Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/82/20240997

Advancements in deep learning for visual object tracking

Jingyi Dai

Beijing-Dublin International College, Beijing University of Technology, Beijing, China

200162@yzpc.edu.cn

Abstract. Since contemporary information-retrieval systems rely heavily on the content of titles and abstracts to identify relevant articles in literature searches, great care should be taken in constructing both. This comprehensive review delves into the transformative impact of deep learning on the domain of visual object tracking. Since the inception of AlexNet in 2012, deep learning has revolutionized feature extraction, leading to significant advancements in tracking accuracy and robustness. The article explores the integration of deep learning with various tracking algorithms, including deep correlation filters, classification-based approaches, Siamese networks, gradient-based methods, and the innovative application of Transformer architectures. Moreover, the role of tracking datasets in fostering algorithm development and innovation is highlighted, with an emphasis on the expansion in scale, diversity, and annotation quality. Furthermore, the article also examines the multifaceted evaluation metrics for tracking algorithms, encompassing precision, robustness, efficiency, generalization, and real-time capabilities. Looking ahead, the review outlines future research directions, such as algorithm optimization for lightweight and accelerated performance, enhancing generalizability, leveraging multimodal data fusion, and refining Transformer models for improved temporal information processing. The challenges of long-term tracking and the growing importance of algorithm interpretability and transparency are also discussed. In summary, the article underscores the promising trajectory of deep learning in visual object tracking, with ongoing research poised to make tracking technologies smarter, more efficient, and robust, catering to a wide array of practical applications and environments.

Keywords: Deep Learning, Visual Object Tracking, Feature Extraction, Tracking Algorithms, Deep Correlation Filters.

1. Introduction

Deep learning techniques have revolutionized the field of visual object tracking since AlexNet's breakthrough in image classification in 2012. The core strength of deep learning lies in its exceptional feature extraction capabilities, which have become particularly evident in object tracking. By utilizing deep convolutional neural networks, researchers can automatically learn multi-level feature representations that range from raw pixels to high-level semantics. These features are then employed for the accurate representation and recognition of targets.

Compared to traditional hand-crafted feature extraction methods, deep learning offers stronger expressiveness and robustness, thereby significantly enhancing tracking performance. End-to-end deep tracking frameworks further simplify the tracking process by directly learning the target's dynamic

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/82/20240997

model from input video frames, eliminating the need for multiple independent steps in the traditional tracking pipeline.

Over the past few years, deep learning has fostered the emergence and development of various classic tracking algorithms. For instance, the HCF algorithm combines deep features with correlation filters, achieving multi-scale feature fusion; the MDNet algorithm treats object tracking as a binary classification problem, learning robust feature representations through multi-domain training; and the SiamFC algorithm employs a Siamese network framework, generating a response map through related operations for efficient tracking [1][2].

Additionally, deep learning has driven improvements in traditional tracking algorithms by integrating deep features into correlation filters, which enhances the robustness of these algorithms in complex scenarios. Moreover, with continuous innovation in network structures and optimization strategies—such as using residual networks to deepen the network architecture or adopting attention mechanisms to improve model focus—deep learning applications in object tracking have become more diverse and efficient.

To address the storage and computational demands of deep learning models, model compression and acceleration techniques like knowledge distillation and network pruning have been widely applied to maintain tracking performance while reducing resource consumption. Moreover, unsupervised and semi-supervised learning methods have shown great potential in object tracking, enabling effective feature learning and tracking with limited labeled data[3].

The application of deep learning in multimodal tracking is an emerging research direction, involving the fusion of data from different sensors or modalities to improve tracking accuracy and robustness. Furthermore, the introduction of the Transformer architecture has brought a new perspective to visual object tracking, with its attention mechanism capable of fully leveraging temporal information to provide richer dynamic representations for tracking models.

Despite the remarkable achievements of deep learning in visual object tracking, researchers continue to explore ways to further enhance algorithm efficiency, interpretability, and generalization capabilities. As technology continues to advance, this paper have reason to believe that future tracking technologies will become more practical, efficient, reliable, and versatile, meeting a variety of tracking scenarios and demands.

2. Classification and Development of Tracking Algorithms

The application of deep learning technology in the field of visual object tracking has propelled the classification and development of various tracking algorithms. These algorithms are categorized into different types based on their design philosophies and implementation methods, each demonstrating unique advantages in performance and efficiency.

Deep correlation filter tracking algorithms combine features extracted by deep learning with traditional correlation filter algorithms, either by integrating them or through end-to-end joint training of deep networks and correlation filters. These algorithms leverage the powerful representational capabilities of deep features. At the same time, they maintain the high computational efficiency of correlation filters in the frequency domain. Figure 1 shows an example of deep correlation filter tracking [4][5].

DOI: 10.54254/2755-2721/82/20240997

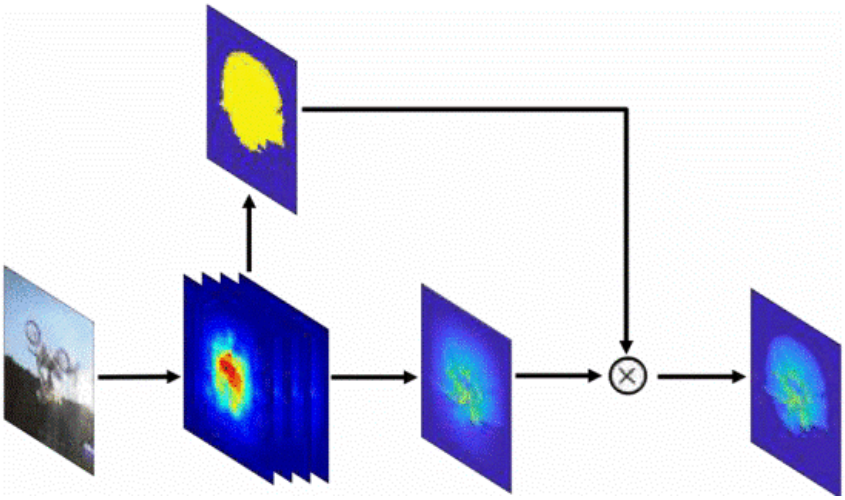

Figure 1. Spatial reliable map construction from the fifth convolutional layers in deep correlation filter tracking [4].

Classification-based tracking algorithms treat the tracking task as a binary classification problem, distinguishing between the target object and the background in each frame. These algorithms typically use pre-trained convolutional layers to extract general features and train fully connected layers to adapt to specific scenes for target discrimination. Although these algorithms have made progress in accuracy, the process of online model fine-tuning also presents challenges in terms of efficiency.

Siamese network tracking algorithms as shown in Figure 2 regard the tracking task as a template matching problem, searching for the most similar candidate sample to the initial target in each frame. Algorithms like SiamFC are widely noted for their concise and efficient framework. However, due to the potential neglect of background information, subsequent research has enhanced the model's ability to discern distractors by incorporating correlation filters [6].

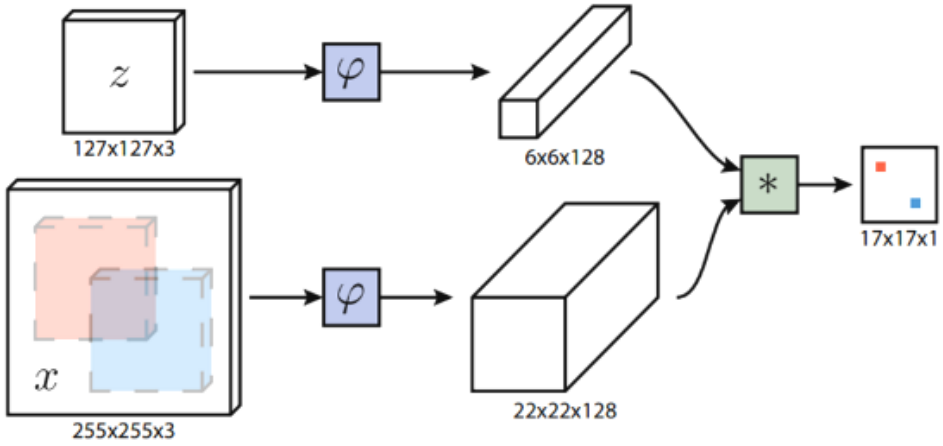

Figure 2. Fully-Convolutional Siamese Networks for Object Tracking [6]

Gradient-based tracking algorithms use gradient descent methods to quickly solve the regression problem in tracking, thereby obtaining tracking models with discriminative power. Inspired by correlation filters, these algorithms solve for filters with the ability to distinguish between the foreground and background, leveraging background information to improve the model's ability to discern distractors. Moreover, recently, deep trackers based on the Transformer architecture have emerged, utilizing attention mechanisms for tracking model modeling and achieving leading performance.These methods,

DOI: 10.54254/2755-2721/82/20240997

which leverage attention mechanisms to utilize temporal information or model the tracker, have shown great potential in handling tracking tasks [7].

As deep learning technology continues to advance, these tracking algorithm frameworks continue to evolve, enhancing the accuracy, robustness, and real-time performance of tracking through various innovative methods. Although existing frameworks still have room for improvement in some aspects, the application of deep learning in visual object tracking is promising, and future research will continue to drive technological progress in this field.

In summary, the application of deep learning in the field of visual object tracking has propelled the classification and development of various tracking algorithms. These algorithms are categorized into different types based on their design philosophies and implementation methods, each demonstrating unique advantages in performance and efficiency. Deep correlation filter tracking algorithms combine deep learning-extracted features with traditional correlation filter algorithms, leveraging the powerful representational capabilities of deep features while maintaining the high computational efficiency of correlation filters in the frequency domain.

3. The development trend of the tracking dataset

In recent years, tracking datasets have played an essential role in the field of visual object tracking. They provide not only a standard platform for evaluating algorithm performance but also greatly promote model training and algorithm innovation. With the introduction of deep learning techniques, the scale and diversity of datasets have increased significantly. Large-scale datasets such as TrackingNet and LaSOT offer tens of thousands of videos, covering a variety of target categories and scenarios, providing ample material for the training of deep learning models. Additionally, to enhance the generalization capabilities of algorithms and address complex scenes, new datasets continuously add more challenging video sequences. These include those with high dynamic ranges, rapid motion, and complex backgrounds.

Besides expanding in scale and diversity, the quality of dataset annotations has also been improving. Detailed labeling information such as precise bounding boxes, occlusion states, and motion patterns provide more training signals for algorithms. Moreover, the development of cross-modal datasets has provided a platform for handling data from multiple sensors. These datasets integrate information from various sensors, including RGB, depth, and infrared, advancing the study of multimodal tracking algorithms. As the importance of real-time tracking becomes increasingly evident, some datasets like NFS offer high frame rate videos to assess the real-time performance of algorithms, and some datasets also provide interactive tools to help researchers analyze and understand tracking scenarios in depth [7][8].

The internationalization and openness of datasets have also facilitated academic exchanges and technological advancements worldwide. Researchers have started to pay attention to ethical and privacy issues in dataset collection and use, ensuring that they comply with legal regulations and ethical standards. These rich datasets enable researchers in the field of visual object tracking to design and test more robust, accurate, and practical tracking algorithms to meet the growing industrial and commercial needs. With the continuous development of technology, this paper can anticipate that future tracking datasets will be more diverse, refined, and challenging, providing a stronger impetus for the development of tracking algorithms.

In recent years, tracking datasets have played an essential role in the field of visual object tracking, providing a standard platform for evaluating algorithm performance and greatly promoting model training and algorithm innovation. With the introduction of deep learning techniques, the scale and diversity of datasets have increased significantly. Large-scale datasets such as TrackingNet and LaSOT offer tens of thousands of videos, covering a variety of target categories and scenarios.

4. Evaluation of algorithm performance

Algorithm performance evaluation is a crucial measure of the progress in visual object tracking research. It provides a comprehensive reflection of an algorithm's capabilities and adaptability across various

DOI: 10.54254/2755-2721/82/20240997

complex scenarios. To thoroughly assess the performance of tracking algorithms, researchers have adopted a multi-dimensional set of evaluation metrics. These metrics encompass not only traditional precision and success rates but also robustness, efficiency, generalization ability, and real-time performance.

Cross-dataset evaluations have tested the performance of algorithms across different scenarios and challenges, with datasets that encompass a wide range of video sequences and object categories designed to pose varying degrees of difficulty in tracking tasks, such as rapid motion, scale variation, occlusion, illumination changes, and background interference [9].

Furthermore, in challenging scenario analyses, the robustness of algorithms is rigorously tested, especially their performance when faced with rapid motion or scale changes of the target.Real-time tracking performance assessments focus on the speed of response of the algorithm in practical applications, which is particularly important for scenarios requiring swift processing. Additionally, user studies and feedback provide a qualitative perspective on algorithm evaluation, helping to understand user satisfaction and expectations regarding tracking outcomes.

The interpretability of an algorithm has emerged as a novel dimension in evaluation, involving the transparency and comprehensibility of the decision-making process of the algorithm, aiding in building user trust and further optimization of the algorithm. Long-term tracking capabilities are assessed to determine an algorithm's ability to maintain stable tracking over extended sequences, especially when the target undergoes significant changes. Model compression and acceleration technologies enable the deployment of algorithms in resource-constrained environments, which is crucial for real-time tracking on mobile or embedded devices.

In terms of multimodal data fusion, the evaluation examines how algorithms integrate data from various sensors and how such integration enhances tracking accuracy and robustness. These evaluation methods not only reveal the strengths and limitations of existing algorithms but also guide the future development of tracking technologies, propelling visual object tracking technology towards greater precision, robustness, and efficiency. With continuous improvement and innovation in evaluation methods, this paper can anticipate even more outstanding performance of tracking algorithms in practical applications, better meeting a variety of application needs [10].

Algorithm performance evaluation is a crucial measure of the progress in visual object tracking research, providing a comprehensive reflection of an algorithm's capabilities and adaptability across various complex scenarios. To thoroughly assess the performance of tracking algorithms, researchers have adopted a multi-dimensional set of evaluation metrics, including traditional precision and success rates, as well as robustness, efficiency, generalization ability, and real-time performance.

5. Future research directions

Future research directions in the field of visual object tracking are multifaceted, encompassing the refinement of algorithms, optimization of models, and exploration of new theoretical avenues. Building on past advancements, as deep learning technology continues to evolve, future research will continue to push the boundaries of tracking algorithms to adapt to more complex and dynamic visual environments. On one hand, future studies will focus on the lightweight and acceleration of algorithms by designing more efficient network architectures and optimization algorithms to reduce computational resource requirements. This includes not only the development of new network structures to reduce the number of parameters but also the exploration of dynamic network adjustment strategies to maintain tracking accuracy while increasing algorithm speed.

At the same time, researchers are exploring how to enhance the generalizability of algorithms, enabling them to adapt to various unknown environments and targets. This may involve research into cross-domain adaptability and how to learn effectively from limited data. Additionally, reinforcement learning and other machine learning techniques may be integrated into tracking algorithms to improve their decision-making capabilities in complex environments.

The integration of multimodal data will also become a hot research direction. Moreover, with the advancement of sensor technology, how to effectively combine information from different sensors, such

DOI: 10.54254/2755-2721/82/20240997

as vision, infrared, radar, etc., to achieve more accurate and robust tracking results will be an important topic.

Moreover, the potential of Transformer models in visual object tracking will continue to be explored. Researchers will investigate how to improve the Transformer structure to better handle temporal information and enhance its performance in long-term tracking. This includes designing new attention mechanisms and how to adapt the Transformer model to target appearance changes.

The challenges in long-term tracking, such as long-term appearance changes, occlusions, and out-of-view problems, will also be key focuses of future research. Researchers will seek new methods for modeling target appearances and developing effective update strategies to address these challenges.

Finally, with the widespread application of artificial intelligence technology, the interpretability and transparency of tracking algorithms will also receive more attention. Researchers will strive to enhance the interpretability of algorithms, allowing users to understand and trust the decision-making processes of AI systems.

Considering all these aspects, in summary, future research will continue to advance the development of visual object tracking technology, making it smarter, more efficient, more robust, and capable of meeting the growing demands of various applications and providing reliable performance in a range of practical environments.

6. Conclusion

Our survey provides a comprehensive review of the application and development of deep learning technology. It specifically focuses on the field of visual object tracking. Since the breakthrough of AlexNet in image classification in 2012, deep learning has significantly enhanced the performance of object tracking with its powerful feature extraction capabilities. In this context, the article discusses in detail the various applications of deep learning in tracking algorithms, including deep correlation filter tracking, classification-based tracking, Siamese network tracking, gradient-based tracking, and Transformer-based tracking algorithms. With the increase in the scale and diversity of datasets, as well as the improvement in the quality of annotations, rich resources have been provided for algorithm training and innovation. In addition, the article explores the multidimensional metrics for evaluating algorithm performance, including accuracy, robustness, efficiency, and real-time performance, and emphasizes the importance of cross-dataset evaluations and user studies. Looking to the future, research will continue to advance the development of tracking algorithms. This will be achieved by focusing on algorithm lightweighting, acceleration, enhancing generalizability, multimodal data fusion, and improving Transformer models. Meanwhile, challenges in long-term tracking, such as appearance changes, occlusions, and out-of-view problems, as well as the interpretability and transparency of algorithms, will also be key research focuses. Overall, deep learning has shown tremendous potential and prospects in the field of visual object tracking, and future research will further promote technological progress to meet the growing industrial and commercial needs.

References

[1] J. Gao et al. ‘Recursive Least-Squares Estimator-Aided Online Learning for Visual Tracking’, 2024 IEEE Trans. Pattern Anal. Mach. Intell. 46 1881-1897

[2] B. Babenko, M.-H. Yang and S. Belongie, ‘Robust Object Tracking with Online Multiple Instance Learning’, 2011 IEEE Trans. Pattern Anal. Mach. Intell. 33 1619-1632

[3] T. Wang et al. ‘GSC: A Graph and Spatio-Temporal Continuity Based Framework for Accident Anticipation’, 2024 IEEE Trans. Intell. Veh. 9 2249

[4] Wang Ning et al. ‘Recent advance in deep visual object tracking’, 2021 J. Univ. Sci. Technol. China 51 335-344

[5] Z. Wang et al. ‘A Dynamic Model-Based Doppler-Adaptive Correlation Filter for Maritime Radar Target Tracking’, 2024 IEEE Trans. Geosci. Remote Sens. 62 5101415

DOI: 10.54254/2755-2721/82/20240997

[6] B. Babenko, M.-H. Yang and S. Belongie 2009 'Robust Object Tracking with Online Multiple Instance Learning' In: 2009 IEEE Conference on Computer Vision and Pattern Recognition Miami 983-990
[7] Zhang K, Zhang L and Yang M H, 'Real-Time Compressive Tracking',2012 In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y and Schmid C (eds) Computer Vision – ECCV 2012 Lecture Notes in Computer Science 7574 (Berlin: Springer)
[8] D.A. Ross, J. Lim, R.S. Lin et al. 'Incremental Learning for Robust Visual Tracking', 2008 Int. J. Comput. Vis. 77 125-141
[9] S. Hare et al. 'Struck: Structured Output Tracking with Kernels', 2016 IEEE Trans. Pattern Anal. Mach. Intell. 38 2096-2109
[10] K. Kalal, K. Mikolajczyk and J. Matas, 'Tracking-Learning-Detection', 2012 IEEE Trans. Pattern Anal. Mach. Intell. 34 1409-1422