-Objects dataset with different thresholds for tracklet co-selection.

Threshold	0.35	0.45	0.55	0.65	0.75	0.85
Overlap ratio	56.8	56.9	58.0	58.5	**59.6**	58.1
F-measure	79.4	79.1	80.0	82.4	**84.2**	83.5

2 Runtime Performance

With the MATLAB implementation, the runtime performance on the Youtube-Objects dataset is shown in Table 3. All the timings are measured on a PC with 3.4 GHz Intel i7 CPU and 32 GB memory. In the feature extraction step, it includes all the needed features in the following three steps. To extract FCN features and outputs, we use a Titan X GPU with 12 GB memory (0.57 second per frame). For optical flow, we use the method of [1], which takes 3.2 seconds per pair of frames on average. Note that during tracklet co-selection, graphs can be solved in parallel.

Table 3. Runtime performance on the Youtube-Objects dataset.

Stage	Time (second)
Feature extraction (per frame)	0.82
Segment clustering (per frame)	0.13
Segment tracking (per frame)	5.74
Tracklet co-selection (per graph)	3.78

3 Segmentation Results

We show more qualitative results and comparisons on the Youtube-Objects (Fig. 1 and 2), MOViCS (Fig. 3 and 4) and Safari (Fig. 5 and 6) datasets. The results show that our method is able to track and segment (multiple) objects under challenges such as occlusions, fast movements, deformed shapes, scale changes and cluttered backgrounds. In Fig. 2, we also show that our method is capable to segment different semantic objects in one video. More results are provided in the video.

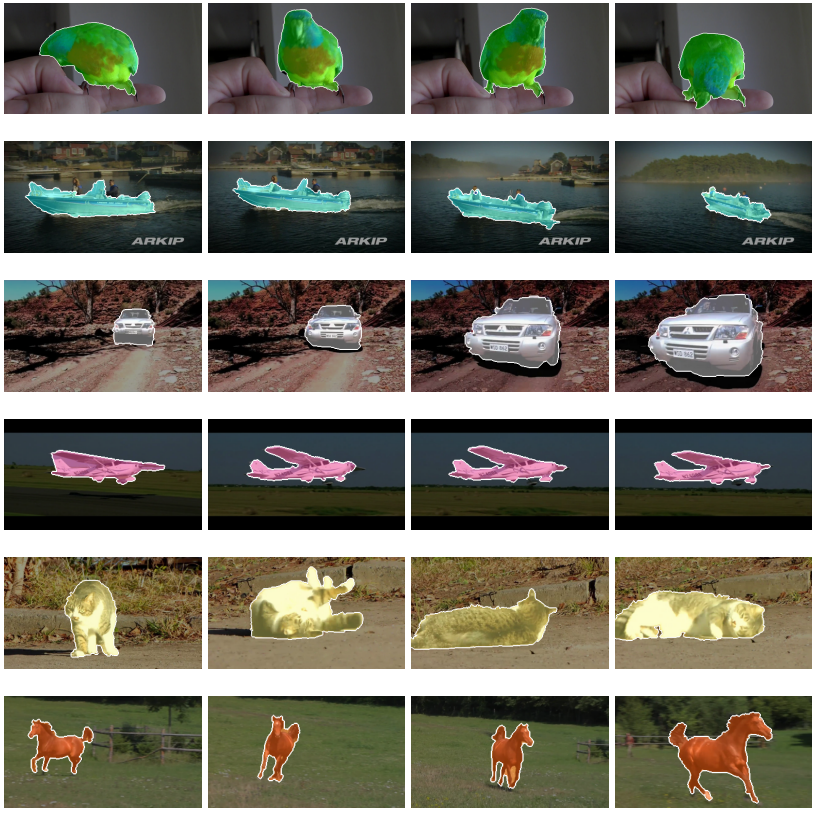
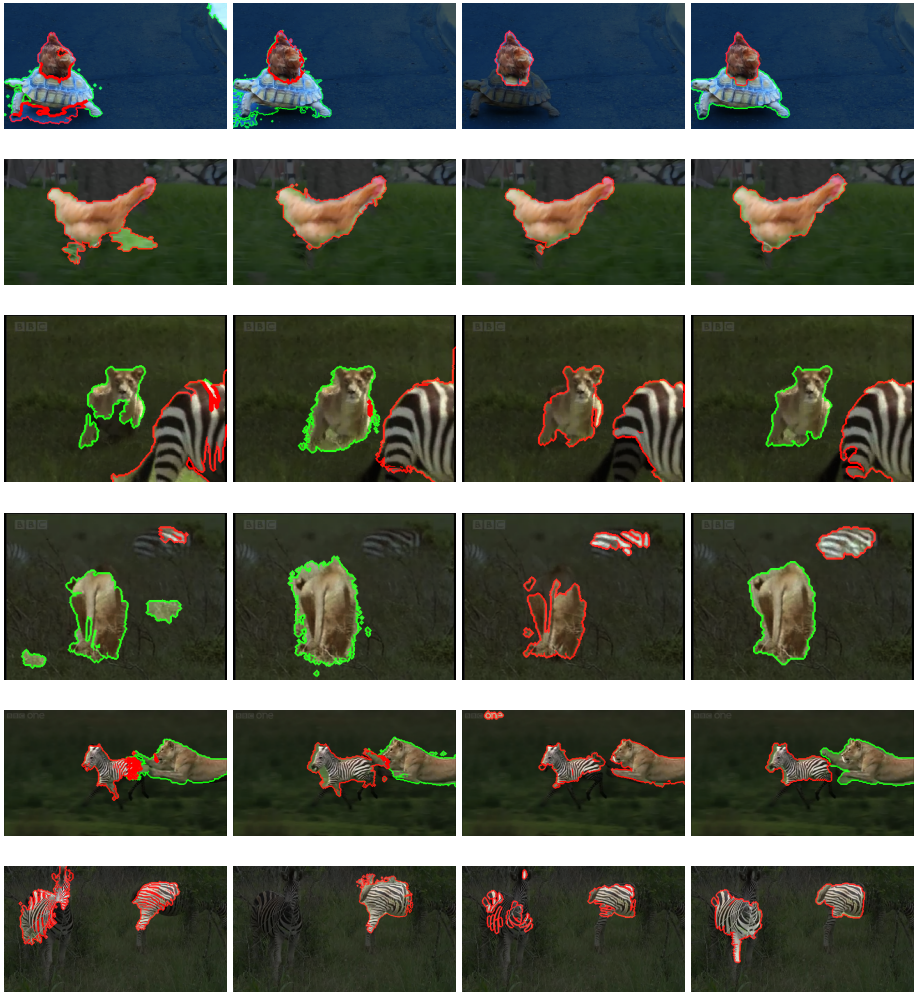


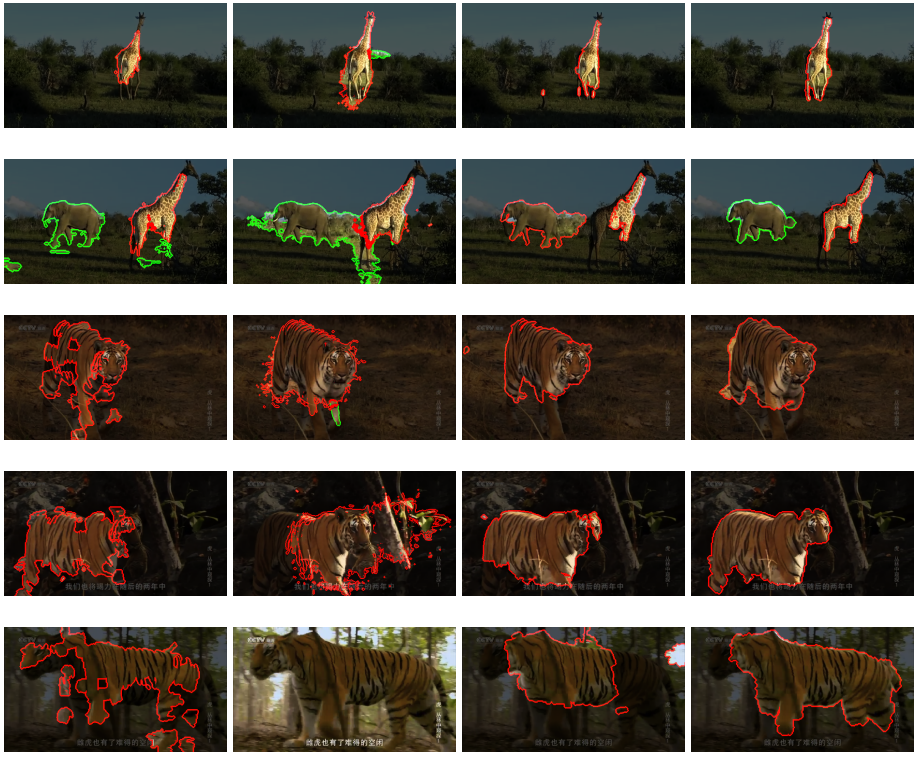
Fig. 1. Example results for semantic co-segmentation on the Youtube-Objects dataset (without knowing object categories). The colors overlapping on the objects indicate different semantic labels. The results show that our method is able to track and segment objects under challenges such as fast movements, deformed shapes, scale changes and cluttered backgrounds. Best viewed in color with enlarged images.

Fig. 2. Example results for semantic co-segmentation on the Youtube-Objects dataset (without knowing object categories). The colors overlapping on the objects indicate different semantic labels. The results show that our method is able to track and segment (multiple) objects under various challenges. Note that multiple objects with different semantic categories can be segmented in one video (see the last two rows). Best viewed in color with enlarged images.



VCS [2] RMWC [3] Baseline [4] Our results

Fig. 3. Example results for object co-segmentation on the MOViCS dataset. Segmentation outputs are indicated as colored contours, where each color represents an instance.



VCS [2] RMWC [3] Baseline [4] Our results

Fig. 4. Example results for object co-segmentation on the MOViCS dataset. Segmentation outputs are indicated as colored contours, where each color represents an instance.

RMWC [3] Baseline [4] Our results

Fig. 5. Example results for object co-segmentation on the Safari dataset. Segmentation outputs are indicated as colored contours, where each color represents an instance.

Fig. 6. Example results for object co-segmentation on the Safari dataset. Segmentation outputs are indicated as colored contours, where each color represents an instance.

References

1. Wulff, J., Black, M.J.: Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In: CVPR. (2015) 2
2. Chiu, W.C., Fritz, M.: Multi-class video co-segmentation with a generative multi-video model. In: CVPR. (2013) 5, 6
3. Zhang, D., Javed, O., Shah, M.: Video object co-segmentation by regulated maximum weight cliques. In: ECCV. (2014) 5, 6, 7, 8
4. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV. (2013) 5, 6, 7, 8