Transferring Rich Feature Hierarchies for Robust Visual Tracking

mingzailao

2016-09-08 Thu

Outline

Introduction

Goal

Given this single (labeled) instance, the goal is to track the movement of the object in an online manner.

The Contributions

- To alleviate the overfitting and drifting problems during online tracking, we pre-train the CNN to distinguish objects from non-objects instead of simply reconstructing the input or performing categorical classification on large-scale datasets with object-level annotations.

Introduction

The Contributions

- The output of the CNN is a pixel-wise map to indicate the probability that each pixel in the input image belongs to the bounding box of an object. The key advantages of the pixel-wise output are its induced structured loss and computational scalability.
- We evaluate our proposed method on an open benchmark as well as a challenging non-rigid object tracking dataset and obtain very remarkable results. In particular, we improve the area under curve (AUC) metric of the overlap rate curve from 0.529 to 0.602 for the open benchmark.

Overview

Two Stages

- offline pre-training stage
- online fine-tuning and tracking stage

Pre-training Stage

We train a CNN to learn generic object features for distinguishing objects from non-objects.

Tracking Stage

Fine-tuning the parameters so that CNN can adapt to the target.

Overview

For Robustness

Running two CNNs concurrently during online tracking to account for possible mistakes caused by model update.

Objectness Pre-training

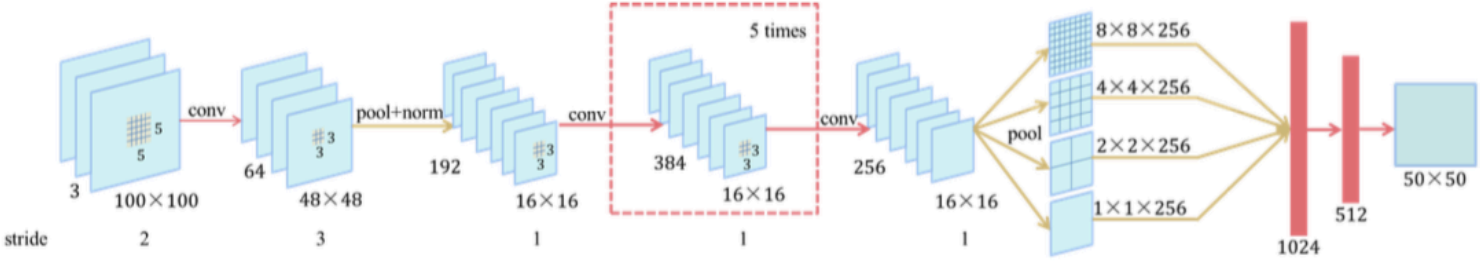

Figure: Structure of the pre-training CNN

Objectness Pre-training

Difference between the tracking model CNN and the classification CNN

The output of the CNN is a 50×50 probability map rather than a single number. Each output pixel corresponds to a 2×2 region in the original input,with its value representing the probability that the corresponding input region belongs to an object

Objectness Pre-training

Mathematically Loss Function

$$\min_{p_{i,j}} \sum_{i=1}^{50} \sum_{j=1}^{50} -(1 - t_{ij}) \log(1 - p_{ij}) - t_{ij} \log(p_{ij}) \quad (1)$$

1. p_{ij} : prediction of (i, j) position
2. t_{ij} : binary variable denotes the ground truth of (i,j) position

Online Tracking

Before

Fine-tuning the network using the annotation in the first frame.

Online Tracking

Basic online tracking pipeline

Two CNNs which use different model update strategie.
After fine-tuning using the annotation in the first frame, we crop some image patches from each new frame based on the estimation of the previous frame.

- Input : the estimation of the previou frame
- Output : some image patched from each new frame

CNN forward out : probability map for each of the image patches.
The final estimation is then determined by searching for a proper bounding box

Bounding Box Determination

The first step of the tracker when a new frame comes

Determine the best location and scale of the target.

1. specify the possible regions that may contain the target and feed the regions into the CNN.
2. decide the most probable location of the bounding box based on the proba- bility map.