An Investigation of Deep Tracking Methods

Zan Huang
School of Computer Science
Beijing Institute of Technology
Beijing, China
dreaming_hz@hotmail.com

***Abstract*—The emerging technology revolution empowered by artificial intelligence related research has brought us closer to the implementation of systems only seen in science-faction movies before. Artificial vision systems built in complex modern artefacts are very important for supporting their execution and perception based on camera captured raw pixels is one of the keys to open the machine intelligence. Deep learning and visual object tracking are both hot topics for computer vision research in recent years. Many visual object tracking algorithms based on deep learning technique appeared. The effectiveness of deep features for visual tracking has been shown on benchmarks, these big guys showed more intelligent behaviour capturing target object in bounding box but the high computational cost and deep hungry for data contradicts our initial requirements for practical tracking algorithm. In this paper, we investigated several deep trackers to help discovering current trends in this research area and hope it could help researchers and practitioners to better understand and further enhance existing methods.**

***Index Terms*—Deep learning, Visual object tracking.**

I. Introduction

VISUAL object tracking algorithm is essential for many computer vision applications. Trackers serves in a wide range of systems including video surveillance, self-driving cars and human computer interaction. For single-object tracking, trackers need locating the object of interest marked in initial video frame continuously. Given a bounding box, best tracker need to overcome challenges like fast motion, serious occlusion, object deformation and achieve real-time performance with high precision.

To win the challenge, many tracking algorithms have been proposed including ones with powerful deep learning engines over the past decade. There is a trend to exploit automatically extracted deep features for visual tracking task and construct an end-to-end system to track. While high computational cost and execution speed hold trackers back from practical use, problems like over-fitting also perplex these approaches.

In this paper, we focus on single-object tracking, the context of this task could be described as: taking a video sequence and a bounding box localising the target object in the initial frame as input, implement an algorithm to generate consecutive bounding boxes as output.

Building a brain by programming then utilise it to do anything like visual object tracking is very straightforward idea but not realistic enough in current time to our knowledge. The popular deep learning methodology provide an replacement to put that idea into practice to some extent and there are more and more deep trackers appearing.

Many existing deep trackers are based on tracking-by-detection framework and most of them are constructed with pretrained artificial neural network prepared for image recognition and object detection task. We picked representative ones for investigating, further analysis and discussion has been presented in following sections and some trackers discussed in this paper are listed in Table I for easier reference.

The overwhelming number of parameters and a bunch of selectable structures support building computer vision system in the way resembles playing lego blocks. Great flexibility with many reusable components inspires us to try variety of methods to reconfigure existing trackers, the potential of deep tracking methodology has not been fully released and it would be fun to further exploit this research topic.

II. Background

Visual object tracking has been an active research area for decades, various approaches have been attempted and there are many trackers and experimental results available for reference. But the problem named visual object tracking has not been solved completely till now, in fact, the problem consists of various sub-problems liking occlusion and motion blur, overwhelming data represented by raw pixels in videos is difficult to deal with for a simple computational model. Deep learning has been deeply investigated by researchers and widely used in various computer vision tasks since the winning of AlexNet [29] on imagenet competition in 2012. Availability of large volume datasets [30], open source toolkits [31], [32] and publicly pre-trained models [33] have accelerated computer vision research. It was the best of times, you can do things that could not be done in data and computation power constraint, it was the worst of times, you could not even totally understand the method you adopted in experiments. Especially for visual object tracking and deep learning, both of which are application driven research topics, the detailed discussion of current methods would help discovering facts that have not been mentioned in former works yet.

A. The Visual Tracking Challenge

In current visual tracking benchmarks [34], [35], the typical input data for the algorithms including video frames and a bounding box marked the target on the first frame. Techniques like stereo vision could not be applied on such benchmark. Meanwhile, the total information given by a bounding box could be represented by four numbers: x, y, width and

TABLE I
TRACKER LIST

Year	Tracker
2013	DLT [1]
2014	DeepTrack [2]
2015	CNN-SVM [3], SO-DLT [4], FCNT [5], DeepSRDCF [6], CF2 [7],
2016	MDNet [8], STCT [9], HDT [10], RTT [11], SINT [12], GOTURN [13], SiameseFC [14], DeepTracking [15], ROLO [16], DMSRDCF [17], Learnet [18], CNT [19], TCNN [20], C-COT [21]
2017	DRT [22], DNT [23], SANet [24], ECO [25], CFNet [26], ADNet [27], ACFN [28]

height(and rotation angle in some cases), putting a visual tracker in seriously constrained context. At the same time, to meet requirements coming from real applications, a visual tracker do need to be model-free, fast, accurate, robust which means it has to overcome challenges like serious occlusion, fast motion, camera shifting [35]. The tracker is required to learn the action of tracking without knowing which object the target is, it should be able to track any possible object outside the training video sequences. We are giving seriously constrained information to algorithms and expect they could give out ideal performance, trackers do need to carry extra information inside its built-in models in this situation and deep neural network models are pretty good choice in recent years.

B. Power of Deep Learning

Due to theoretical and technical reasons, although the artificial neural network has been used for object tracking back to 1990s [36], the popularity of deep learning in visual object tracking research field is the most recent thing. The practice of this popular connectionism approach renamed deep learning is far from trivial, availability of large volume data, powerful parallel computing frameworks and hardware, easy to use toolkits, pretrained models and bag of tricks are all constraints for the success of applying a deep model to solve a challenging real-world problem. At the same time, the various architecture to learn from unstructured data in deep learing toolbox makes it very rewarding to try a deep titled algorithms for classical problems.

Visual object tracking especially single camera single object short-term visual object tracking algorithms runs in seriously constrained context, the tracker need information or even kind of intelligence to catch the target by bounding box, it do not need to know exactly which object it is at the same time, leaving the task for visual recognition systems. Luckily, there are some existing works attempting to build tracker using deep features and progress has been made in recent years.

C. Related Work

There are many ways to deal with the visual object tracking challenge, both low-level vision and high-level vision techniques could be applied for this application driven task. Approaches like superpixel tracking [37], tracking by segmentation [38], tracking by utilising the saliency map [3] have all been tested in the past. Tracking-by-detection [39] is most commonly adopted approach in few recent years, the representative trackers we picked mostly belong to this category. Before the wining of AlexNet [29] on ImageNet [30] 2012, convolutional neural network has already been used for human tracking task [40], the authors proposed the method to use a two convolutional layer for human tracking. Due to trends and availability of data and computational resources, the DLT [1] tracker was the frist one to integrate a deep model into visual tracking framework for general single object tracking to our knowledge. By merging the deep autoencoder into the particle filter framework, the proposed tracker is competitive to many state-of-art trackers at that time. Several important points was emphasised in [1] including using large volume dataset to pretrain the model, this work left enough space for later deep trackers as the more complex models like deep convolutional neural networks, recurrent neural networks has not been used yet while popular theories like deep reinforcement learning has not been covered too.

Accuracy was the first break-point for demonstrating power of deep trackers. As demonstrated in CF2 [7] algorithm. Different layers tend to encode the semantic information from different levels, low-level features may be embedded in first several layers of the artificial neural network while high-level features could be captured in deeper ones. Combining the functioning of low-level feature for localisation and high-level feature for distinguishing the target object from the background is effective for creating a state-of-art tracker. Smaller network [8] cross-trained on two commonly adopted benchmarks while eliminating the shared video sequence during training would generate a more accurate tracker with mulit-domain learning, hard negative mining and bounding box regression used, an obvious silhouette of object detection algorithm could be seen from such work.

An end-to-end system for visual tracking is more suitable for releasing the power of deep learning techniques. Like SINT [12] which adopted siamese network to fully exploit the usage of initial bounding box for visual tracking, this approach owns a pretty concise framework.

Speed became the focus of deep tracking algorithms later, an adventurous approach named GOTURN [13] directly trained a network to generate bounding box parameters to efficiently localise object. The network structure is similar to siamese networks despite the conjunction of the output of two separate convolutional components by connecting them to fully connected layers, it requires large volume of data for training while the speed is nearly 100 FPS empowered by GPU which is really a positive speed acceleration.

Fast real-time trackers built with deep features is feasible

and the parameters of the network even could be generated by one-shoot learning as demonstrated by learnet [18], the tracker was a variant of another siamese network tracker [14] proposed by the same group while the performance of learnet is inferior comparing to newly appeared trackers, the idea behind it is inspiring. Many correlation tracker also adopted deep features for object tracking but the speed is slowing down by referencing to recent works, but ECO [25] have accelerated the speed of these kinds of approaches drastically, high speed running of convolutional operators on CPU has been achieved. Also, the motion feature would be helpful for object tracking and recurrent neural networks are expected to be better at handling sequential data like videos. How to teach a tracker to ignore concrete details of the object and track it via motion information with help of temporal information may be the next move in this research area.

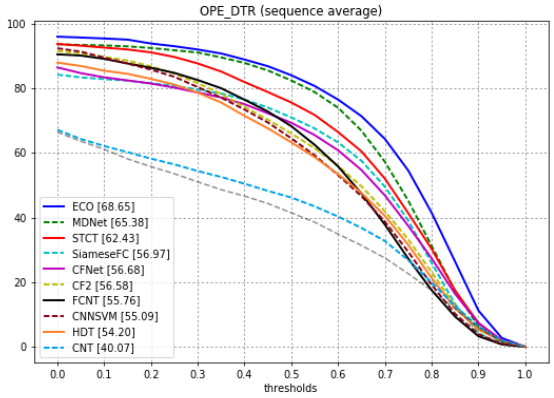

Fig. 1. The success plot of OPE test on OTB-100 [35] benchmark.

III. Trends and Facts of Current Deep Tracking

There is always a tradeoff in experiment while researchers sometimes tend to emphasis the advantages of proposed algorithm and leaving the drawbacks out. Hands on testing and cross comparison of various algorithms could help us get more information about the algorithms.

A. Experimental Analysis

For better understanding of current deep tracking algorithms, we collected some publically available deep trackers code and tested on the widely used visual tracking benchmark [35]. Due to page size limits, just one figure was included in this paper 1 which shows the success rate varies with overlap threshold. It is a commonly adopted plot for tracker comparison and the AUC(Aera Under Curve) score is the key for referring tracker performance, the higher it is, the better the tracker would be according to the evaluation rules of the benchmark [35].

The evaluation results shown in 1 is approximately same to results reported in original papers for each visual tracker. ECO [25] is the best performer in our experiment and deep trackers are made to be faster, more accurate, simpler each year by efforts of researchers. At the same time, we found that the benchmark has not been fully utilised for tracker evaluation in most circumstances, the SRE and TRE test on [35] were not adopted for recent deep trackers, it is partially because of slowness of some deep trackers for which it may take several days to run the experiment for a single tracker, it may reveal the fact that some algorithms are selectively using benchmarks for evaluation at the same time. Fully utilising object tracking benchmarks may help better analysis of current algorithms. And some benchmarks like PTB [41] waits to be exploited for deep trackers. Meanwhile, the pipeline of visual trackers is becoming more complex with deep learning introduced into this field. Faster, light-weight algorithms would be more popular.

B. Current State of Deep Tracking Research

There are several common to-do list items for deep tracking community and the public pipeline for carrying next generation deep tracker seems intuitive. First, there were experiments for testing usability of deep learning for visual tracking [1], [40]. Later the focus has been how to utilise convolutional neural network for extracting robust features for visual tracking and further improve tracker performance. Inevitably, popular research topics such as recurrent neural network and deep reinforcement learning would be introduced into this field. The current focus has been deep reinforcement learning and there has been some works [27] published. Unsupervised learning topics such as generative adversarial network maybe seen in the near future in visual tracking research community.

C. Unexplored Topics

The application of recurrent neural networks in visual tracking has not shown the superiority over other methods yet, some of them are too slow [24] and some algorithms has not been tested on real-world data [15]. A better way to learn temporal information for creating better tracker waits to be discovered.

Visual tracking is an application driven research field and more applications are in need of better tracking system like drone control [42], [43]. At the same time, few trackers have the knowledge of real-3D-world, most trackers are processing matrices which are data mapping from real world to 2D arrays. Teaching the algorithm about the physics of real-world would definitely enhance tracker reliability and make use of unlabelled data for tracker training at the same time as demonstrated in [44].

Bounding boxes are raw representation of target object location, the lack of context information may lead to difficulty in solving the visual tracking problem. The more accurate groundtruth segmenting target from the background given by VOT [34] challenge since 2016 would be meaty for training trackers. Meanwhile, the conjunction with natural language processing field could also help with visual tracking research as demonstrated in [45].

IV. Conclusion

Tracking is middle level computer vision task, but current trackers are actually mixtures of a hierarchy of computer vision techniques. Current works are pushing the visual tracking research forward. Meanwhile, when building up a visual tracker using deep features, sometimes we are reversely using high-level vision product for middle-level vision task due to the advanced success of deep learning in those fields. Visual object tracking is a different challenge comparing to object recognition and detection and further exploration of deep tracking would help us stepping further on the way of creating intelligent vision system.

References

[1] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Advances in Neural Information Processing Systems 26*, 2013.

[2] H. Li, Y. Li, and F. Porikli, "Deeptrack: Learning discriminative feature representations by convolutional neural networks for visual tracking," in *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014*, 2014.

[3] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," pp. 597–606, 2015.

[4] N. Wang, S. Li, A. Gupta, and D.-Y. Yeung, "Transferring rich feature hierarchies for robust visual tracking," *arXiv preprint arXiv:1501.04587*, 2015.

[5] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3119–3127.

[6] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 58–66.

[7] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3074–3082.

[8] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," *arXiv preprint arXiv:1510.07945*, 2015.

[9] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Stct: Sequentially training convolutional networks for visual tracking." CVPR, 2016.

[10] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, and J. L. M.-H. Yang, "Hedged deep tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[11] Z. Cui, S. Xiao, J. Feng, and S. Yan, "Recurrently target-attending tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1449–1458.

[12] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016.

[13] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 fps with deep regression networks," in *European Conference Computer Vision (ECCV)*, 2016.

[14] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 850–865.

[15] P. Ondruska and I. Posner, "Deep tracking: Seeing beyond seeing using recurrent neural networks," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, 2016, pp. 3361–3368.

[16] G. Ning, Z. Zhang, C. Huang, Z. He, X. Ren, and H. Wang, "Spatially supervised recurrent convolutional neural networks for visual object tracking," *arXiv preprint arXiv:1607.05781*, 2016.

[17] S. Gladh, M. Danelljan, F. S. Khan, and M. Felsberg, "Deep motion features for visual tracking," *CoRR*, 2016.

[18] L. Bertinetto, J. F. Henriques, J. Valmadre, P. H. S. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016.

[19] K. Zhang, Q. Liu, Y. Wu, and M.-H. Yang, "Robust visual tracking via convolutional networks without training," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1779–1792, 2016.

[20] H. Nam, M. Baek, and B. Han, "Modeling and propagating cnns in a tree structure for visual tracking," *CoRR*, vol. abs/1608.07242, 2016.

[21] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," *arXiv preprint arXiv:1608.03773*, 2016.

[22] J. Gao, T. Zhang, X. Yang, and C. Xu, "Deep relative tracking," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1845–1858, 2017.

[23] Z. Chi, H. Li, H. Lu, and M. Yang, "Dual deep network for visual tracking," *IEEE Transaction on Image Processing*, vol. 26, pp. 2005–2015, 2017.

[24] H. Fan and H. Ling, "Sanet: Structure-aware network for visual tracking," *CoRR*, vol. abs/1611.06878, 2016.

[25] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: efficient convolution operators for tracking," *CoRR*, 2016.

[26] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," *CoRR*, 2017.

[27] J. C. Sangdoo Yun, Y. Yoo, K. Yun, and J. Y. Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in *CVPR*, 2017.

[28] J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, and J. Y. Choi, "Attentional correlation filter network for adaptive visual tracking," in *CVPR*, 2017.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," pp. 1097–1105, 2012.

[30] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Feifei, "Imagenet: A large-scale hierarchical image database," pp. 248–255, 2009.

[31] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[32] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," in *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[34] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," 2016.

[35] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[36] S. J. Nowlan and J. C. Platt, "A convolutional neural network hand tracker," in *Advances in Neural Information Processing Systems 7*, 1995, pp. 901–908.

[37] S. Wang, H. Lu, F. Yang, and M.-H. Yang, "Superpixel tracking," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1323–1330.

[38] A. Milan, L. Leal-Taix, K. Schindler, and I. Reid, "Joint tracking and segmentation of multiple targets," in *CVPR*, 2015.

[39] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2012.

[40] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *IEEE Transactions on Neural Networks*, vol. 21, no. 10, pp. 1610–1623, 2010.

[41] S. Song and J. Xiao, "Tracking revisited using rgbd camera: Unified benchmark and baselines," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[42] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2016.

[43] S. Li and D.-Y. Yeung, "Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models," in *AAAI*, 2017.

[44] R. Stewart and S. Ermon, "Label-free supervision of neural networks with physics and domain knowledge," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, 2017, pp. 2576–2582.

[45] Z. Li, R. Tao, E. Gavves, C. G. M. Snoek, and A. W. Smeulders, "Tracking by natural language specification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.