Unsupervised Deep Tracking

Vitalii Duma

August 9, 2019

Paper

- **Title:** Unsupervised Deep Tracking
- **Authors:** Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, Houqiang Li1
- **Link:** https://arxiv.org/pdf/1904.01828.pdf
- **Tags:** visual tracking, unsupervised learning, deep neural networks
- **Year:** 2019

Summary

- What
  - They propose unsupervised tracking method based on the Siamese correlation filter backbone
  - They propose multiple-frame validation method and a cost-sensitive loss
  - They show extensive experiments on the standard benchmarks
- How
  - Randomly draw bounding boxes in unlabeled videos to perform forward and backward tracking. Track forward to predict boxes location in the subsequent frame.
  - Perform forward tracking
    * Given two consecutive frames P_1 and P_2. Build a Siamese correlation filter network to track the initial bounding box region in frame P_1.
    * Build target template using Siamese correlation filter.
    * Calculate response map of the search path S from frame P_2.
  - Perform backward tracking
    * After generating response map R_s from P_2, they treat S as the template patch, and generate a target template W_S using pseudo Gaussian labels.

– Calculate consistency loss having response map from backward and forward tracking (R_t). Their loss function:

$$L_{un} = ||R_T - Y_T||_2^2$$

where Y_T is the originaly given label after backward and forward tracking

– They propose next unsupervised learning improvements:

  * A multiple frames validation approach to alleviate the inaccurate localization issue that is not penalized by loss function.
  * Cost-sensitive loss.
    · During unsupervised learning, they construct multiple training pairs from the training sequences. They found that few training pairs with extremely high losses prevent the network training from convergence.
    · They introduce A_{motion} weight vector which indicates that the target undergoes a larger movement in this continuous trajectory.
  * After offline unsupervised learning, they online track the target object following forward tracking as :

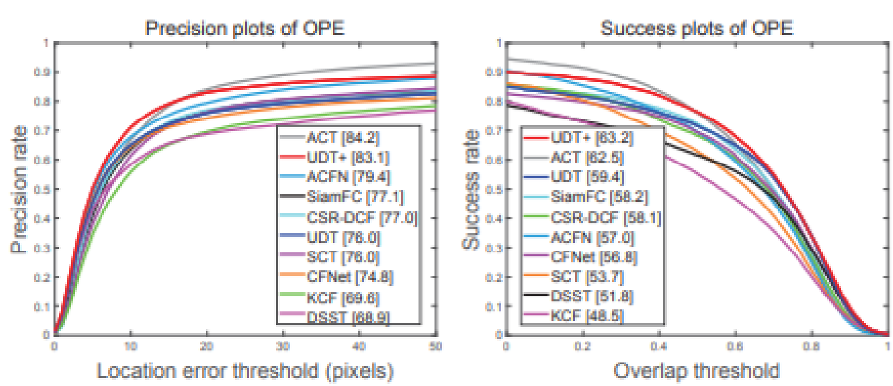
$$W_t = (1 - t)W_{t-1} + tW$$

– Training:

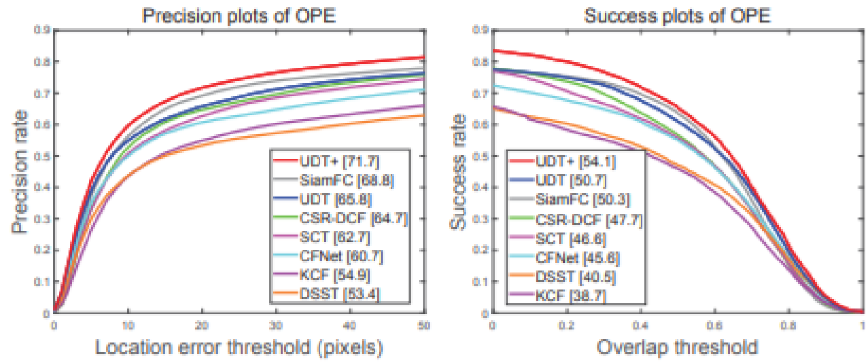
  * They choose the widely used ILSVRC 2015 as training data to fairly compare with existing supervised trackers.
  * They do not preprocess any data and simply crop the center patch in each frame.
  * They randomly choose three cropped patches from the continuous 10 frames in a video for training. This is based on the assumption that the center located target objects are unlikely to move out of the cropped region in a short period.

– Results:

  * Precision and success plots on the OTB-2015 dataset for recent real-time trackers.



* Precision and success plots on the Temple-Color dataset for recent real-time trackers.



* Comparison with state-of-the-art and baseline trackers on the VOT2016 benchmark. The evaluation metrics include Accuracy, Failures (over 60 sequences), and Expected Average Overlap (EAO).

Trackers	Accuracy (↑)	Failures (↓)	EAO (↑)	FPS (↑)
ECO [7]	0.54	-	0.374	6
C-COT [11]	0.52	51	0.331	0.3
pyMDNet [37]	-	-	0.304	2
SA-Siam [15]	0.53	-	0.291	50
StructSiam [60]	-	-	0.264	45
MemTrack [58]	0.53	-	0.273	50
SiamFC [1]	0.53	99	0.235	86
SCT [5]	0.48	117	0.188	40
DSST [8]	0.53	151	0.181	25
KCF [16]	0.49	122	0.192	170
UDT (Ours)	0.54	102	0.226	70
UDT+ (Ours)	0.53	66	0.301	55

* Limitations:
  - Unsupervised feature representation may lack the objectness information to cope with complex scenarios.
  - Since the approach involves both forward and backward tracking, the computational load is another potential drawback.