One-Shot Video Object Segmentation with Iterative Online Fine-Tuning

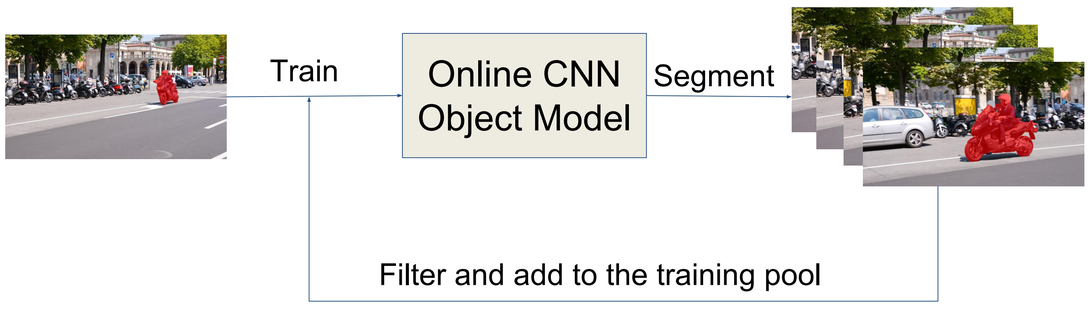
Amos Newswanger and Chenliang Xu
anewswan@u.rochester.edu, chenliang.xu@rochester.edu
Department of Computer Science, University of Rochester

Introduction

Motivation: Semi-supervised video object segmentation has recently attracted much attention. The OSVOS[1] model achieves state of the art results on the DAVIS 2016 dataset[2] by first fine-turning a CNN model on the first pre-segmented frame of a video and then independently segmenting the rest of the frames in that video. However, the model lacks the ability to learn new information about the object as it evolves throughout the video displaying features that were not present in the first frame. In the sequence shown below (segmented by OSVOS), the scooter shows relatively few features in the first frame, and as a result, the OSVOS model fails to segment it correctly in the later frames. **The segmentations in later frames can provide additional information about the object.**

Objective: We propose an iterative online training method whereby the model is fine-tuned on the first frame, segments several consecutive frames independently, and then gets updated on its own output segmentation. This process is repeated until all frames of a video are segmented. To segment multiple similar objects in a video, we use an object tracker to filter the output of the individually trained CNN object models before being used for iterative fine-turning. This reduces the possibility for error propagation, and helps the model increase its discriminative power as it is being iteratively fine-tuned. Our method shows improvement over the standard OSVOS model on both DAVIS 2016 and 2017 datasets.

Method



Iterative Online Training We first fine-tune the CNN object model on the first frame that comes with ground-truth object mask. We then use this model to segment the a few following consecutive frames. These segmented frames are added back into the training set, and the object model is further updated by fine-tuning on its own output. This process is repeated until all frames in a video are segmented.

Selective Training In some cases, this causes errors made early on in the process to be propagated forward. To mitigate this issue, we filter the output before being used for further training by 1) taking the largest blob in the segmentation or 2) using object tracker to filter out regions outside of the tracked object bounding box.

Multiple Objects We handled the multi-object masks in DAVIS 2017 dataset by first splitting them into separate binary masks, and then running our method independently on each object, and finally, merging the results by taking the maximum over all the masks with boundaries snapped to contours. For objects with similar appearances, we use the OpenCV KCF tracker to filter the output and assist the selective training, which helps the model differentiate between similar objects.

References:
[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-Shot Video Object Segmentation. In CVPR, 2017.
[2] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In CVPR, 2016.

Experiments

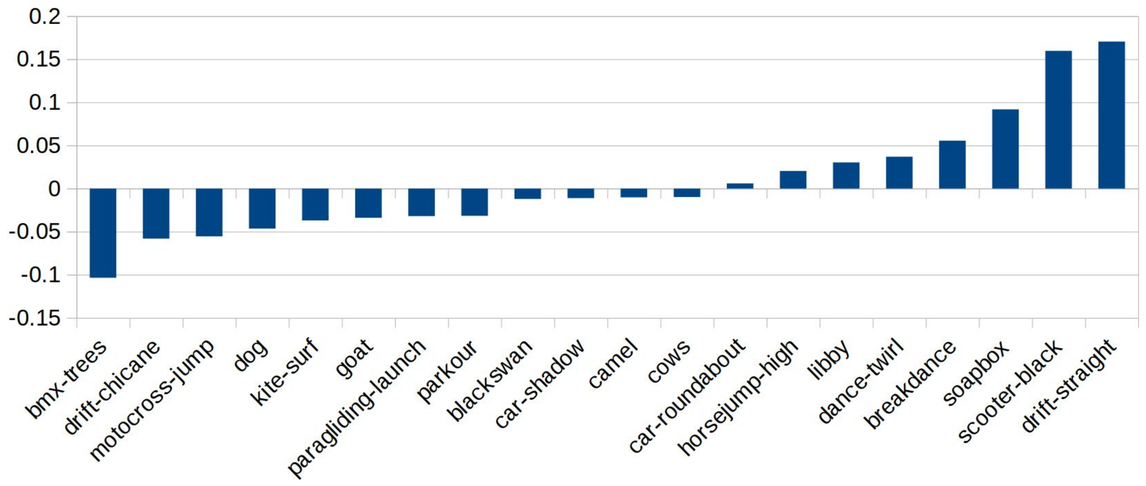
We conduct experiments on both the DAVIS 2016 and 2017 datasets. Our main metrics are Intersection-over-Union (IoU or J) and contour accuracy (F).

Results on DAVIS 2016 Dataset

We compare three variants of our method: iterative training (IT), iterative training on the largest blob (IT+LB) and iterative training with bounding box filtering (IT+Box).

	Measure	OSVOS	IT	IT+LB	IT+Box
J	Mean (↑)	**0.798**	**0.804**	0.794	0.777
	Recall (↑)	**0.936**	**0.974**	0.926	0.911
	Decay (↓)	0.149	**0.091**	**0.138**	0.158
F	Mean (↑)	**0.806**	**0.809**	**0.806**	0.799
	Recall (↑)	0.926	**0.934**	**0.937**	0.915
	Decay (↓)	**0.150**	**0.124**	0.152	0.157

IT shows improvement over the standard OSVOS model in all measures. The following chart shows the relative difference in IoU of IT and OSVOS on each of the sequences in the validation set.



Most improved sequence:

Most harmed sequence:

Results on DAVIS 2017 Dataset

Results on the DAVIS 2017 dataset show a notable reduction over that on the 2016 dataset. The table on the right shows the performance on the test challenge videos for our model variants and OSVOS. In this case, simply running IT reduces the performance compared to OSVOS, possible due to excessive error propagation. Filtering by bounding box improves the results. Qualitatively, the images on the right show the method's ability to track similar objects, and learn to discriminate against features outside of the tracked object bounding boxes.

	Measure	OSVOS	IT	IT+LB	IT+Box
	Overall (↑)	0.488	0.471	**0.500**	**0.509**
J	Mean (↑)	0.462	0.448	**0.481**	**0.490**
	Recall (↑)	0.515	0.479	**0.533**	**0.551**
	Decay (↓)	0.253	0.286	**0.222**	**0.213**
F	Mean (↑)	0.514	0.494	**0.519**	**0.528**
	Recall (↑)	**0.582**	0.524	0.576	**0.583**
	Decay (↓)	0.257	0.291	**0.240**	**0.237**