Hierarchical Convolutional Features for Visual Tracking
Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang

Introduction

Challenges:
- Significant appearance variations caused by deformation, abrupt motion, illumination changes, background clutter, heavy occlusion, out-of-view, etc.

Spatial Details
- Early layers of CNNs: e.g., intensity, Gabor filter
- Last layer of CNNs: e.g., fc7 in AlexNet

Our Approach
- Exploiting both spatial details and semantics

Semantics
- Our Approach

Contributions:
- Use the rich feature hierarchies of CNNs as target representations for visual tracking, where both semantics and fine-grained details are simultaneously exploited to handle large appearance variations and avoid drifting;
- Adaptively learn linear correlation filters on each CNN layer to alleviate the sampling ambiguity, and infer the target location using the multi-level correlation response maps in a coarse-to-fine fashion.

Method Overview

Correlation Filters

A correlation filter \( w \) is trained from all the circularly shifted input \( x \) with a Gaussian function label \( y_{m,n} \):

\[
\min_{m,n} \sum_{m,n} |w \cdot x_{m,n} - y_{m,n}|^2 + \lambda \|w\|^2.
\]

Where \( w \cdot x_{m,n} = \sum_{d=1}^{D} w_{m,n,d} x_{m,n,d} \). Using the FFT trick, (1) is minimized in the frequency domain on the \( d \)-th \( (d \in \{1, \ldots, D\}) \) channel as:

\[
W_d = Y \bigotimes \hat{X}^2 + \lambda.
\]

Given an new image on the \( l \)-th layer with feature \( z \) of size \( M \times N \times D \), the response is:

\[
f_l = X^{-1} \left( \sum_{d=1}^{D} W_d \otimes \hat{Z}^d \right).
\]

CNN Features

- (d) is with semantic abstraction, and is robust to appearance changes;
- (b) (c) contains more fine-grained spatial details, and is helpful for precise localization;
- It is important to exploit the merits of all layers for robust visual tracking.

Overall Performance

- Precision plots of OPE:
- Success plots of TRE:
- Precision plots of SRE:
- Success plots of SRE:

Discussion

Our method performs favorably against state-of-the-art methods:
- CNN features (e.g., VGG-Net) learned with category-level supervision are effective in discriminating targets from background;
- The deeper layer (conv5-4) is insensitive to appearance changes, and is weighted more than earlier layers (conv3-4 and conv4-4).

Reference


Code available at https://github.com/jbhuang0604/CF

Fig. 1. Visualization of the CNN features using VGG-Net-19 [1].
- The conv5-4 layer is less effective to locate the step edge due to its low spatial resolution;
- The conv3-4 layer is more useful for precise localization.

Fig. 2. Main steps of the proposed algorithm.
- Crop the search window;
- Compute the response map for each layer;
- Estimate translation hierarchically.

Fig. 3. Visualization of convolutional layers.
- (d) is with semantic abstraction, and is robust to appearance changes;
- (b) (c) contains more fine-grained spatial details, and is helpful for precise localization;
- It is important to exploit the merits of all layers for robust visual tracking.

Fig. 4. Distance precision and overlap success plots on OTB-100 [3] using OPE, TRE and SRE.

Fig. 5. Results with VGG-Net and AlexNet. c5,c4 and c3: each single convolutional layer; c5-c4: the combination of conv5 and conv4; c543: the concatenation of three convolutional layers.

More Results

Table 1. Results of distance precision (DP) rate at a threshold of 20 pixels, overlap success (OS) rate at a threshold of 0.5 and center location error (CLE) on OTB-50 (I) [2] and OTB-100 (II) [3].

Ablation Study