Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 4, No. 3, 2023

Survey of AR Object Tracking Technology Based on Deep Learning

Kaihao Xu *

Zhuhai College of Science and Technology, Zhuhai, 519040, China
*** Corresponding author Email:** kcx8821@gmail.com

Abstract: Augmented reality (AR) is the process of "augmenting" computer information into the real or physical world. This helps us better understand the process and has numerous application-level AR has an almost unlimited range of applications in today's world, and if implemented, will help solve many complex problems in a simple way. This paper presents a comprehensive survey of the different types of tracking technique. As used to develop object tracking algorithms for AR applications. This paper compares and analyses different types of sensor-based and vision-based tracking technologies, and finally draws conclusions. The mixed use of different types of tracking technologies is one of the best solutions to achieve accurate and powerful tracking to meet the strict requirements of AR applications.

Keywords: Visual Object Tracking; Deep Neural Network; Generative Adversarial Networks.

1. Introduction

The demand for the interaction between virtual Reality and the real world has gradually increased, and the Augmented Reality (AR) technology has emerged. Augmented reality combines computer-generated virtual information with the real-world environment to provide users with a new type of interactive experience that is interesting and practical. Object tracking is one of the core technologies in many AR application scenarios. By detecting and tracking objects in the real environment in real time, object tracking realizes the accurate correspondence between virtual information and the real world. AR object tracking technology based on deep learning has shown strong competitiveness in practical applications due to its high accuracy, robustness and real-time performance. This paper aims to investigate and compare the existing AR object tracking technologies based on deep learning, in order to provide a comprehensive and systematic reference for research in related fields.

This paper aims to provide a high-quality research review for researchers by classifying and analysing the object tracking algorithms in the AR field, so as to help them better conduct further research. The structure of this paper mainly includes the following sections:

The first chapter of this paper first introduces the application and direction of tracking technology in the field of AR, reviews the development process of AR object tracking technology, combs the traditional tracking technology and its limitations, and lays a foundation for the application of deep learning technology in the field of AR tracking. In Chapter 2, a general overview of the development, status and classification of visual object tracking technology is given. The third chapter compares and analyses the visual multi-object tracking methods with excellent performance in public datasets in recent years. The fourth chapter looks forward to the possible future research directions based on the previous discussion.

In conclusion, this paper comprehensively and systematically reviews the research and comparison of AR object tracking technologies based on deep learning from many aspects, and provides valuable reference and guidance for researchers and engineers in related fields.

2. The Exploration of Tracking Technology in the AR Field

2.1. The Development of AR

Augmented reality is the real-time interaction between the real world and the virtual world, that is, accurate three-dimensional registration [1, 2].

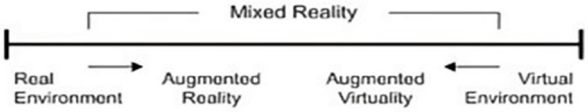

Figure 1. Figure with short caption (caption centred).

Today's mobile augmented reality systems typically use multiple motion tracking technologies, aiming at tracking the user's head position and orientation, while tracking the user's hand or handheld input device can provide more accurate 6-DOF interaction techniques. In addition, many different sensors cooperate with each other to form hybrid tracking. [2] Hybrid tracking is a head pose tracking technology that merges or fuses the output pose data from different types of trackers to form a hybrid tracker to achieve higher accuracy and stronger tracking robustness. Although the AR system using hybrid tracking technology will increase a certain complexity, it can effectively maintain high-precision pose tracking and enhance the robustness of tracking.

2.2. Application of Tracking Registration and Computer Vision in AR Field

The commonly used methods include tracker-based registration, machine vision-based tracking registration, and hybrid tracking registration technology based on wireless network. The trackers-based registration technology uses a pre-trained tracker to track the movement of the object, so as to achieve tracking registration. Machine vision-based tracking and registration techniques use computer vision algorithms to extract the feature points or contours of the

object and track these feature points or contours for registration. The hybrid tracking and registration technology based on wireless network fuses multiple tracking technologies to improve the accuracy and robustness of tracking and registration.

3. Overview of Target Tracking Algorithms based on Deep Learning

3.1. Development History of Tracking Algorithms

Traditional target tracking algorithms are the earliest algorithms, including optical flow method, Kalman filter, particle filter, mean shift and so on, and they laid for the vigorous development of the target tracking field, these algorithms can be divided into generative model and discriminative model. Since the generative model only focuses on the object itself, the defect of this method is particularly obvious when the object undergoes huge deformation or change. Unlike generative models, discriminative models take into account not only the object information but also the background information. Discriminative models treat the object tracking problem as a classification or regression problem, and the goal is to find a discriminant function to distinguish the object from the background in order to track the object. Therefore, discriminative models pay more attention to how to extract features from the input data to better realize the classification or regression of the target. In object tracking, discriminative models can track the target by learning a classifier or regressor, which can achieve better results in practical applications.

Discriminative methods are more robust and gradually occupy the mainstream position in object tracking algorithms. At present, many deep learning objects tracking algorithms also belong to the discriminative framework.

During the period of 2013 to 2022, the model of the target tracking algorithm has undergone a transformation from the traditional feature-based method to the deep learning-based method, and the fusion method is gradually used to improve the performance and robustness of the tracking. The continuous evolution and development of these methods provide strong support for the application and development of target tracking technology.

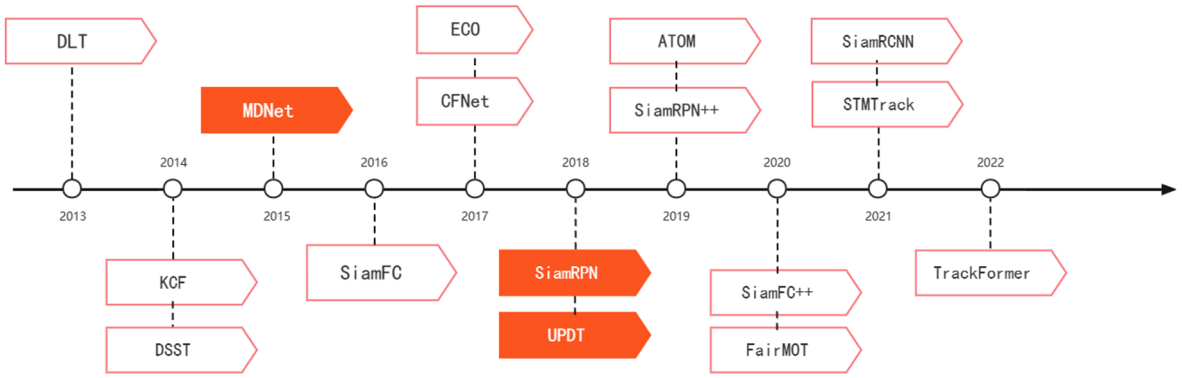

Figure 2. Representative target tracking algorithms generated from 2013 to 2022

3.2. Research Status of Deep Learning-based Tracking Algorithms

Deep Learning-based Tracking Algorithms (DLTA) have received extensive attention in recent years. This method uses deep learning technology to extract and characterize the features of the target object, so as to achieve accurate tracking of the target. The main aspects are as follows:

Feature extraction and representation: Effective feature representation is the key to achieve object tracking. Researchers have tried to use deep learning models such as convolutional Neural Network (CNN) and recurrent neural Network (RNN) to extract and represent the features of the target object. These features can include the shape, texture, motion and other information of the object.

Data association and filtering: During tracking, it is necessary to determine the data association of the object between adjacent frames. Data association methods include matching, Kalman filter, particle filter and so on. Deep learning can improve the accuracy of data association by learning the complex relationship between the target and the background.

Online learning and updating: Online learning and model updating are important parts of the tracking algorithm. Researchers adopt an end-to-end training strategy to enable the model to adapt to the changes of the target in real time during the tracking process.

The main algorithm models are as follows:

Siamese network: The Siamese network is a commonly used tracking algorithm model, which uses two convolutional neural networks with the same structure to extract the features of the target and the search area, and then determines the location of the target by comparing these features. Representative algorithms include SiamFC and SiamRPN.

Rnn-based tracking algorithms: These algorithms exploit the temporal characteristics of RNN to capture the dynamic changes of the target in the video sequence. For example, the MDNet algorithm uses RNN for multi-domain feature extraction and combines an online learning strategy to achieve real-time tracking.

Integrated object detection and tracking: This type of method unifies the problem of object detection and tracking, and uses an end-to-end training strategy for joint optimization. For example, D&T algorithm combines Faster R-CNN with Kalman filter to realize target detection and tracking.

4. Comparison of Tracking Algorithms based on Deep Learning

4.1. Visual SLAM based Methods

In augmented reality applications, it is necessary to accurately align virtual objects with the real scene to ensure the correct presentation and interactive experience of virtual objects. At the same time, since AR applications usually need

to sense and track object s and camera positions in the scene in real time, an efficient, accurate and real-time method is required to accomplish this task. In this case, the method based on SLAM (Simultaneous Localization and Mapping) can be used as an effective solution, which use deep learning to improve the accuracy of visual Simultaneous localization and mapping (SLAM) algorithms. SLAM algorithms are used to track the position and orientation of devices in the real world. Deep learning can improve the accuracy of SLAM algorithms by recognizing objects in the environment and tracking their movements.

LIFT-SLAM combines deep learning-based feature descriptors with traditional geomete-based systems. And it extends the pipeline of the ORB-SLAM system to use CNNS to extract features from images, providing denser and precise matching based on the learned features. The figure below compares LIFT-SLAM with various SLAM algorithms on the KITTI dataset.

Table 1. Formatting sections, subsections and subsubsections

Algorithm	**Type**	**Metric**	**00 01 02 03 04 05 06 07 08 09 10**
LIFT- SLAM	Hybrid	ATE (m) RPE_{trans} (%) RPE_{rot} (deg/ m)	8.06 X 40.04 2.23 0.51 13.55 30.38 3.63 184.43 59.62 29.87 3.18 X 8.73 1.46 2.22 6.09 12.24 2.42 47.10 19.91 9.72 2.99 X 2.49 0.34 0.48 3.11 2.91 4.02 2.02 2.14 2.24
ORB- SLAM*	Traditional	ATE (m) RPE_{trans} (%) RPE_{rot} (deg/ m)	11.54 X X 15.13 4.29 7.74 20.26 13.47 39.51 49.67 19.94 4.46 X X 9.75 3.71 3.35 8.11 7.43 12.16 26.51 8.65 3.28 X X 2.78 2.15 3.57 2.88 3.58 3.05 11.13 3.62
ORB-SLAM''	Traditional	ATE (m) RPE_{trans} (%) RPE_{rot} (deg/ m)	5.33 X 21.28 1.51 1.62 4.85 12.34 2.26 46.68 6.62 8.80 -
DeepVO**	End-to-end	ATE (m) RPE_{trans} (%) RPE_{rot} (deg/ m)	- - - - - - - - - - - 8.49 7.19 2.62 5.42 3.91 - - 8.11 - - 6.89 6.97 3.61 5.82 4.60 - - 8.83
NeuralBundler	Hybrid	ATE (m) RPE_{trans} (%) RPE_{rot} (deg/ m)	- - - - - - - - - - - 3.24-4.85-1.83, 2.74-6.23-3.53 1.35-1.60 -- 0.7 2.6 2.02-2.11 -

4.2. Based on Object Detection and Tracking Algorithms

In the field of AR, especially the behaviours and operations that pursue real-time performance and accuracy, such as interaction, games, fast tracking and positioning, etc., all aim to improve the user experience, which especially needs the support of excellent algorithms. The following is the performance of various state-of-the-art VOT methods for single object tracking on four data sets until recently. In Masked Autoencoders with Spatial-Attention Dropout for Tracking Tasks [3], the masked Autoencoder (MAE) was pre-trained on the video for VOT and VOS. Their DropMAE is a video frame reconstruction method that aims to solve the problem of excessive spatial cues and ignoring temporal relationships in traditional MAE (Mean Absolute Error) methods. This method facilitates temporal correspondence learning in videos by adaptively performing spatial attention dropout during frame reconstruction.

Table 2. Comparison with state-of- theart VOT methods on four large-scale challenge datasets. Data from [3]

Method	Source	AO	GOT -10k[4]		T NL2K [5]		LaSOT $_{ext}$ [6]			AUC	LaSOT [7]	P
			$SR_{0.5}$	$SR_{0.75}$	AUC	P	AUC	P_{Norm}	P		P_{Norm}	
SiamFC	ECCVW16	34.8	35.3	9.8	29.5	28.6	23.0	31.1	26.9	33.6	42.0	33.9
MDNet	CVPR16	29.9	30.3	9.9	-	-	27.9	34.9	31.8	39.7	46.0	37.3
ECO	ICCV17	31.6	30.9	11.1	32.6	31.7	22.0	25.2	24.0	32.4	33.8	30.1
SiamPRN++	CVPR19	51.7	61.6	32.5	41.3	41.2	34.0	41.6	39.6	49.6	56.9	49.1
DiMP	ICCV19	61.1	71.7	49.2	44.7	43.4	39.2	47.6	45.1	56.9	65.0	56.7
SiamR-CNN	CVPR20	64.9	72.8	59.7	52.3	52.8	-	-	-	64.8	72.2	-
LT MU	CVPR20	-	-	-	48.5	47.3	41.4	49.9	47.3	57.2	-	57.2
Ocean	ECCV20	61.1	72.1	47.3	38.4	37.7	-	-	-	56.0	65.1	56.6
T rDiMP	CVPR21	67.1	77.7	58.3	-	-	-	-	-	63.9	-	61.4
T ransT	CVPR21	67.1	76.8	60.9	50.7	51.7	-	-	-	64.9	73.8	69.0
AutoMatch	ICCV21	65.2	76.6	54.3	47.2	43.5	37.6	-	43.0	58.3	-	59.9
STARK	ICCV21	68.8	78.1	64.1	-	-	-	-	-	67.1	77.0	-
KeepT rack	ICCV21	-	-	-	-	-	48.2	-	-	67.1	77.2	70.2
MixFormer-L	CVPR22	70.7	80.0	67.8	-	-	-	-	-	70.1	79.9	76.3
SBT	CVPR22	70.4	80.8	64.7	-	-	-	-	-	66.7	-	71.1
UAST	ICML22	63.5	74.1	51.4	-	-	-	-	-	57.1	-	58.7
SwinT rack-384	NeurIPS22	72.4	80.5	67.8	**55.9**	**57.1**	49.1	-	55.6	**71.3**	-	76.5
AiAT rack	ECCV22	69.6	80.0	63.2	-	-	47.7	55.6	55.4	69.0	79.4	73.8
CIA50	ECCV22	67.9	79.0	60.3	50.9	57.6	-	-	-	66.2	-	69.6
SimT rack-L	ECCV22	69.8	78.8	66.0	55.6	55.7	-	-	-	70.5	79.7	-
OST rack-384	ECCV22	**73.7**	**83.2**	**70.8**	**55.9**	**56.7**	**50.5**	**61.3**	**57.6**	**71.1**	**81.1**	**77.6**
DropTrack	**C VPR23**	**75.9**	**86.8**	**72.0**	**56.9**	**57.9**	**52.7**	**63.9**	**60.2**	**71.8**	**81.8**	**78.1**

It can be seen that the overall performance efficiency of target tracking algorithms in recent years is getting higher and higher, among which the Drop Track algorithm is the brightest. In the table, the widely used average overlap (AO) and success rate (SR), and the area under the Curve (AUC) metrics are selected as indicators, which can be seen that the target tracking algorithm based on deep learning accounts for a large proportion in recent years. Although there is a certain difference in calculation and time consumption, it can be seen that deep learning has occupied half of the tracking algorithm.

4.3. Deep Learning-based Pose Estimation Method

Pose estimation can be used to determine the position and orientation of a camera or device in order to correctly present a virtual object in a real scene. Therefore, pose estimation is one of the key steps to achieve virtual -real fusion and accurate alignment in AR applications.

Pose estimation can be achieved by using sensor data such as accelerometers, gyroscopes or magnetometers, among others, and computer vision techniques such as feature matching or structured light scanning, among others. The sensor data can provide information such as the acceleration, angular velocity, and magnetic field of the device to infer its orientation and position. While computer vision techniques can use images or point cloud data to estimate the pose of a device.

Since the AlexNet network was proposed in 2012, deep learning has developed rapidly and brought new development impetus to the field of human pose estimation. In 2014, convolutional neural networks were first successfully applied to solve the single-person pose estimation problem. Since then, the backbone structure based on convolutional neural network has been the mainstream method in the field of human pose estimation. Subsequently, the Transformer structure achieved great success in the field of time series, and some researchers began to introduce it into the field of computer vision, and the human pose estimation algorithm based on the Transformer structure has become a new research hotspot. Therefore, at present, pose estimation is roughly divided into convolution-based algorithms and Transformer-based algorithms [8].

However, pose estimation is a complex problem that involves multiple factors, such as sensor accuracy, algorithm efficiency, and environmental noise. Therefore, pose estimation needs to consider multiple factors and combine techniques such as algorithm optimization and parameter adjustment to improve accuracy and performance.

5. Holistic Analysis

5.1. Problems in Deep Learning Object Tracking

Deep learning technology provides the possibility to construct a more robust object appearance model, so it has become a trend to be widely used in object tracking tasks. The current tracking algorithms still have many challenges in high accuracy, high robustness and real-time performance in complex environments. The target tracking task has its own particularity, so there are some problems when applying deep learning technology, such as offline training data, real-time online training, and object occlusion.

5.2. Advantages and Disadvantages of Deep Learning Object Tracking in the AR Field

5.2.1. Advantages

High accuracy: Target tracking algorithms based on deep learning can learn and build an accurate target model, thereby improving the accuracy of tracking. Real-time: They can usually complete target tracking in a short period of time and achieve real-time operation. Adaptability: They can be adjusted adaptively according to the changes of the scene and environment, so as to improve the robustness of tracking. Scalability: Since deep learning algorithms are highly scalable, they can be applied to a wider range of AR scenarios and tasks.

5.2.2. Cons

High demand for training data: Object tracking algorithms based on deep learning require a large amount of training data to build an accurate object model, which may be difficult to meet in some AR application scenarios. High hardware requirements: They require a large amount of computation, so they need more powerful hardware support, which may limit their application on some mobile devices. Sensitivity to light and background: They are sensitive to changes in light and background, which may lead to unstable tracking performance under different lighting and background.

6. Conclusion

The object tracking algorithm of deep learning has undergone many changes in recent years, which have significantly improved the accuracy and robustness of the algorithm. One of the most important changes has been the shift from single-stage to two-stage object tracking algorithms. Single-stage algorithms, such as the Siamese tracker, use a single neural network for object tracking. While two-stage algorithms, such as Faster R-CNN and Mask R-CNN, use two neural networks: one for detecting the object and the other for tracking it. Two-stage algorithms are generally more accurate than single-stage algorithms, but they are also more computationally intensive. Another important change is the use of attention mechanisms in object tracking algorithms. The attention mechanism enables the algorithm to focus on the most important parts of the image, thus improving the tracking accuracy. Object tracking technology plays a vital role in the field of AR. Coupled with the development of large language models in recent years, it is hoped that through the summary and discovery of knowledge, people can develop a richer target tracking model similar to large language models in the field of tracking, such as multimodal fusion, joint learning, unsupervised learning, transfer learning, edge computing, and multi-modal fusion. The combination of reinforcement learning and other technologies is also one of the most important research directions in the future development.

References

[1] Azuma, R., Baillot, Y., Behringer, R., et al. (2001) Recent advances in augmented reality. IEEE Computer Graphics and Applications, 21(6): 34-47.

[2] Luo, B., Wang, Y., Shen, H., et al. (2013) Augmented reality hybrid tracking technology review. Journal of Automation, 39 (8): 17.

[3] Wu, Q., Yang, T., Liu, Z., et al. (2023) DropMAE: Masked Autoencoders with Spatial-Attention Dropout for Tracking Tasks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14561-14571.

[4] Huang, L., Zhao, X., Huang, K. (2019) Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE transactions on pattern analysis and machine intelligence, 43(5): 1562-1577.

[5] Wang, X., Shu, X., Zhang, Z., et al. (2021) Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13763-13773.

[6] Fan, H., Bai, H., Lin, L., et al. (2021) Lasot: A high-quality large-scale single object tracking benchmark. International Journal of Computer Vision, 129: 439-461.

[7] Fan, H., Lin, L., Yang, F., et al. (2019) Lasot: A high-quality benchmark for large-scale single object tracking. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 5374-5383.

[8] Feng, J., Zheng, J. (2023) Comparative Study of human Pose estimation methods based on convolution and Transformer. Software Engineering, 26(3):18-24.