Introduction:

➢ Single object tracking:
  Target localization in the video frames.

➢ Existing frameworks:
  • Tracking by detection vs. discriminative correlation filter.

➢ Insights on the Discriminative Correlation Filter (DCF):
  Pros:
  • Efficient correlation operation in the Fourier domain.
  • Dense prediction for target locations.
  Cons:
  • Boundary effect via Fourier transform.
  • The whole framework is empirically designed (i.e., filter weights training, model update, feature integration).

Our formulations:

➢ The objective function of DCF is ridge regression:
  \[ W^* = \arg\min_W ||W \ast X - Y||^2 + \lambda \cdot ||W||^2 \]

➢ We use single convolutional layer \( W \) to replace DCF.
  ✓ End-to-end integration with convolutional features.
  ✓ Filter weights optimization via gradient descent.

➢ We adopt residual learning to measure the difference between the convolutional layer output and the ground truth.
  \[ H(x) = F(x, \{W_i\}) + W \ast x \]
  where \( H \) is the ground truth optimal mapping and \( F \) is the residual mapping.

Our contributions:

✓ We formulate feature extraction and response generation in an end-to-end form via CNN. We adopt back propagation for model update and fully exploit the deep architecture.

✓ We use residual learning to handle large appearance variations, which alleviates model degradation.

✓ State-of-the-art performance on the prevalent benchmarks.

Framework:

Base mapping: DCF

Feature extraction

Frame T

Feature maps

Residual mapping

Response map

Base output

Residual output

Network output

Visualization:

Input frame

Base output

Internal evaluations

External evaluations

Experiments:

Evaluations on the OTB 2013 dataset.

We show evaluations on the OTB 2015 and VOT 2016 datasets in the paper. Our implementation is available online.