2019/06/22 3:42:50 PM

Learns the similarity between pairs of patches by training on the ImageNet video data set and then generalizing to other tracking data sets

Output of the network is a similarity/score map where each value corresponds to the similarity of the patch underneath the corresponding sub window of the input image with the template
tracking then consists of simply choosing the sub window with the highest corresponding score

A mapping is fully convolutional if the translated version of the result of applying the mapping to an input signal is identical to the result of applying the mapping to the translated version of the input signal

Training is done using exemplar and search images extracted from the data set where each one as the target object centred
ground truth score map is a +1,-1 thing where the threshold on the distance from the center is used to decide which of the pixels would be +1 and which -1

The network is fully symmetric on the exemplar and search images so that switching them does not change the output feature map

The algorithm described in the paper apparently does not incorporate any kind of model update but the implementation apparently does

One of the contributions of this paper is to show that training on ImageNet video data set is enough to generalize to the mostly unrelated scenes from the tracking benchmark data sets – apparently training and testing on the same benchmark has now been forbidden by VOT to avoid overfitting to the scenes in the benchmark