DeepTrack: Learning Discriminative Feature Representations by Convolutional Neural Networks for Visual Tracking

Hanxi Li[12]
hanxi.li@nicta.com.au
Yi Li[1]
http://users.cecs.anu.edu.au/~yili/
Fatih Porikli[1]
http://www.porikli.com/

[1] NICTA and ANU, Canberra, Australia
[2] Jiangxi Normal University, Jiangxi, China

Defining hand-crafted feature representations needs expert knowledge, requires time-consuming manual adjustments, and besides, it is arguably one of the limiting factors of object tracking.

In this paper, we propose a novel solution to automatically relearn the most useful feature representations during the tracking process in order to accurately adapt appearance changes, pose and scale variations while preventing from drift and tracking failures. We employ a candidate pool of multiple Convolutional Neural Networks (CNNs) as a data-driven model of different instances of the target object. Individually, each CNN maintains a specific set of kernels that favourably discriminate object patches from their surrounding background using all available low-level cues (Fig. 1). These kernels are updated in an online manner at each frame after being trained with just one instance at the initialization of the corresponding CNN.

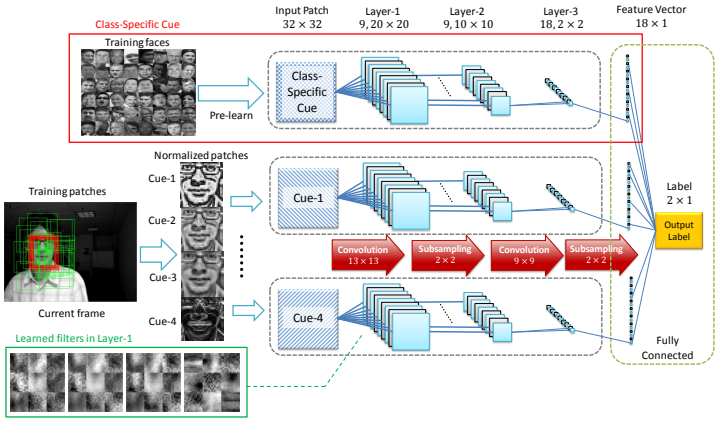

Figure 1: Overall architecture with (red box) and without (rest) the class-specific version.

Instead of learning one complicated and powerful CNN model for all the appearance observations in the past, we chose a relatively small number of filters in the CNN within a framework equipped with a temporal adaptation mechanism (Fig. 2). Given a frame, the most promising CNNs in the pool are selected to evaluate the hypothesises for the target object. The hypothesis with the highest score is assigned as the current detection window and the selected models are retrained using a warm-start back-propagation which optimizes a structural loss function.

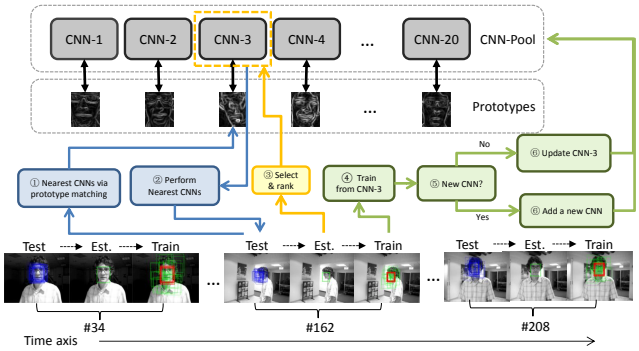

Figure 2: Illustration of the temporal adaptation mechanism.

Our experiments on a large selection of videos from the recent benchmarks demonstrate that our method outperforms the existing state-of-the-art algorithms and rarely loses the track of the target object. We evaluate our method on 16 benchmark video sequences that cover most challenging tracking scenarios such as scale changes, illumination changes, occlusions, cluttered backgrounds and fast motion (Fig. 3).

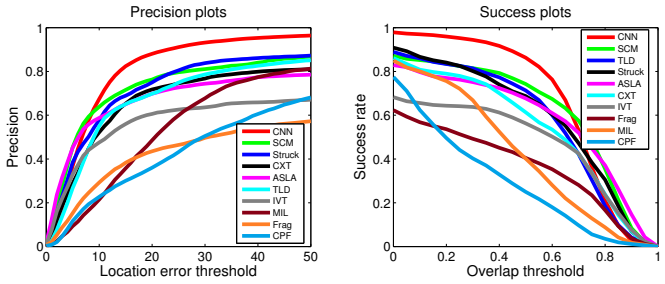

Figure 3: The Precision Plot (left) and the Success Plot (right) of the comparing visual tracking methods.

In certain applications, the target object is from a known class of objects such as human faces. Our method can use this prior information to leverage the performance of tracking by training a class-specific detector. In the tracking stage, given the particular instance information, one needs to combine the class-level detector and the instance-level tracker in a certain way, which usually leads to higher model complexity (Fig. 4).

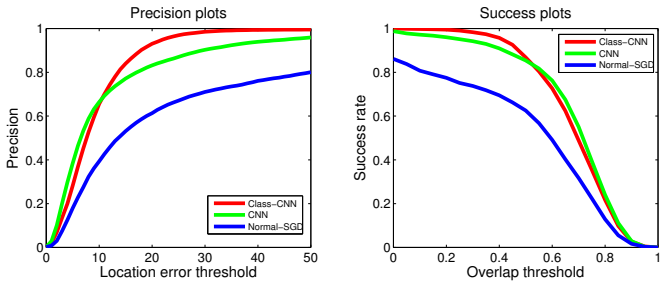

Figure 4: The Precision Plot (left) and the Success Plot (right) of the comparing visual tracking methods.

To conclude, we introduced DeepTrack, a CNN based online object tracker. We employed a CNN architecture and a structural loss function that handles multiple input cues and class-specific tracking. We also proposed an iterative procedure, which speeds up the training process significantly. Together with the CNN pool, our experiments demonstrate that DeepTrack performs very well on 16 sequences.