Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network
Supplementary Material

This document describes comprehensive experimental results that can not be accommodated in the main manuscript due to space limitation and provides codec information about the attached video.

1. Evaluation of Segmentation Performance

The proposed algorithm produces pixel-wise target segmentation using target-specific discriminative saliency map. To evaluate segmentation accuracy, we select 9 video sequences from the online tracking benchmark dataset[1] and annotate ground-truth segmentation for each sequence. The selected sequences cover various attributes in tracking challenges, and the list of sequences with associated attributes are summarized in Table 1.

The segmentation performance of the proposed algorithm is evaluated based on the overlap ratio—intersection over union—between ground-truth and identified target segmentation. As other trackers used for comparison may not be able to generate pixel-wise segmentation, we employ their bounding box outputs as segmentation masks and compute the overlap ratio with respect to the ground-truth segmentation. The results are presented by success plot as in Figure 1, where Ours$_{\text{seg}}$ denotes the proposed algorithm with target segmentation. According to Figure 1, our method outperforms all other trackers with substantial margin. Especially, we can observe a large performance improvement of the proposed target segmentation algorithm over our bonding box trackers denoted by Ours and Ours$_{\text{SVM}}$. It suggests that the proposed target-specific saliency map is sufficiently accurate to estimate the target area in a video thus can be utilized to further improve tracking.

2. Codec for Supplementary Video

The video is encoded by MPEG-XVid codec. You can download the codec from `http://www.xvid.org`.

Table 1. List of sequences and their attributes used for segmentation performance evaluation. The set of sequences contains 10 attributes (out of 11 altogether) such as illumination variations (IV), out-of-plane rotation (OPR), scale variations (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), background clutter (BC) and low resolution (LR). The numbers in parentheses denote the number of frames.

Sequence name	Attributes
Bolt (350)	OPR, OCC, DEF, IPR
Coke (291)	IV, OPR, OCC, FM, IPR
Couple (140)	OPR, SC, DEF FM, BC
Jogging (307)	OPR, OCC, DEF
MotorRolling (164)	IV, SC, MB, FM, IPR, BC, LR
MountainBike (228)	OPR, IPR, BC
Walking (412)	SC, OCC, DEF
Walking2 (500)	SC, OCC, LR
Woman (597)	IV, OPR, SC, OCC, DEF, MB, FM

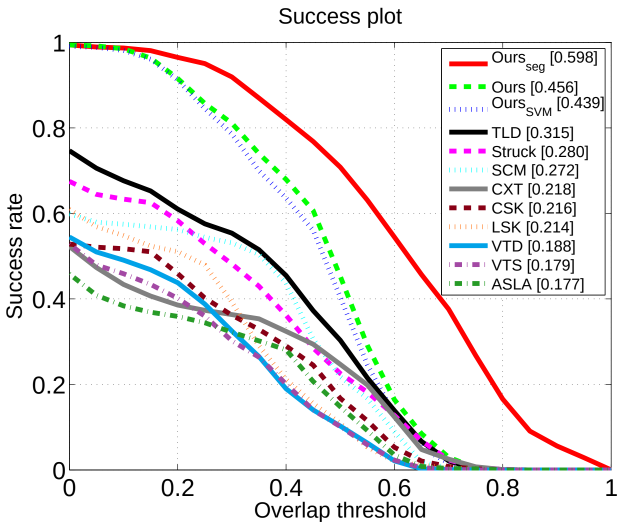

Figure 1. Average success plot over 9 selected sequences. Numbers in the legend indicate overall score of each tracker calculated by area under curve (AUC).

[1] Since accurate annotation of segmentation is labor intensive and time consuming, we selected a subset of sequences (typically short ones) for evaluation. We plan to complete the annotation for the rest of the sequences in the near future.