Real-time Visual Tracking Based on Convolutional Neural Networks

Rui Li¹, Jirong Lian²

¹College of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China
²College of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China
jirong.lian@qq.com

Abstract. Traditional target tracking is based on target detection. When the target changes significantly, such as occlusion, scale change, the update of the tracking model will waste a lot of space and time resources, resulting in a very slow tracking speed, which cannot meet the actual engineering needs. In view of the above situation, an end-to-end tracking strategy is proposed, which is simpler and faster than the existing technology. The proposed tracker only needs to detect the first frame image and use it as the input of the model, and set the multi-task loss function to predict the position of the next frame of the target and the size of the bounding box. This paper constructs a lightweight network architecture with an additional selection mechanism to avoid wasting resources for global search and matching. Through experiments, good results can be achieved on the standard data set, and tracking speeds close to one hundred frames per second are achieved, which is very competitive with existing advanced trackers.

1. Introduction

As an important branch of computer vision, visual tracking has great research value and is widely used in robotics, intelligent transportation, automatic driving, monitoring and other fields [1,2]. Since visual tracking needs to establish the corresponding relationship between the target and the frame, there are great challenges, such as: background clutter, scene change, target scale change, lighting change, occlusion. In single target tracking, you can arbitrarily select the object of interest as the tracking target, mark the target with a rectangular bounding box in the first frame of the video and initialize it, and accurately locate the marked target in the subsequent frame and then lay the foundation for the behavior recognition of later targets.

In recent years, the target tracking algorithm has developed rapidly. Early tracking algorithms based on correlation filters used convolution operations to separate the background from the target and obtained the final estimated position of the target by maximizing the response map [3]. With the emergence of neural networks, with its powerful feature learning and representation capabilities, target tracking has been brought into a new era, greatly improving the accuracy and success rate of target tracking. Convolutional Neural Network (CNN) [4] has good robustness to a certain degree of deformation and deformation. Due to the existence of local receptive fields, the number of training parameters is greatly reduced. With the above advantages, the method based on CNN has made a great breakthrough, but due to the complex scenes of target tracking and the changeable characteristics of
the target, the entire learning process has a huge amount of calculation and long training time, and it is difficult to achieve a good balance between tracking speed and accuracy [5,6].

2. Related work

2.1. Visual target tracking
From a positioning point of view, visual tracking is a dynamic search process, which requires positioning the target in the current frame based on a priori results. Until now, the most used method is still based on detection tracking, which is to train a classifier with the real information marked in the first frame of the video, and then update it online [7,8]. However, a large amount of calculation and global search are required to cause the observation model to become very slow. Early Correlation Filter (CF) technology used filters to speed up the sampling of circulant matrices, and improved kernelized Correlation Filters (KCF) [9], introducing kernel methods to avoid complex non-linear Linear calculation, the effect is obvious, get amazing tracking speed. Later, SRDCF (Spatially Regularized Correlation Filters) [10] and C-COT (Continuous Convolution Operators for Visual Tracking) [11] proposed by Martin Danelljan and others have improved tracking accuracy, but the speed is very slow, only a few FPS.

2.2. Convolutional neural network
Recent research shows that the features learned by CNN in vision are far superior to the traditional manual features, which can greatly improve the accuracy and accuracy of tracking. Unlike traditional neural networks, convolutional neural networks add convolutional layers and down-sampling layers for feature extraction to ensure the completeness of learned features [12,13]. At the same time, convolutional neural networks can also share the weight of neurons, reducing the complexity of the calculation. A typical convolutional neural network (LeNet-5) structure is shown in Figure 1.

![Figure 1. Structure diagram of convolutional neural network LeNet-5](image)

Literature [14] conducted an in-depth study on CNN, and proposed that different convolutional layers characterize the target from different angles. Higher convolutional layers learn more abstract and advanced semantic information, which is very strong for different types of objects Distinguishing ability, on the contrary, has poor ability to distinguish similar objects; low-level learning of more detailed local features helps to distinguish the target from similar objects with similar appearance. Based on this, we construct a shallow network framework with an additional judgment mechanism, which automatically matches the learned features according to the number and number of occurrences of interfering objects.

2.3. Contribution of this paper
The contribution of this article mainly has the following three aspects:
(1) We constructed a shallow network with an additional judgment mechanism, which can automatically select the matching of features based on the number and frequency of interferences.
When a frame image contains only targets or contains a small number of analogs, low-level features can be selected Mark targets easily; otherwise use advanced semantic features. In this way, it can accurately locate and reduce the calculation complexity and improve the tracking speed.

(2) Due to the close relationship between frames in the video sequence, the position of the object in the current frame is not shifted from the position of the previous frame. With the above features, we can easily establish frames and frames The relationship between them, quickly and accurately predict the position of the target in the next frame.

(3) On the standard data set, experiments show that our method is faster than the previous method under the premise of ensuring accuracy.

3. Research methods

Target tracking is usually to specify the target in the initial frame of the video sequence and mark the position. In the subsequent frames, predict the Bounding box represented by a vector [X, Y, Height, Width] to describe the position information of the target. Where (X, Y) indicates the current coordinate point, Height indicates the height away from the coordinate point, and Width indicates the width away from the coordinate point. Using a single labeled image as the input of the network, you can perform feature extraction, feature matching, model update, trajectory prediction, and position determination. The tracking process is shown in Figure 2.

![Figure 2. Tracking flow chart](image)

To avoid redundant calculations, we agreed to fine-tune the first frame of the image when marking the annotation target in the initial frame. The size of the bounding box should not be too large to
prevent it from containing a lot of background information. Nor can it be too small to prevent the inclusion of a complete goal. Set the size of the image block Path to 1.5 times the target size as shown in Figure 3.

Figure 3. Targets marked with different sizes (A: black box is too large and contains too much background information. B: red box contains the appropriate amount of background and complete target. C: blue box does not completely contain the target)

3.1. Feature extraction
The neurons on the convolutional layer are used for feature extraction. Not all neurons are involved in a complete extraction process. For this reason, under the premise that the main neuron gets a response, selectively remove some of the layers. A part of neurons enhances the robustness of the network, reduces the number of training parameters, and reduces the complexity of the network. Training through classic feedforward and feedback. At the same time, the parameters of the network are very important. In the selection of the activation function, the ablation study shows that the ReLU activation function as equation (1) has the characteristics of fast and simple, and During the training process, some neurons can be randomly inactivated, which is a good match with our method and further improves the performance of the network. Because the Adam (Adaptive Moment Estimation) optimization algorithm has less memory and efficient calculation, the parameters are set to $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$, and each learning rate has a clear range, making the parameter change very stable, which is consistent with ReLU. When the learning rate is large, partial neuron deactivation will complement each other, which can achieve better results.

$$\text{ReLU}(x) = \max(0, x) = \begin{cases} 0, & x < 0 \\ x, & x > 0 \end{cases}$$

3.2. Feature matching, Model update
Our method is different from the past, common ones include heat map, score, global search, etc. According to the characteristics of video sequence images, that is, the movement of objects in a period of time is continuous, and there will be no excessive displacement as shown in Figure 4. We propose that the coordinate position of the object in the current frame $I_t$ must be near the previous frame $I_{t-1}$, and this range is very small. Based on the above observations, it is only necessary to search and match within a small range of the current position of the target, and to train by minimizing the square loss as shown in equation (2), which greatly reduces the global search and redundant calculation.

$$L_m = \| \hat{I} - I_t \|$$

Where $\hat{I}$ represents the predicted position and $I_t$ represents the target real position.
It can be seen from Figure 4 that the method based on our observation and proposal is feasible. The displacement of the target between the position of the current frame A and the previous frames B and the following frames C will not be large, that is, no large jump will occur. Therefore, we propose that you only need to search for matches in the black box and do not need to spend a lot of time searching in the red box, which greatly reduces the waste of resources for searching and matching, reduces the calculation complexity, and improves the speed of the network.

3.3. Trajectory prediction, Position determination
In 2D images, the motion of the target is based on a plane. In our proposed visual tracking method, the prediction of the target's movement trajectory is added. By predicting the possible direction of the target's next time movement as shown in Figure 5, it can greatly reduce the search range during feature matching and improve the tracking efficiency.

From the initial frame of the marked video sequence, find the center $c = (x, y)$ of the bounding box, where $c$ represents the coordinates of the center point, set the upper left corner of the image as the coordinate origin, $x$ represents the horizontal distance from the coordinate origin, and $y$ represents the vertical distance from the coordinate origin. By calculating the displacement vector $\vec{v} = (\Delta x, \Delta y)$:

$$\vec{v} = c_i - c_{i-1}$$
$$\Delta x = x_i - x_{i-1}$$
$$\Delta y = y_i - y_{i-1}$$
Where $c_t$ represents the center coordinate of the target in the $t$ frame. Further, we learn this displacement vector through Equation 6 regression:

$$L_v = \frac{1}{N} \sum_{i=1}^{N} |v_i^f - (g_i^f - g_i^{-1})|$$

Among them $v_i^f$ represents the displacement of the target, $g_i^f$ and $g_i^{-1}$ represent the real position of the ground respectively. Through the above regression, the moving direction of the target can be obtained, and the position coordinates of the target can be further quickly determined.

4. Experiment and Analysis

This method is implemented in Python based on the TensorFlow framework. The experimental equipment hardware is a computer equipped with NVDIA Titan X display processing core and equipped with i7-7700k processor. In order to evaluate the proposed method, we use the standard tracking benchmark Online Tracking Benchmark (OTB) [15,16] and Visual Object Tracking Challenge 2016 (VOT2016) [17], while using success rate, accuracy, tracking speed (FPS) Three indicators compare our tracker with the classic tracker. Some representative tracking effects are shown in Figure 6.

![Figure 6. The tracking effect of our tracker (Ours) and a typical tracker on a standard tracking benchmark](image)

The above results show that, in the challenge of different types of data, our tracker (Ours) can successfully track the target. Other trackers are limited by different influencing factors, which results in poor tracking of the target. Error tracking.

Quantitatively compare the proposed tracker with typical trackers, including ECO, SiameFC, SRDCF, PTAV, KCF, DSST, MEEM, among which are trackers based on early manual features and trackers based on neural network feature extraction. In order to be fair and equitable, the quantitative data is expressed by the average value obtained on the standard test set. The results are shown in Table 1.

|      | Accuracy | Success rate | Speed |
|------|----------|--------------|-------|
| Ours | 0.882    | 0.672        | 92    |
| ECO  | 0.874    | 0.643        | 6     |
It can be seen from the data in Table 1 that the method proposed in this paper achieves a good balance between tracking accuracy and speed. The early trackers used for manual features tracked quickly, but the accuracy and success rate were not satisfactory, and they could not be tracked very well. The method based on neural network can track the target well, but the real-time performance is too poor, usually only a few FPS or a dozen FPS. The method in this paper greatly improves the tracking speed while ensuring the tracking accuracy and success rate, which is in line with the application of actual engineering.

At the same time, in the experiment, our method is not robust to some complex scene changes and high-speed motion of objects, and there will be missing targets or tracking errors.

5. Conclusion
In this paper, we propose a simplified visual tracker based on convolutional neural networks. Based on an observation that the displacement of the target in a short time is extremely small, the additional judgment mechanism reduces the search range and greatly reduces the computational complexity. At the same time, the model will calculate the target's movement direction, further predict the target's movement trajectory, improve tracking efficiency and tracking speed. The whole process is simple and clear, but it is very effective, and the tracking speed of nearly one hundred frames per second on the standard data set is very impressive. A large number of experiments show that our proposed tracker has better performance than many trackers.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (No. 61761028).
This work was supported by the Gansu Provincial Key R&D Program-Industrial (18YF1GA060).

References
[1] Lu Huchuan, Li Peixia, Wang Dong. Overview of target tracking algorithms [J]. Pattern Recognition and Artificial Intelligence, 2018, 31(01): 61-76.(in Chinese).
[2] Xin Changhao, Chang Yuke, Zhao Hongzhi, Zhang Bin. Real-time tracking of UAV dynamic target based on monocular camera[J]. Science and Technology Innovation and Application, 2020(08): 46-48.(in Chinese).
[3] Wang L, Liu T, Wang G, et al. Video Tracking Using Learned Hierarchical Features[J]. IEEE Transactions on Image Processing, 2015, 24(4):1424-1435.
[4] Chen Qianqian. Scale adaptive correlation filtering target tracking algorithm based on convolutional neural network [C]. Process Control Professional Committee of the Chinese Society of Automation. Summary of the 30th China Process Control Conference (CPCC 2019). Process Control of the Chinese Society of Automation Professional Committee: Chinese Automation Society Process Control Professional Committee, 2019: 69.(in Chinese).
[5] Shi Yaqian. Application Research of Deep Learning and Deep Reinforcement Learning in Video Target Tracking [D]. Xi’an University of Technology, 2019.(in Chinese).
[6] Wang Q, Teng Z, Xing J, et al. Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018:4854-4863.
[7] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.
[8] Du Sheng. Research on outdoor vehicle tracking algorithm based on deep learning [D]. Xi’an University of Technology, 2017. (in Chinese).

[9] Henriques J F, Caseiro R, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3):583-596.

[10] Danelljan M, Häger, Gustav, Khan F S, et al. Learning Spatially Regularized Correlation Filters for Visual Tracking[J]. 2016:4310-4318.

[11] Danelljan M, Robinson A, Khan F S, et al. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking[J]. 2016:1-16.

[12] Ping-yue Lv, Chang-qing Lin, Sheng-li Sun. Dim small moving target detection and tracking method based on spatial-temporal joint processing model[J]. Infrared Physics and Technology, 2019, 102.

[13] Wang Q, Zhang L, Luca Bertinetto, et al. Fast Online Object Tracking and Segmentation: A Unifying Approach[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019:1328-1338.

[14] Wang, Lijun, et al. Visual Tracking with Fully Convolutional Networks.[C]// The IEEE International Conference on Computer Vision (ICCV) IEEE. 2016:3119-3127.

[15] Wu, Y., Lim, J., Yang, M.: Object tracking benchmark[J]. IEEE Trans. Pattern Anal. Mach. Intell. 37(9), 2015:1834–1848.

[16] Wu Y, Lim J, Yang M H. Online Object Tracking: A Benchmark[C]// Computer Vision and Pattern Recognition (CVPR). 2013 IEEE Conference on. IEEE, 2013:2411-2418.

[17] Battistone F, Santopietro V, Petrosino A. The Visual Object Tracking VOT2016 challenge results[C]// European Conference on Computer Vision. Springer, Cham, 2016:98-111.