Combining Self Training and Active Learning for Video Segmentation

Alireza Fathi[1]
afathi3@gatech.edu
Maria Florina Balcan[1]
ninamf@cc.gatech.edu
Xiaofeng Ren[2]
xiaofeng.ren@intel.com
James M. Rehg[1]
rehg@cc.gatech.edu

[1] College of Computing
Georgia Institute of Technology
[2] Intel Labs Seattle

Video object segmentation is an important problem in video analysis, and it has many applications, including post-production, special effects, object recognition, object tracking, and video compression. In particular, the rapidly-growing numbers of videos which are available from the web represent an opportunity for new video applications and analysis methods.

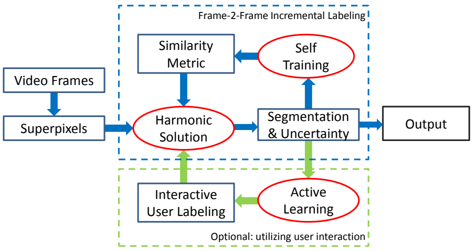
Frame-2-Frame Incremental Labeling
Similarity Metric
Self Training
Video Frames
Superpixels
Harmonic Solution
Segmentation & Uncertainty
Output
Interactive User Labeling
Active Learning
Optional: utilizing user interaction

Figure 1: A flow chart illustration of our approach.

A key motivation for this paper is the observation that video-based object tracking and interactive video object segmentation are closely-related problems. In both cases, the goal is to segment an object from the video with a minimum number of annotations provided by the user. In this paper we address both of these problems within the same framework. In a tracking context, our approach makes it easy to fix up an existing solution by carving the object labeled in the first frame through out the video. Applications to biotracking are a motivating example, as there is a strong need for a general purpose tool for tracking a wide range of animals with different morphologies. In contrast to video post-production, in biotracking applications segmentations which are not pixel-perfect but which delineate the limbs of the animal (for example) can still be useful for animal behavior experiments.

Likewise, in the context of interactive video object cutout, our approach leverages constraints on video data and an active learning approach to minimize the number of annotations that must be supplied by the user. In case of biotracking, it is necessary to minimize the need for guidance by the user in order to have a useful tool for biologists. As a result our method aims at getting the most benefit from each user click. To achieve this goal, we develop an active learning method which chooses the most important frames to be labeled, and which guides the user in each frame about where to click.

Our framework is based on casting video segmentation as a semi-supervised learning problem with video-specific structures such as temporal coherence which makes the following contributions:

A framework for object cutout and interactive segmentation: In this work we present a framework which addresses video object cutout and has a natural extension to interactive segmentation. We use semi-supervised learning to propagate labels from known locations to unknown, and *uncertainty* is a key in a effective label propagation. Furthermore, uncertainty is incorporated into active learning to guide the user in annotating the video.

Incremental self-training: We develop an iterative solution to semi-supervised video segmentation. At each step, we pick the least uncertain frame, fix all labels in the frame, and update system parameters (e.g. object appearance models). We show that incremental self-training is very effective in adapting to video content, outperforming standard semi-supervised learning and state-of-the-art tracking systems.

Intelligent user guidance: We develop a systematic way to provide intelligent guidance to users in an interactive setting. This is by selecting the most informative frames to be labeled by user, and guiding the user while labeling each frame.

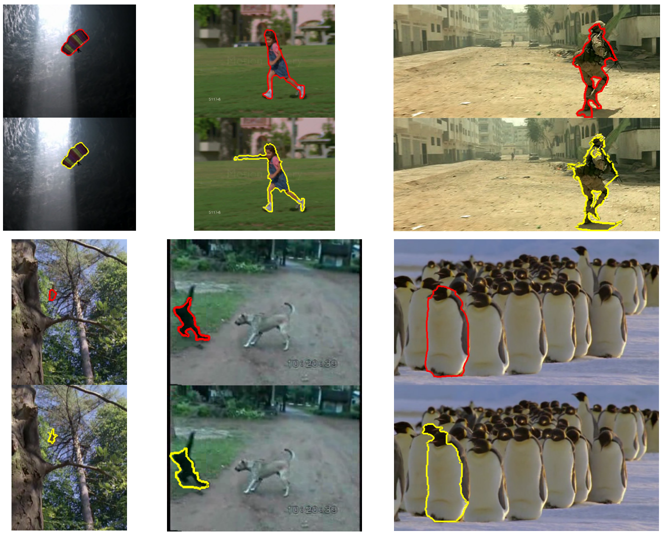

Figure 2: We qualitatively compare our results with [2]. Our results are shown with red contours, and the results from [2] with yellow contours.

Sequence	GraphCut	GraphCut + Self Training	Ours	[1]	[2]
Parachute	254	253	251	502	**235**
Girl	4121	1616	**1206**	1755	1304
Monkey-dog	1312	3727	598	683	**563**
Penguin	19569	19569	**1367**	6627	1705
Bird-fall	454	766	342	454	**252**
Cheetah	1961	898	**711**	1217	1142
Soldier	3344	2484	**1368**	2984	2228
Monkey-water	1306	1266	**1009**	4142	2814

Table 1: We compare our tracking results with [1, 2] using the average number of errors (pixels) per frame.

Figure 1 gives the flow chart for our approach. We formulate video segmentation as a semi-supervised learning problem on a graph of superpixels. Harmonic functions provide an efficient solution to the graph labeling problem and produce soft labels, which we use to measure labeling uncertainties at both the super-pixel and the frame level. We then iteratively choose the least uncertain (or most certain) frame in the video, discretize the soft labels, and apply self-training to update similarity metrics and the affinity graph. We will demonstrate that incremental self-training significantly improves segmentation accuracy in comparison to standard baselines and state of the art segmentation-based tracking methods. We have compared our results with previous work both qualitatively in Fig 2 and quantitatively in Table 1.

Our approach has a natural extension to interactive segmentation. We can present the current segmentation to the user and ask for more input. To minimize user effort, we devise an effective scheme to intelligently suggest the user which frame and which super-pixel to label. Our scheme is empirically validated through simulation.

[1] P. Chockalingam, N. Pradeep, and S. T. Birchfield. Adaptive fragments-based tracking of non-rigid objects using level sets. In *ICCV*, 2009.

[2] D. Tsai, M. Flagg, and J. M. Rehg. Motion coherent tracking with multi-label mrf optimization. In *BMVC*, 2010.