Copyright © 2019 Tech Science Press JIHPP, vol.1, no.1, pp.23-33, 2019

Deep Learning Trackers Review and Challenge

Yongxiang Gu[1], Beijing Chen[1], Xu Cheng[1, *], Yifeng Zhang[2, 3] and Jingang Shi[4]

Abstract: Recently, deep learning has achieved great success in visual tracking. The goal of this paper is to review the state-of-the-art tracking methods based on deep learning. First, we categorize the existing deep learning based trackers into three classes according to network structure, network function and network training. For each categorize, we analyze papers in different categories. Then, we conduct extensive experiments to compare the representative methods on the popular OTB-100, TC-128 and VOT2015 benchmarks. Based on our observations. We conclude that: (1) The usage of the convolutional neural network (CNN) model could significantly improve the tracking performance. (2) The trackers with deep features perform much better than those with low-level hand-crafted features. (3) Deep features from different convolutional layers have different characteristics and the effective combination of them usually results in a more robust tracker. (4) The deep visual trackers using end-to-end networks usually perform better than the trackers merely using feature extraction networks. (5) For visual tracking, the most suitable network training method is to per-train networks with video information and online fine-tune them with subsequent observations. Finally, we summarize our manuscript and highlight our insights, and point out the further trends for deep visual tracking.

Keywords: Deep learning, CNN, object tracking, online learning.

1 Introduction

Visual tracking has numerous realistic applications in navigation, surveillance, robotics, augmented reality, to name a few. Many efforts have been done in last decades. However, it is still a challenging task to develop a robust tracker to handle the complex scenes.

Traditional trackers usually focus on developing robust appearance model from the perspectives of hand-crafted features, online learning algorithms or both. Some milestones include IVT [Ross, Lim, Lin et al. (2008)], MIL [Babenko, Yang, Belongie et al. (2011)], TLD [Kalal, Mikolajczyk and Matas (2012)], APGL1 [Bao, Wu, Ling et al.

[1] School of Computer and Software, Nanjing University of Information Science and Technology, 210044, China.

[2] School of Information Science and Engineering, Southeast University, 210096, China.

[3] State Key Laboratory for Novel Software Technology，Nanjing University, China.

[4] The Center for Machine Vision and Signal Analysis, University of Oulu, FI-90014 Oulu, Finland.

[*] Corresponding Author: Xu Cheng. Email: xcheng@nuist.edu.cn.

JIHPP. doi:10.32604/jihpp.2019.05938 www.techscience.com/jihpp

(2012)], SCM [Zhong, Lu and Yang (2014)], ASLAS [Jia, Lu and Yang (2016)], STRUCK [Hare, Golodetz, Saffari et al. (2016)] and KCF [Wu, Lim and Yang (2013)]. However, the reports on large-benchmark evaluations (both OTB-100 [Wu, Lim and Yang (2015)], TC128 [Liang, Blasch and Ling (2015)] and VOT2015 [Kristan, Matas, Leonardis et al. (2015)]) suggest that the performance of these traditional algorithms is far from the requirement of realistic applications.

Recently, deep learning [LeCun, Bengio and Hinton (2015)] have achieved an impressive suite of results thanks to their success on automatic feature extraction via multi-layer nonlinear transformations, especially in computer vision [Girshick, Donahue, Darrell et al. (2016); Schroff, Kalenichenko and Philbin (2015); Zhou, Liang, Li et al. (2018); Li, Jiang and Cheslyar (2018)], speech recognition [Kim, Hori and Watanabe (2017); Wu, Valentini-Botinhao, Watts et al. (2015)] and natural language processing [Vinyals, Kaiser, Koo et al. (2015); Bahdanau, Cho and Bengio (2014); Zhang, Wang, Li et al. (2018)]. Motivated by these breakthroughs, several deep-learning-based trackers (e.g., FCNT [Wang, Ouyang, Wang et al. (2015)], MDNet [Nam and Han (2016)], STCT [Wang, Ouyang, Wang et al. (2016)], SINT [Tao, Gavves and Smeulders (2016)], SiameFC [Bertinetto, Valmadre, Henriques et al. (2016)], C-COT [Danelljan, Robinson, Khan et al. (2016)], GO-TURN [Held, Thrun and Savarese (2016)], TCNN [Nam, Baek and Han (2016)], ADNet [Yun, Choi, Yoo et al. (2017)] and SANet [Fan and Ling (2017)]) have demonstrated the potential advantages for significantly improving the tracking performance. The performance on the OTB-100 [Wu, Lim and Yang (2015)] dataset is constantly refreshed by the tracking methods based on deep learning (such as DeepSRDCF [Danelljan, Häger, Khan et al. (2015)], HCFT [Ma, Huang, Yang et al. (2015)] and HDT [Qi, Zhang, Qin et al. (2016)]). The MDNet [Nam and Han (2016)] tracker is the winner of VOT2015 competition. All the top-4 trackers in the VOT2016 competition, including C-COT [Danelljan, Robinson, Khan et al. (2016)], TCNN [Nam, Baek and Han (2016)], SSAT and MLDF, are based on deep neural networks. In this work, we review existing deep learning based tracking algorithms and evaluate them on recent benchmarks.

The rest of the paper is organized as follows. In Section 2, we review the existing trackers based on deep learning. In Section 3, we report the experiment evaluations of different trackers and give some discussions and analysis on them. Finally, we summarize this paper and point out some further directions in Section 4.

2 Deep object tracking algorithms

2.1 Network structure

Networks with different structures focus on solving different tasks. The convolutional neural network (CNN) has been demonstrated to be effective for feature extraction and achieved great success on image classification. While the recurrent neural network (RNN) is able to remember previous states and establish temporal connection, which is suitable for sequence modeling.

The CNN model is very suitable for developing robust appearance model in the tracking task, due to its powerful ability on feature extraction and image classification. For the VGGNet [Simonyan and Zisserman (2014)] model, Ma et al. [Ma, Huang, Yang et al.

(2015)] find that the outputs of the last convolutional layer encode the semantic information and such representations are robust to significant appearance variations. However, their spatial resolution is too coarse to precisely localize the tracked objects. The DeepSRDCF [Danelljan, Häger, Khan et al. (2015)] method combines activations from the convolutional layer of a CNN model with the SRDCF framework. Some similar ideas have been presented in Danelljan et al. [Danelljan, Robinson, Khan et al. (2016); Yun, Choi, Yoo et al. (2017); Simonyan and Zisserman (2014)]. Tao et al. [Tao, Gavves and Smeulders (2016)] propose a Siamese network model to match the object template and candidates for visual tracking. After that, Bertinetto et al. [Bertinetto, Valmadre, Henriques et al. (2016)] develop a fully connected Siamese network to match the object template and current search region in a convolutional manner. Recently, some trackers apply the satisfactory performance of correlation filter to deep neural network. Valmadre et al. [Valmadre, Bertinetto and Henriques (2017)] interpret correlation filter as a differentiable layer in a Siamese network.

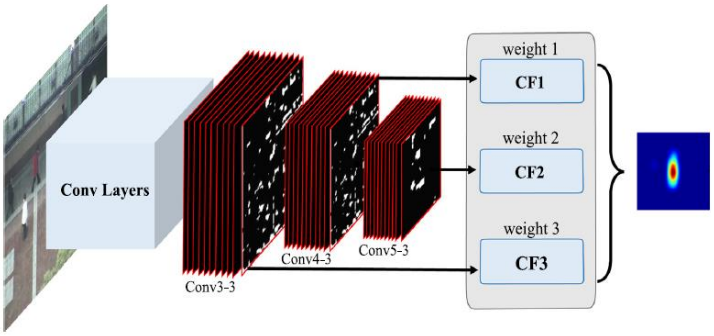

Figure 1: The structure of HCFT

2.2 Network function

For visual tracking, deep networks can be not only used for extracting effective features but also adopted for evaluating the candidates of the tracked object.

Feature extraction network (FEN): Feature extraction network merely uses deep networks to extract deep features and then adopts the traditional method to learn the appearance model and locate the target. The DeepSRDCF [Danelljan, Häger, Khan et al. (2015)] method extracts features from the first layers of the VGG network, and combines deep features with the SRDCF framework to improve the tracking performance. Ma et al. [Ma, Huang, Yang et al. (2015)] observe that different convolutional layers of a typical CNN model provide multiple levels of abstraction in the feature hierarchies. Features in earlier layers retain higher spatial resolution for precise localization with low-level visual information. While features in latter layers capture more semantic information and less fine-grained spatial details. Thus, they extract features from three different layers and use a fix weight to combine the feature maps generated by those layers. Ma et al. [Ma, Xu, Ni et al. (2016)] use a designed network to replace the Conv3-3 layer in Ma et al. [Ma, Huang, Yang et al. (2015)] and therefore improve the tracking performance. Qi et al. [Qi, Zhang, Qin et al. (2016)] extract features from six convolutional layers and combine these layers using an adaptive weight scheme. The C-COT [Danelljan, Robinson, Khan et al. (2016)] algorithm proposes a joint learning framework to fuse deep features from different spatial pyramids.

End to end network (EEN): This network only uses deep networks for feature extraction but also for candidate evaluation. The outputs of the EEN methods can be in terms of probability map, heat map, candidate's score, object position or even bounding determined directly. In Tao et al. [Tao, Gavves and Smeulders (2016)], the SINT method generates a lot of particles and calculates their similarity scores using the Siamese network. The optimal state can be determined by the particle with the highest score. In [Wang, Ouyang, Wang et al. (2015)], two deep networks are designed with the Conv4-3 and Conv5-3 layers of the VGG-16 model and then used to calculate the response maps. The SiameFC [Bertinetto, Valmadre, Henriques et al. (2016)] tracker utilizes a pre-trained fully convolutional Siamese network, the inputs of which are the object template and current search region. In each frame, this method generates a response map regarding the tracked object with convolution operation.

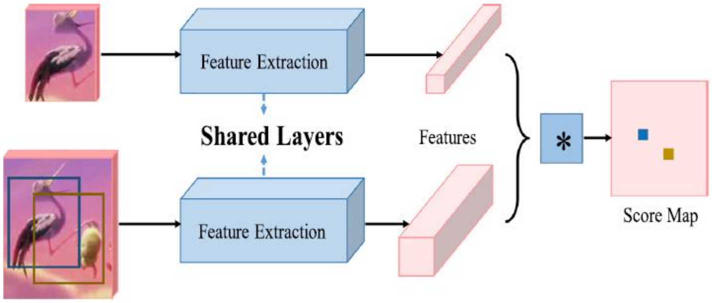

Figure 2: The structure of SiameFC

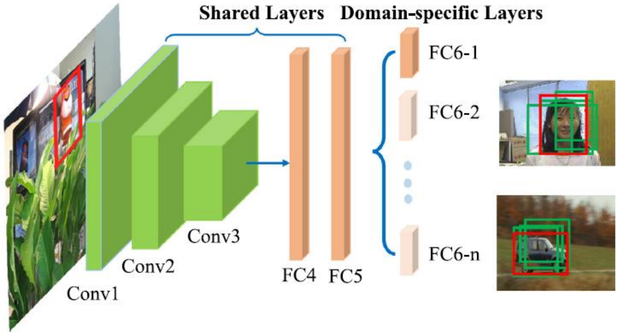

Figure 3: The structure of MDNet

2.3 Network training

The network training is also a critical issue for developing a robust deep learning based tracker, which may be used to transfer visual prior, online learning, or both. The CNT method collects a series of positive and negative patches to learn many image filters, and builds a robust appearance model using convolutional operators. The lack of labeled data in online learning limits the performance of deep networks. Thus, it is reasonable to transfer the visual prior using existing networks pre-trained by massive natural images. The well-known VGGNet [Simonyan and Zisserman (2014)] (offline trained on the large-scale ImageNet dataset) is widely used to design the tracking methods because of their satisfactory performance on image classification. The network in the FCNT [Wang, Ouyang, Wang et al. (2015)] method is comprised of VGG-16 convolutional layers and an additional net- work with three designed layers. The parameters of former layers are pre-trained with the ImageNet dataset and fixed during the tracking process. The later layers are flexible enough to capture the appearance change with online update. Nam et al. effectively combine pre-trained convolutional layers and multiple fully-connected layers with a tree structure to achieve good tracking performance. SANet [Fan and Ling (2017)] method exploits the same idea with the MDNet tracker, and introduces an additional RNN-based structure to further enhance object representation.

Recently, deep reinforcement learning has drawn more attention in visual tracking, which is suitable when we lack of training labels or have delayed labels. Yun et al. [Yun, Choi, Yoo et al. (2017)] attempt to learn an agent which could evaluate the moving direction of bounding box through unsupervised learning in a frame. The part of whole model is pre-trained offline with videos. Then, the network is updated online.

3 Experiment and evaluation

In our experiment, we collect 16 deep visual trackers, the source codes or benchmark results of which have been publicly available already. These methods include MEEM [Zhang and Ma (2014b)], KCF [Henriques, Caseiro, Martins et al. (2015)], TGPR [Gao, Ling, Hu et al. (2014)], SCM [Zhong, Lu and Yang (2014)], ASLA [Jia, Lu and Yang (2016)], SRDCF [Danelljan, Häger, Khan et al. (2015)], ECO [Danelljan, Bhat, Khan et al. (2017)], CFNet [Valmadre, Bertinetto, Henriques et al. (2017)], MCPF [Zhang, Xu and Yang (2017)], DNT [Chi, Li, Lu et al. (2017)], DCFNet [Wang, Gao, Xing et al. (2017)], MDNet [Nam and Han (2016)], SANet [Fan and Ling (2017)], TCNN [Nam, Baek and Han (2016)], C-COT [Danelljan, Robinson, Khan et al. (2016)], STCT [Wang, Ouyang, Wang et al. (2016)], FCNT [Wang, Ouyang, Wang et al. (2015)], HCFT [Ma, Huang, Yang et al. (2015)], HDT [Qi, Zhang, Qin, et al. (2016)] and SiameFC [Bertinetto, Valmadre, Henriques et al. (2016)]. These methods have achieved top performance on three benchmarks. In our experiments, we run the source codes (with same parameters) or use tracking results provided by the original authors to conduct experimental comparisons. Specially, all trackers are retested on TC-128 [Liang, Blasch and Ling (2015)] for fair speed comparison.

3.1 Evaluation benchmark

This subsection introduces the adopted dataset (OTB-100 [Wu, Lim and Yang (2015)], TC-128 [Liang, Blasch and Ling (2015)], VOT2015 [Kristan, Matas, Leonardis et al. (2015)] and deep-learning based trackers to be evaluated in this paper.

OTB-100: The OTB-100 [Wu, Lim and Yang (2015)] dataset is presented by Wu et al., which has been one of most commonly used benchmarks in evaluating online visual trackers. This dataset includes 100 challenging video clips annotated with different attributes, such as Illumination Variation (IV), Scale Variation (SV), Occlusion (OCC), Deformation (DEF), Motion Blur (MB), Fast Motion (FM), In-Plane Rotation (IPR), Out-of-Plane Rotation (OPR), Out-of-View (OV), Background Clutters (BC), and Low Resolution (LR). By the 11 different attributes, we can analyze the performance of trackers in different aspects. The performance of the 23 trackers is quantitatively validated using two metrics including distance precision and success rate.

VOT2015: The VOT2015 [Kristan, Matas, Leonardis et al. (2015)] dataset consists of 60 short sequences annotated with 6 different attributes including occlusion, illumination change, motion change, size change, camera motion and unassigned. The major difference between VOT2015 and OTB-100 is that the VOT2015 challenge provides a reinitialization protocol (i.e., trackers are reset with ground-truths in the middle of evaluation if tracking failures are observed).

TC-128: The TC-128 [Liang, Blasch and Ling (2015)] dataset is presented by Liang et al. and focus on color information. This benchmark contains 128 color sequences with ground truth and challenge factor annotations, including Illumination Variation (IV), Scale Variation (SV), Occlusion (OCC), Deformation (DEF), Motion Blur (MB), Fast Motion (FM), In-Plane Rotation (IPR), Out-of-Plane Rotation (OPR), Out-of-View (OV), Background Clutters (BC), and Low Resolution (LR). The 11 challenge factors as well as the evaluation metrics here are the same with the OTB-100 [Wu, Lim and Yang (2015)] dataset.

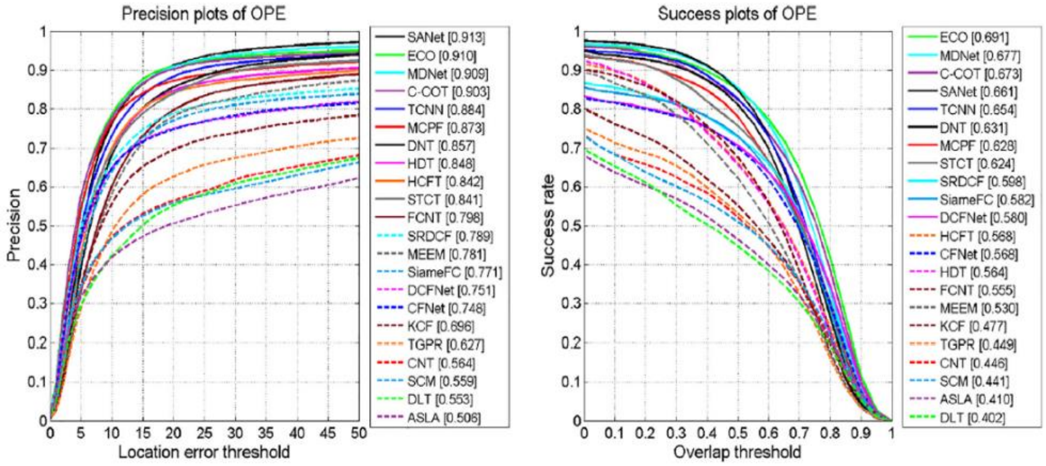

Figure 4: Distance precision plots and success rate plots over OTB100 benchmark on 22 trackers using one-pass evaluation (OPE)

Table 1: Detailed information of deep visual trackers evaluated in our paper. Abb Abbreviation, NS-Network Structure, NF-Network Function, C-Code (M-Matlab, m-Matconvnet, c-Caffe)

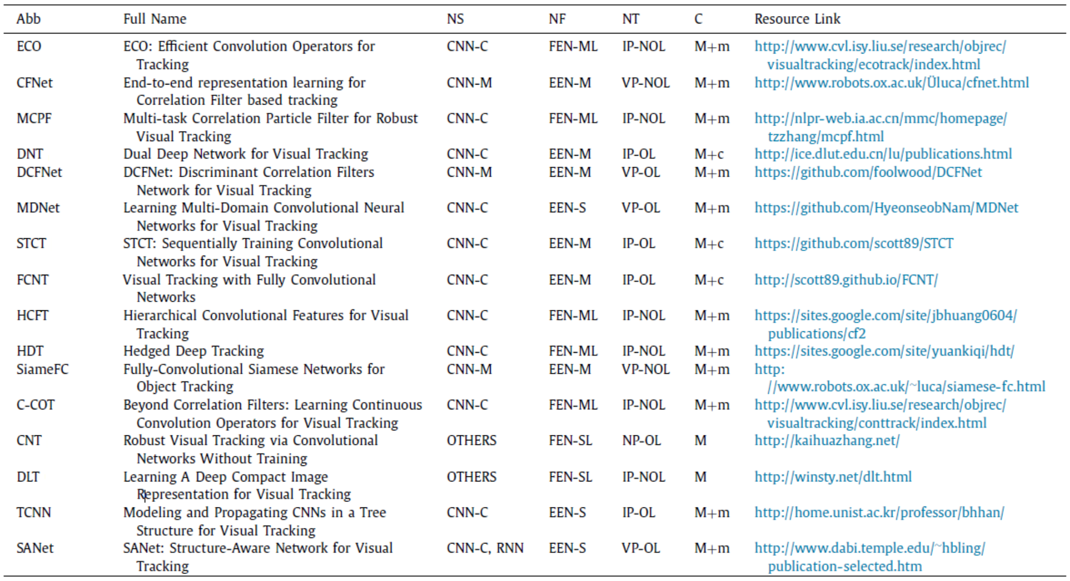

Abb	Full Name	NS	NF	NT	C	Resource Link
ECO	ECO: Efficient Convolution Operators for Tracking	CNN-C	FEN-ML	IP-NOL	M+m	http://www.cvl.isy.liu.se/research/objrec/visualtracking/ecotrack/index.html
CFNet	End-to-end representation learning for Correlation Filter based tracking	CNN-M	EEN-M	VP-NOL	M+m	http://www.robots.ox.ac.uk/Üluca/cfnet.html
MCPF	Multi-task Correlation Particle Filter for Robust Visual Tracking	CNN-C	FEN-ML	IP-NOL	M+m	http://nlpr-web.ia.ac.cn/mmc/homepage/tzzhang/mcpf.html
DNT	Dual Deep Network for Visual Tracking	CNN-C	EEN-M	IP-OL	M+c	http://ice.dlut.edu.cn/lu/publications.html
DCFNet	DCFNet: Discriminant Correlation Filters Network for Visual Tracking	CNN-M	EEN-M	VP-OL	M+m	https://github.com/foolwood/DCFNet
MDNet	Learning Multi-Domain Convolutional Neural Networks for Visual Tracking	CNN-C	EEN-S	VP-OL	M+m	https://github.com/HyeonseobNam/MDNet
STCT	STCT: Sequentially Training Convolutional Networks for Visual Tracking	CNN-C	EEN-M	IP-OL	M+c	https://github.com/scott89/STCT
FCNT	Visual Tracking with Fully Convolutional Networks	CNN-C	EEN-M	IP-OL	M+c	http://scott89.github.io/FCNT/
HCFT	Hierarchical Convolutional Features for Visual Tracking	CNN-C	FEN-ML	IP-NOL	M+m	https://sites.google.com/site/jbhuang0604/publications/cf2
HDT	Hedged Deep Tracking	CNN-C	FEN-ML	IP-NOL	M+m	https://sites.google.com/site/yuankiqi/hdt/
SiameFC	Fully-Convolutional Siamese Networks for Object Tracking	CNN-M	EEN-M	VP-NOL	M+m	http://www.robots.ox.ac.uk/~luca/siamese-fc.html
C-COT	Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking	CNN-C	FEN-ML	IP-NOL	M+m	http://www.cvl.isy.liu.se/research/objrec/visualtracking/conttrack/index.html
CNT	Robust Visual Tracking via Convolutional Networks Without Training	OTHERS	FEN-SL	NP-OL	M	http://kaihuazhang.net/
DLT	Learning A Deep Compact Image Representation for Visual Tracking	OTHERS	FEN-SL	IP-NOL	M	http://winsty.net/dlt.html
TCNN	Modeling and Propagating CNNs in a Tree Structure for Visual Tracking	CNN-C	EEN-S	IP-OL	M+m	http://home.unist.ac.kr/professor/bhhan/
SANet	SANet: Structure-Aware Network for Visual Tracking	CNN-C, RNN	EEN-S	VP-OL	M+m	http://www.dabi.temple.edu/~hbling/publication-selected.htm

3.2 Tracking speed

Speed is also an important aspect for online visual tracking, which is mainly effected by model complexity and update frequency. Generally, deep feature extraction needs more time than hand-crafted feature extraction. But the truth is not the deeper the network is, the better the features are. Because the resolution of tracking video sequence is lower than that of image classification or object detection, so very deep network will make much lost of information. Most state-of-the-art trackers (MDNet, SANet, ADNet) use VGG-M instead of VGG-16 or VGG-19. It makes tracker faster and more accurate. Besides, the lack of prior information in visual tracking need the tracker conserve template or update model to fit different sequences. But it is often time consuming to update the deep network frame by frame. Thus, some trackers (FCNT, ECO) update the network every few frames and others (SiameFC, DCFNet, CFNet) apply Siamese network in visual tracking. They use one path of network to model template, replacing online update of classifier. We can observe from Fig. 4 that these methods are much faster than others.

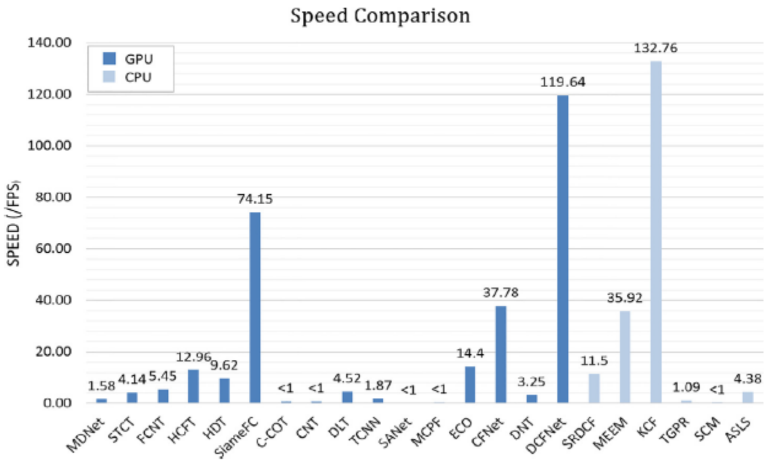

Figure 5: Speed comparison over TC-128 Benchmark

4 Conclusion and further work

In this work, we review the recently proposed visual trackers based on deep learning and conduct extensive experiments to evaluate the existing deep learning based tracking methods. The main contributions of this work are two-folds. First, we review the existing deep visual trackers in three aspects including network structure, network function and network training. Second, we conduct extensive experiments to compare the representative methods on the popular OTB-100, TC-128 and VOT2015 benchmarks. Although deep learning has used in visual tracking and achieved promising improvements compared with traditional trackers, there also exist many topics to be investigated. First, the deep features have much redundancy which limits both speed and accuracy improvement. It will be a promising direction to reduce the redundancy in deep visual tracking. Second, most trackers use VGG network. Developing more effective network structures should be noticed. Third, the lack of training data needs more focus on unsupervised or weakly supervised learning. Reinforcement learning or exploiting generative adversarial networks to generate more training samples will improve the tracking performance. Besides, the transfer ability of model is pretty important in visual tracking. In conclusion, improving tracking efficiency and solving the lack of training data will be new directions.

Acknowledgement: This paper is supported by the Natural Science Foundation of China no. 61802058. State Key Laboratory of Novel Software Technology (Nanjing University) no. KFKT2017B17. Nanjing University of Information Science and Technology Start-up fund No. 2018r057.

References

Babenko, B.;Yang, M. H.; Belongie, S. J. (2011): Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619-1632.

Bahdanau, D.; Cho, K.; Bengio, Y. (2014): Neural machine translation by jointly learning to align and translate. arXiv:1409.0473.

Bao, C.; Wu, Y.; Ling, H.; Ji, H. (2012): Real time robust L1 tracker using accelerated proximal gradient approach. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1830-1837.

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; Torr, P. H. S. (2016): Fully convolutional siamese networks for object tracking. *Proceedings of the European Conference on Computer Vision Workshops*, pp. 850-865.

Chi, Z.; Li, H.; Lu, H.; Yang, M. (2017): Dual deep network for visual tracking. *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 2005-2015.

Danelljan, M.; Häger, G.; Khan, F. S.; Felsberg, M. (2015): Learning spatially regularized correlation filters for visual tracking. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4310-4318.

Danelljan, M.; Bhat, G.; Khan, F. S.; Felsberg, M. (2017): ECO: efficient convolution operators for tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6638-6646.

Danelljan, M.; Robinson, A.; Khan, F. S.; Felsberg, M. (2016): Beyond correlation filters: learning continuous convolution operators for visual tracking. *Proceedings of the European Conference on Computer Vision*, pp. 472-488.

Fan, H.; Ling, H. (2017): Sanet: structure-aware network for visual tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.*

Gao, J.; Ling, H.; Hu, W.; Xing, J. (2014): Transfer learning based visual tracking with gaussian processes regression. *Proceedings of the European Conference on Computer Vision*, pp. 188-203.

Girshick, R. B.; Donahue, J.; Darrell, T.; Malik, J. (2016): Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142-158.

Hare, S.; Golodetz, S.; Saffari, A.; Vineet, V.; Cheng, M. et al. (2016): Struck: structured output tracking with kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2096-2109.

Held, D.; Thrun, S.; Savarese, S. (2016): Learning to track at 100 FPS with deep regression networks. *Proceedings of the European Conference on Computer Vision*, pp. 749-765.

Henriques, J. F.; Caseiro, R.; Martins, P.; Batista, J. (2015): High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583-596.

Jia, X.; Lu, H. C.; Yang, M. (2016): Visual tracking via coarse and fine structural local sparse appearance models. *IEEE Xplore: IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4555-4564.

Kalal, Z.; Mikolajczyk, K.; Matas, J. (2012): Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409-1422.

Kim, S.; Hori, T.; Watanabe, S. (2017): Joint ctc-attention based end-to-end speech recognition using multi-task learning. *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, pp. 4835-4839.

Kristan, M.; Matas, J.; Leonardis, A.; Felsberg, M.; Cehovin, L. et al. (2015): The visual object tracking VOT2015 challenge results. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 564-586.

LeCun, Y.; Bengio, Y.; Hinton, G. E. (2015): Deep learning. *Nature*, pp. 436-444.

Li, C. L.; Jiang, Y. M.; Cheslyar, M. (2018): Embedding image through generated intermediate medium using deep convolutional generative adversarial network. *Computers, Materials & Continua*, vol. 56, no. 2, pp. 313-324.

Liang, P. P.; Blasch, E.; Ling, H. (2015): Encoding color information for visual tracking: algorithms and benchmark. *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5630-5644.

Ma, C.; Huang, J. B.; Yang, X.; Yang, M. H. (2015): Hierarchical convolutional features for visual tracking. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3074-3082.

Ma, C.; Xu, Y.; Ni, B.; Yang, X. (2016): When correlation filters meet convolutional neural networks for visual tracking. *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1454-1458.

Nam, H.; Han, B. (2016): Learning multi-domain convolutional neural networks for visual tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4293-4302.

Nam, H.; Baek, M.; Han, B. (2016): Modeling and propagating cnns in a tree structure for visual tracking. arXiv:1608.07242.

Qi, Y.; Zhang, S.; Qin, L.; Yao, H.; Huang, Q. et al. (2016): Hedged deep tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4303-4311.

Ross, D. A.; Lim, J.; Lin, R.; Yang, M. (2008): Incremental learning for robust visual tracking. *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125-141.

Schroff, F.; Kalenichenko, D.; Philbin, J. (2015): Facenet: a unified embedding for face recognition and clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815-823.

Simonyan, K.; Zisserman, A. (2014): Very deep convolutional networks for large scale image recognition. arXiv:1409.1556.

Tao, R.; Gavves, E.; Smeulders, A. W. M. (2016): Siamese instance search for tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1420-1429.

Valmadre, J.; Bertinetto, L.; Henriques, J. F.; Vedaldi, A. (2017): Torr PHS. End-to-end representation learning for correlation filter based tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2805-2813.

Vinyals, O.; Kaiser, L.; Koo, T.; Petrov, S.; Sutskever, I. et al. (2015): Grammar as a foreign language. *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2773-2781.

Wang, L.; Ouyang, W.; Wang, X.; Lu, H. (2015): Visual tracking with fully convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3119-3127.

Wang, L.; Ouyang, W.; Wang, X.; Lu, H. (2016): STCT: sequentially training convolutional networks for visual tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1373-1381.

Wang, Q.; Gao, J.; Xing, J.; Zhang, M.; Hu, W. (2017): DCFNet: discriminant correlation filters network for visual tracking. arXiv:1704.04057.

Wu, Y.; Lim, J.; Yang, M. (2013): Online object tracking: a benchmark. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2411-2418.

Wu, Y.; Lim, J.; Yang, M. (2015): Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834-1848.

Wu, Z. C.; Valentini-Botinhao, Watt O. S.; King, S. (2015): Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, pp. 4460-4464.

Yun, S.; Choi, J.; Yoo, Y.; Yun, K.; Choi, J. Y. (2017): Action-decision networks for visual tracking with deep reinforcement learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2711-2720.

Zhang, J.; Ma, S. (2014b): MEEM: robust tracking via multiple experts using entropy minimization. *Proceedings of the European Conference on Computer Vision*, pp. 188-203.

Zhang, T.; Xu, C.; Yang, M. H. (2017): Multi-task correlation particle filter for robust object tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4335-4343.

Zhang, Y. H.; Wang, Q. Q.; Li, Y. L.; Wu, X. D. (2018): Sentiment classification based on piecewise pooling convolutional neural network. *Computers, Materials & Continua*, vol. 56, no. 2, pp. 285-297.

Zhong, W.; Lu, H. C.; Yang, M. (2014): Robust object tracking via sparse collaborative appearance model. *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2356-2368.

Zhou, S. R.; Liang, W. L.; Li, J. G.; Kim, J. K. (2018): Improved VGG model for road traffic sign recognition. *Computers, Materials & Continua*, vol. 57 no. 1, pp. 11-24.