Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Visual Object Tracking Using Deep Learning Techniques: A Comparison

Enzheng Su*

Sendelta International Academy Shenzhen, Shenzhen, China

*Corresponding author: Wilsonsu_academic@hotmail.com

Abstract. Visual object tracking is one of the important topics in computer vision: given a target object in the first frame of a video, follow this target object on subsequent frames. Its applications are found in autonomous driving, human-computer interaction, military operations, and game development. However interesting these applications may be, the implementation remains very challenging because occlusions happen among other factors like scale variations or rotations that require a tracker to be robust and generalize well to give accurate results. In this paper, we provide an extensive review of visual object tracking methods divided into three primary frameworks: correlation filter-based trackers, Siamese network-based trackers, and Transformer-based trackers. We provide insights into a total of 12 state-of-the-art techniques for each category through large-scale experimentation on various benchmark data sets, aiming to reveal the research frontier of the visual object tracking field and point out some potentially useful directions for future investigations.

Keywords: Visual Object Tracking, Deep Learning, Correlation Filter, Siamese Network.

1. Introduction

In recent years, with the emerging development of deep learning-based methods [1-4], many methods with novel backbones have been proposed, such as AlexNet [5], VGG [6], ResNet [7], DenseNet [8], and the Inception [9] network. Specifically, visual object tracking is key to a large number of computer vision applications where it acts as a fundamental technology, including but not limited to autonomous driving, surveillance systems, and human-computer interaction. The problem is defined as being able to spot an object in successive frames given its initial frame location. Despite considerable improvements in recent years, visual tracking remains quite challenging due to several factors affecting target appearance: (1) Occlusion: when target objects are partially or fully occluded by another scene object; (2) Changes in illumination: lighting variations that alter the subject's appearance; (3) Motion blur: caused by rapid movement of the camera or target object; and (4) Deformations: changes in the shape or form of the object.

To overcome these challenges, researchers have come up with tracking algorithms of increased sophistication. The evolution of these algorithms can be generalized to have passed through three phases: (1) Correlation Filter-based Trackers: Initiated by Bolme et al. in 2010 through the development of the Adaptive Correlation Filter (ACF), this marked a key breakthrough in the field. ACF performs its operations by correlating the target features from the first frame with those from the current frame. This produces a response map, with the maximum value location indicating the estimated target position. (2) Siamese Network-based Trackers: A significant improvement was realized in 2016 following Bertinetto et al.'s introduction of SiamFC. It extended the correlation filter idea by replacing correlations with a convolutional neural network. This innovation enabled end-to-end training and had exploitation capabilities over large-scale tracking datasets— aspects that brought about great improvements in performance while keeping track speeds high. (3) Trackeri Basati su Trasformatori: Di recente, l'architettura Transformer, che ha raggiunto un successo notevole in svariati ambiti dell'IA, è stata adattata per il tracking visivo di oggetti. TransT, presentato da Chen et al. nel 2021, è stato uno dei primi lavori a combinare la struttura Transformer con il tracking di oggetti. Esso sostituiva le operazioni convoluzionali nei network di Siamese con meccanismi di self-attention e cross-attention derivati dal modello Transformer.

Intending to give a brief introduction to visual object tracking, the paper reviews all recent developments in this domain with a focus on deep learning-based approaches. We categorize tracking methods into three main frameworks: the correlation filter-based trackers, Siamese network-based trackers, and Transformer-based trackers. For each category, we analyze representative algorithms by discussing their motivations, methodologies, and advantages. This is done in the hope of obtaining an analysis that will provide a clear vision of the current state of visual object tracking and interesting research directions yet to be explored.

2. Deep Learning-based Visual Object Tracking

2.1. Correlation Filter-based Trackers

Bolme et al. [10] in 2010 introduced the Adaptive Correlation Filter (ACF), a notable improvement in visual object tracking. Most traditional tracking methods found it difficult to be real-time trackers and to adapt with changing appearances of the objects. ACF resolved these challenges in tracking by proposing a new formulation for the tracker. The ACF technique functions by learning a very simple correlation filter that produces a sharp peak at the location of the target object and low responses elsewhere for background clutter. This is updated by an efficient method in Fourier domains, hence making it possible for real-time monitoring. The correlation process can be formulated as:

$$G = F \odot H^* \tag{1}$$

where $F = \mathrm{F}(f)$ and $H = \mathrm{F}(h)$, $\mathrm{F}(\cdot)$ represents Fast Fourier Transform, while the G is the response map in Fourier domain and $\odot$ denotes the element-wise multiplication. This approach demonstrated impressive speed and accuracy, capable of tracking hundreds of frames per second while maintaining competitive performance on benchmark datasets. The success of ACF paved the way for further developments in correlation filter-based tracking methods.

Following the success of correlation filter-based approaches, K. Henriques et al. introduced the Kernelized Correlation Filter (KCF [11]) in 2015. KCF tackled a key issue that had plagued earlier trackers: redundant training samples. For correlation filters, this meant a decrease in training efficiency. There were two key innovations that KCF brought about: Circulant matrix formulation: KCF cast the tracking problem into circulant matrices. This enables the realization of fast computations in Fourier domains; thus, it drastically reduces computation complexities during the tracking process.Application of the kernel trick: By adopting kernel regression, KCF increased the discriminative power of features while retaining very high tracking speeds. This allows for a more complex relationship to be captured between data in a non-linear fashion. The integration of these innovations enabled KCF to reach state-of-the-art performance on the OTB50 dataset while keeping its tracking speed well over 100 frames for seconds. This exceptional accuracy and pace propelled KCF into benchmark status within the field, further galvanizing research into correlation filter-based tracking.

In 2014, Li and Zhu introduced the Scale Adaptive Mean Shift (SAMF [12]) tracker to some existing trackers, an important issue: accurate scale estimation. At that time, most of the trackers had a common problem of their performance degrading when the size varies, which is a free phenomenon in actual tracking scenarios.

SAMF integrates the advantages of correlation filter-based tracking within a scale space. The key contributions can be summarized as follows,

1. Multi-scale search: SAMF considers a pool of scaled versions of the target template and performs joint estimation of position and scale.

2. Mean-shift refinement: Following initial localization based on correlation, SAMF applies the mean-shift algorithm for refining target location and scale.

3. Adaptive scale estimation The tracker updates its scale space filter continuously, responding to changes in the size of the target.

This is what enabled SAMF to outperform many other contemporary trackers on standard benchmarks with large-scale variations while keeping computation reasonable.

In 2015, Danelljan et al. introduced the Spatially Regularized Discriminative Correlation Filter (SRDCF [13]) to solve one of the most basic limitations of correlation filter-based trackers; boundary effects. It limited the ability to use large search areas because correlation filters are traditionally said to have such effects as a consequence of making an assumption about training samples being periodic in nature.

SRDCF adds a spatial regularization component to the correlation filter formulation. It is shown that this regularization is analogous to penalizing the location of filter coefficients in the spatial domain, thus allowing larger training and search regions without sacrificing accuracy. The SRDCF approach can be formulated as:

$$S_f(x) = \sum_{l=1}^{d} x^l * f^l \tag{2}$$

$$\varepsilon_t(f) = \sum_{k=1}^{t} \alpha_k \left|\left|S_f(x_k) - y_k\right|\right|^2 + \lambda \sum_{l=1}^{d} ||f^l|| \tag{3}$$

where Sf(x) represents the correlation scores, εt(f) is the objective function to be minimized. w is the spatial regularization weight. xl and fl are the feature channels and corresponding filter channels. αk and λ are weight parameters.

This innovation significantly improved tracking performance, especially in scenarios with fast motion and large appearance changes. By allowing for larger search areas, SRDCF could handle more challenging tracking scenarios while maintaining real-time performance.

2.2. Siamese Network-based Trackers

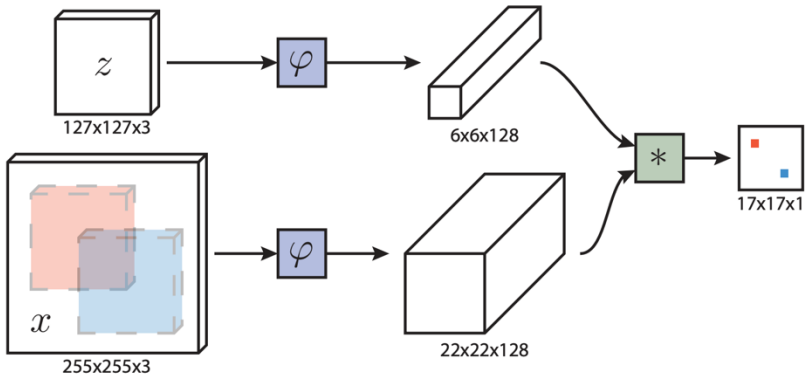

Figure 1. Overall framework of SiamFC.

In 2016, SiamFC was introduced by Bertinetto et al. [14], following up on prior work in visual object tracking approaches (see in Figure 1). The main idea behind SiamFC is to leverage deep learning-based methodologies to enhance tracking performance, achieve high speeds, and address the limitations of correlation filter-based methods in handling complex appearance variations.

SiamFC deploys a fully convolutional Siamese architecture consisting of two identical branches as follows: (1) Template branch, and (2) Search branch. This paper encodes target object appearance from the initial frame. The search branch deals with the search area across subsequent frames. It is the overall optimization over large-scale data that allows it to learn a general similarity function between a template and possible target locations.

Key steps in the SiamFC tracking process are:

(1) Feature extraction: Pass target template and search area through convolutional layers to extract deep features.

(2) Cross-correlation: Cross-correlate the extracted features to generate a similarity score map.

(3) Target localization: Identify the new target location as the position with the highest similarity score.

SiamFC offers several key advantages: (1) Strong performance on multiple benchmarks, (2) Real-time tracking speeds (86 fps on GPU), (3) Good generalization capabilities due to offline pre-training on large datasets, and (4) Robustness to appearance changes.

In demonstrating that it is possible for SiamFC to balance accuracy and speed while offering an approach for end-to-end learning in visual tracking, this further elicited a wave of enthusiasm for Siamese network-based tracking methods.

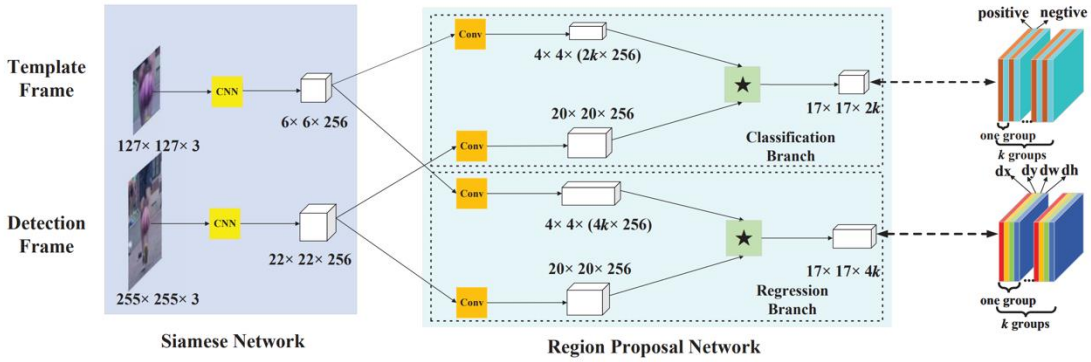

Figure 2. Overall framework of SiamRPN.

Following the SiamFC, in 2018 Li et al. introduced SiamRPN [15], basing it on the framework of Siamese tracking and incorporating detect object ideas. The SiamRPN net is shown in Figure 2. Although the SiamFC proved to be quite effective, it had no dedicated component for scale estimation; multi-scale testing had to be applied, hence decreasing the efficiency.

SiamRPN is applying an RPN to the Siamese tracking framework. The key components of SiamRPN are: (1) Siamese subnetwork; much like in SiamFC, it extracts features for both the target template and search region. (2) Region Proposal Network; this network produces a set of candidates along with corresponding objectness scores. (3) Classification branch; it classifies the proposed regions into target or background. (4) Regression branch: Further refines bounding box coordinates for localization.

Here is a brief of the SiamRPN tracking pipeline:

(1) Siamese subnetwork for feature extraction. (2) RPN for proposal generation. (3) Proposal classification and bounding box regression. (4) Proposals scoring. to have the final tracking output.

This approach attained the best reported performance in accuracy and speed on multiple benchmarks, pushing it to state-of-the-art by achieving 160 frames per second and thus suitable for real-time applications. Integrating RPN with SiamRPN allows the latter to generate high-quality region proposals that get further refined into final tight bounding boxes. The two-stage nature of this process enhances robustness with regard to challenges such as scale variations and partial occlusions, in the visual object tracking field.

In 2017, Valmadre et al. [16] introduced CFNet in an attempt to marry correlation filter- and deep learning-based methods for tracking. The primary motivation of the work is to exploit the strengths of both techniques since correlation filters are fast and can adapt online while deep neural networks offer a powerful feature representation. CFNet innovates by introducing, at its core, a Siamese network architecture into which a correlation filter layer is placed. Making it possible to learn end-to-end optimal features for correlation filter-based tracking. CFNet's main components are:

(1) Siamese network—for extracting features from both the target template and search region. (2) Correlation filter layer—that makes up the trained optimal correlation filter. (3) Correlation operation—that carries out the application of the learned filter on the search region features.

CFNet tracking proceeds from extracting features of the target template and search region by the Siamese network, through applying the correlation filter to search region features, creating a response map indicating probable target positions, and then setting the location with the maximum response as the new target position.

CFNet has the following key advantages: (1) Performance is better than that of traditional correlation filter trackers and original SiamFC. (2) This high maintenance of tracking speeds is because of efficient correlation operations. (3) More generalization to unseen objects is possible through end-to-end training. (4) More robustness to appearance changes is achieved by deep feature plus correlation filter combination.

CFNet showed that hybrid tracking techniques could be developed by integrating the merits of both methodologies, design, and execution.

2.3. Transformer-based Trackers

The development of Transformer-based architectures in visual object tracking is the most recent stride. While native to tasks based on natural language processing, Transformer models have shown excellent performance when applied to vision tasks, among them object tracking.

One of the earliest works in this direction is TransT [17] by Chen et al. 2021. The Transformer structure is applied to tracking by replacing the Siamese correlation operators with self-attention and cross-attention mechanisms based on those defined for the Transformer model.

The feature extraction backbone of TransT uses Convolutional Neural Network to extract features from both the template and search regions. Transformer encoder: To apply self-attention for improving feature representations independently for both template and search regions. Transformer decoder: Herein, cross-attention is used to merge information coming from the template and search regions. Prediction head: It generates the final tracking result by basing this on the fused features.

The TransT tracking process is described as below: (1) Backbone for template and search region feature extraction. (2) Transformer encoder for further improving feature representation. (3) Transformer decoder for information-based interaction between the template and search region. (4) Prediction head generating the final tracking result.

Transformer-based trackers offer several significant advantages, such as:

(1) Long-range dependency in both spatial and temporal dimensions. (2) Better feature representation. (3) Occlusions and appearance changes are learned by the model due to global context consideration i.e. information from all pixels of the image that also leads to its high interpretability. (4) Performance among the state-of-the-art levels across multiple benchmark datasets, often outperforming correlation filter-based as well as Siamese network-based approaches

Yet Transformer-based trackers also bring about new challenges that have to be addressed, especially in computational complexity and requirement of vast amounts of training data. With the evolution of research in this area, more improvements can be expected regarding efficiency and the effectiveness of Transformer-based tracking methods.

3. Experimental Comparison

3.1. Dataset Description

We evaluated various tracking methods using two widely recognized datasets: OTB (Object Tracking Benchmark [18]) and LaSOT (Large-Scale Single Object Tracking [19]).

The OTB dataset comprises OTB-2013 and OTB-2015, with 50 and 100 sequences, respectively, making a total of about 59,000 frames. Challenges include but are not limited to: Variation in illumination Scale Occlusion Deformation Motion blur Fast motion In-plane rotation Out-of-plane rotation Out-of-view Background clutter Low resolution The evaluation is done using both success plot and precision plots: Success Plot (SP): Overlap between predicted and ground truth bounding boxes Precision Plot (PP): Measures the center location error between predicted and ground truth bounding boxes

LaSOT has 1,400 sequences amounting to 3.52 million frames at an average sequence length of 2,512 frames and class number 70. It retains the challenges similar to OTB but with longer sequences and more diverse object categories. Success rate is given in the evaluation metrics as the percentage of frames that are successfully tracked; Precision Rate, i.e., the percentage of frames when the distance between the predicted and ground-truth center is within some threshold; and Normalized Precision Rate (NPR) similar to PR but normalized by target size. These datasets have provided such a holistic evaluation environment setting our ability to make assessment over different challenges for long-term tracking scenarios on trackers' performance.

3.2. Comparison Results

Here we present a comparison of the tracking methods discussed in this paper, evaluated on the OTB and LaSOT datasets:

Table 1. Comparison results on OTB and LaSOT datasets.

Method	**OTB**		**LaSOT**			**Speed (fps)**
	SP	PP	SR	PR	NPR	
ACF	0.58	0.78	-	-	-	292
KCF	0.62	0.85	0.18	0.17	0.20	172
SAMF	0.64	0.78	-	-	-	7
SRDCF	0.67	0.89	0.26	0.25	0.29	5
SiamFC	0.69	0.88	0.34	0.33	0.38	86
SiamRPN	0.73	0.95	0.45	0.43	0.49	160
SiamFC++	0.75	0.97	0.54	0.54	0.60	90
TransT	0.81	0.97	0.64	0.69	0.73	50

Referring to Table 1, several trends can be seen. One is the drastic improvement in tracking performance from earlier methods (ACF, KCF) to the more recent ones (SiamFC++, TransT). This indicates a fast-increasing development rate in the field of visual object tracking. (2) While the older methods such as ACF and KCF provide very high frame rates, they do not track with great accuracy. Newer methods achieve a much better accuracy at some marginal cost in reducing speed. (3) The first appearance of Siamese network-based methods (SiamFC, SiamRPN, SiamFC++) realized a large increase regarding both accuracy and speed when contrasted against traditional correlation filter-based methods. (4) Among Transformer-based trackers up to this point, TransT demonstrates the best overall performance on both datasets but especially on the more complicated LaSOT dataset. But it is at a lower frame rate than some Siamese network-based methods. (5) Additionally, results on LaSOT are consistently lower than those on OTB, indicating the higher complexity of the LaSOT dataset because of its longer sequences and larger variety of object classes.

These results show how far visual object tracking has advanced, where every new tracker is evaluated on how much it can address the limitations of the existing state-of-the-art as well as push the boundary for better performance.

4. Conclusion

This paper provides a very detailed review of recent methods of visual object tracking, which may mostly be classified into three primary frameworks—correlation filter-based trackers, Siamese network-based trackers, and Transformer-based trackers. We discuss 12 influential methods from these categories and perform their performance analysis on two benchmark datasets: OTB and LaSOT. The paper discusses each method's parameter tuning and implementation details for the reported experiments so that the results could be reproduced by other researchers. Additionally, for some of the methods, we have also reported novel experimental results, showing that the reported state-of-the-art performances can be further improved.

Our analysis places TransT, which is a Transformer-based framework, as the most accurate at present. But Siamese network-based approaches—especially SiamRPN—provide the best speed-accuracy trade-off and are thus more suitable for real-time applications.

Future directions of visual object tracking include applications in: (1) Autonomous driving, wherein the robustness and speed of the visual object trackers, especially handling the occlusions and tracking during nighttime, need to be improved; (2) Game design, where the developed object tracker should be capable of tracking any general object, anywhere, to improve generalization across different game environments; and (3) Human-computer interaction, so that a tracker may track arbitrary objects and provide feedback for robotic decision-making assistance. In the future, research directions in visual object tracking are as follows: (1) Transformer-based methods need to be dealt with performance improvement to realize real-time performance. (2) More robust trackers need to be developed that can handle very extreme appearance changes and long-term occlusions. (3) There should be an investigation on ways to decrease the data dependency of deep learning-based trackers—probably through self-supervised methods or even few-shot learning.

We look forward to enhancements in precision tracking, velocity, and robustness as further fine-tuned with time in this ever-dynamic field—inevitably ushering myriad fresh use case scenarios across sectors.

References

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[3] Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., ... & Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR), 51(5), 1-36.

[4] Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. IEEE Access, 7, 53040-53065.

[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[6] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778)

[8] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[9] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[10] Bolme, D. S., Beveridge, J. R., Draper, B. A., & Lui, Y. M. (2010, June). Visual object tracking using adaptive correlation filters. In 2010 IEEE computer society conference on computer vision and pattern recognition (pp. 2544-2550). IEEE.

[11] Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2014). High-speed tracking with kernelized correlation filters. IEEE transactions on pattern analysis and machine intelligence, 37(3), 583-596.

[12] Li, Y., & Zhu, J. (2015). A scale adaptive kernel correlation filter tracker with feature integration. In Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II 13 (pp. 254-265). Springer International Publishing.

[13] Danelljan, M., Hager, G., Shahbaz Khan, F., & Felsberg, M. (2015). Learning spatially regularized correlation filters for visual tracking. In Proceedings of the IEEE international conference on computer vision (pp. 4310-4318).

[14] Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. In Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14 (pp. 850-865). Springer International Publishing.

[15] Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018). High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8971-8980).

[16] Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., & Torr, P. H. (2017). End-to-end representation learning for correlation filter based tracking. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2805-2813).

[17] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., & Lu, H. (2021). Transformer tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8126-8135).

[18] Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: A benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2411-2418).

[19] Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., ... & Ling, H. (2019). Lasot: A high-quality benchmark for large-scale single object tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5374-5383).