Visual object tracking: Progress, challenge, and future

Libo Zhang[1,3,*] **and Heng Fan**[2]

[1]Institute of Software, Chinese Academy of Sciences, Beijing 100190, China

[2]Department of Computer Science and Engineering, University of North Texas, Denton, TX 76206, USA

[3]Nanjing Institute of Software Technology, Nanjing 210000, China

*Correspondence: libo@iscas.ac.cn

Received: January 10, 2023; Accepted: February 16, 2023; Published Online: February 21, 2023; https://doi.org/10.1016/j.xinn.2023.100402

Citation: Zhang L. and Fan H. (2023). Visual object tracking: Progress, challenge, and future. The Innovation **4(2)**, 100402.

Visual object tracking aims to continuously localize the target object of interest in a video sequence. As one of the most fundamental problems in computer vision, visual object tracking has a long list of critical applications including video surveillance, autonomous driving, human-machine interaction, augmented reality, robotics, etc., in which the tracking system provides the capacity to report target positions in real time for subsequent visual analysis. In the past decades, visual object tracking has been extensively explored and has witnessed considerable progress, especially in deep-learning-based tracking. Despite this, robust tracking remains challenging due to many factors. To provide the community an overview, in this commentary, we will discuss visual tracking from different aspects. Specifically, we will first summarize the recent advancements achieved in visual tracking from the perspectives of algorithm and dataset. Then, we will analyze the challenges that the tracking community faces in developing practical tracking systems. Finally, we will discuss several promising directions for future research on visual tracking. It is worth noting that visual object tracking is a broad problem and consists of many specific topics. In this commentary, we focus on the most popular single-modality (i.e., RGB), bounding-box-based tracking. Figure 1 illustrates the task of visual tracking and the organization of this commentary.

PROGRESS ON VISUAL OBJECT TRACKING

Visual object tracking has been largely studied in the past decades. Earlier hand-crafted approaches mainly focus on designing discriminative classification models and/or robust appearance features for tracking. Nevertheless, these trackers usually suffer from various appearance variations under complex scenarios, which limits their performance in achieving high accuracy. Motivated

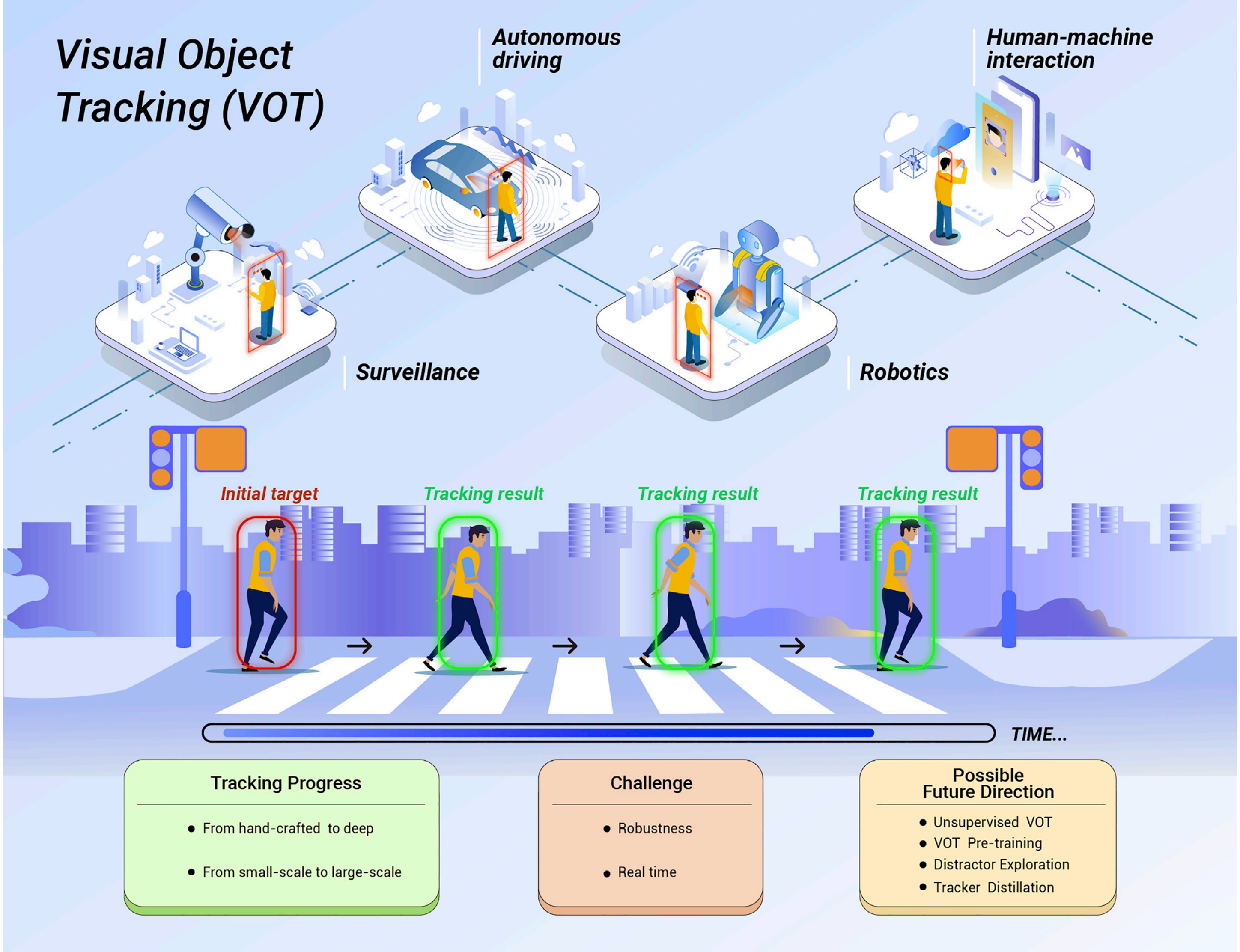

Figure 1. Illustration of visual object tracking and summary of its progress, current challenges, and potential future directions

by deep learning, researchers in the tracking community have leveraged deep neural networks (DNNs) for visual tracking in recent years. Specifically, these trackers propose to extract deep features from pre-trained networks and then employ more robust deep representations for tracking, demonstrating superior performance compared with traditional hand-crafted tracking models. Despite this, earlier deep trackers often undergo a heavy computational burden because they require updating the deep model frequently in an online manner, which may result in slow running speed. Addressing this issue, researchers later formulate the tracking task as a forward matching problem. The main idea is to directly leverage DNNs for learning a generic similarity measurement, which matches the designated target object in the initial frame with the current search region to achieve object tracking. As no model update is required once the training completed, these trackers are able to run efficiently in real time with a video graphics card or GPU. Because of the good balance between accuracy and speed, this matching-based tracking framework has become the major trend in visual tracking with many extensions for further improvements. More recently, due to the capacity in modeling global contextual relation, the Transformer architecture[1] has been introduced to object tracking, which greatly improves the object appearance representation and further pushes the frontier of tracking accuracy. Nowadays, Transformer almost becomes a necessity for state-of-the-art trackers.

In addition to the tracking algorithm, the tracking dataset has seen remarkable progress in recent years. The tracking dataset serves an important role in advancing visual tracking. Previous datasets are usually utilized for fair evaluation and comparison of different trackers and are therefore small in scale. One representative is the VOT Challenge,[2] which has introduced a series of competitions such as short-term tracking since 2013. Owing to its popularity, the VOT Challenge has attracted numerous participants every year and become a commonly used evaluation dataset. In the deep-learning era, previous small-scale datasets are not suitable for training of deep tracking. As a consequence, researchers have been forced to utilize the video data from other fields for training, which degrades the tracking performance due to domain gap. In addition, these small-scale datasets may bring bias in performance evaluation and comparison. In order to mitigate these issues, a few large-scale datasets have been proposed in recent years, which supports large-scale learning of deep trackers. The proposal of large-scale datasets significantly improves the tracking performance. Besides the training purpose, these large-scale datasets provide large-scale evaluation for visual trackers, which, to some extent, can reduce the bias in evaluation and thus faithfully reflect the capacity of tracking algorithms.

Besides the above discussion, more progress and details can be found in surveys. The survey of Li et al.[3] summarizes and compares deep tracking algorithms with extensive experiments. The survey of Marvasti-Zadeh et al.[4] comprehensively discusses different deep trackers, benchmarks, evaluations, etc. A more recent one[5] focuses on discussing long-term tracking including algorithms, datasets, and experiments. We encourage readers to refer to these surveys for more details.

CHALLENGE FOR PRACTICAL TRACKING SYSTEM

Despite decades of research on visual tracking, a practical tracking system for real-world applications remains difficult due to many challenges. In general, these challenges can be categorized into two types: robustness related and efficiency related.

The robustness challenge requires that a visual tracker achieves high-accuracy results. However, due to many factors such as occlusion, deformation, fast motion, motion blur, rotation, out of view, distractor, scale variation, and illumination change in the video sequence, the target appearance may suffer from severe changes, which results in gradual drift and even tracking failure. Although the usage of deep features for tracking alleviates these issues, the tracking performance is still degenerated, especially in the presence of occlusion, out of view, and distractor. In both occlusion and out of view, the target object may completely disappear. Since the target can move anywhere, it is hard to accurately re-locate the object, and tracking failure may happen. When distractors that are visually similar to the target exist, the tracker may drift to the background distractor region because the tracking model heavily relies on appearance information for target localization. Besides the above factors, adversarial attack may pose another threat to tracking robustness. The attack model adds imperceptible noise to the video frames to fool the visual tracker, causing tracking failure.

In addition to the robustness requirement, the other challenge for a practical tracker is high efficiency. As a video task, real-time efficiency is critical for visual tracking. Although most current deep-tracking approaches are able to run in real-time speed, they require huge computing resources such as CPUs and GPUs. However, when deployed on mobile devices, which do not have powerful CPUs and GPUs, these tracking algorithms may suffer from heavy time delays in reporting target position, limiting their applications. To address this problem, more efforts are required for improving tracking efficiency on mobile or edge devices while maintaining the accuracy.

FUTURE RESEARCH DIRECTION

Despite considerable progress in recent years, many problems are unsolved in visual object tracking. In the following, we will discuss several potential research directions for visual tracking.

First, one of the promising directions is to develop unsupervised tracking models. Currently, deep trackers usually require a large set of labeled videos for training. However, the annotation of these videos is expensive and time consuming. Especially, as the model increases in the future, more labeled training data are desired, which may significantly hinder further development of visual tracking. Addressing this, a potential solution is to develop unsupervised tracking models that can automatically learn from videos without human labels. Recently, several attempts have been made for unsupervised tracking. However, the performance of these unsupervised trackers falls far behind the supervised visual trackers. Further study is needed in investigating unsupervised visual tracking. Second, it is worthy to explore an effective pre-training strategy for tracking. Existing tracking models often leverage the pre-trained image classification model for training. However, due to domain gap, it may not be optimal to use the parameters of the image classification model. Instead, a dedicated generic pre-trained tracking model is needed. Self-supervised learning approaches can be borrowed to pre-train a universal large-scale tracking model. When developing new trackers, the pre-trained tracking model can be directly adopted for feature extraction without fine-tuning, simplifying the pipeline for new algorithm design. Third, it is crucial to exploit distractor information in videos for tracking. Currently, most trackers aim to localize the target of interest while ignoring visually similar distractors in the videos, resulting in drift or even failure. To alleviate this issue, a future direction is to simultaneously locate the target object and similar distractors to provide more information for distinguishing the target from the background. Finally, in order to improve the inference efficiency of visual trackers, a feasible solution is network distillation. Network distillation aims to transfer crucial knowledge of the large-scale teacher network to a small-scale student network. For visual tracking, given a trained teacher tracker, our goal is to learn a student tracker that performs similarly during inference while running much faster.

REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. ***30***, 5998–6008.
2. Kristan, M., Matas, J., Leonardis, A., et al. (2016). A novel performance evaluation methodology for single-target trackers. IEEE Trans. Pattern Anal. Mach. Intell. ***38***, 2137–2155.
3. Li, P., Wang, D., Wang, L., et al. (2018). Deep visual tracking: review and experimental comparison. Pattern Recogn. ***76***, 323–338.
4. Marvasti-Zadeh, S.M., Cheng, L., Ghanei-Yakhdan, H., et al. (2022). Deep learning for visual tracking: a comprehensive survey. IEEE Trans. Intell. Transport. Syst. ***23***, 3943–3968.
5. Liu, C., Chen, X.F., Bo, C.J., et al. (2022). Long-term visual tracking: review and experimental comparison. Mach. Intell. Res. ***19***, 512–530.

ACKNOWLEDGEMENTS

Libo Zhang was supported by the Key Research Program of Frontier Sciences, CAS, Grant No. ZDBSLY- JSC038, CAAI-Huawei MindSpore Open Fund and Youth Innovation Promotion Association, CAS (2020111).

DECLARATION OF INTERESTS

The authors declare no competing interests.