Scholar Publishing Group
International Journal of Neural Network
https://doi.org/10.38007/NN.2022.030402
ISSN 2790-2692 Vol. 3, Issue 4: 9-16

Algorithm Research of Neural Network for Feature Target Tracking

Rashid Almulihi*

Jawaharlal Nehru University, India

**corresponding author*

Keywords: Computer Vision, Convolutional Neural Network, Target Tracking, Tracking Accuracy

Abstract: In recent years, with the explosive growth of research in the field of computer vision, TT has an important application in visual recognition tasks. It can assist target detection and improve the speed of recognition, which has certain theoretical value and research significance. However, in practical application scenarios, target tracking (TT) still faces the problems of inaccurate tracking, poor robustness, and low overall system speed caused by scene changes. Since the TT algorithm based on convolutional neural network (CNN) was proposed, it has attracted the attention of a large number of researchers with the advantages of both speed and accuracy. The multi-layer convolution features extracted from the input image through the CNN have a good appearance representation ability for the target under a variety of complex interference factors. Therefore, this paper uses the CNN to build a feature TT algorithm model, and experiments prove that the accuracy of the model can be improved by increasing the order of magnitude and deepening the ResNet backbone network design. After comparing the tracking effects of the CNN algorithm and other algorithms, it shows that the CNN algorithm has the highest tracking accuracy and success rate.

1. Introduction

After years of research by researchers, the TT algorithm has achieved good tracking results in simple application scenarios. However, in actual scenes, there are a series of challenging interference factors such as occlusion, target deformation and motion blur, which requires Establishing a more complex appearance representation model and a reasonable model update strategy are also the key to the success of TT [1].

So far, domestic and foreign researchers have proposed many excellent and innovative TT algorithms, laying a solid foundation for the practical application of TT technology in people's daily

life. For example, the DLT algorithm proposed by a scholar is the first successful attempt of deep learning in the field of tracking. DLT adopts the idea of transfer learning. When the tracking data is very limited, the training set of other tasks is used to assist training to obtain the abstract representation of the target. During tracking, the positioning image of each frame is used to fine-tune the network to adapt it to the actual situation. Environment and Task [2]. Some scholars proposed the FCNT algorithm. By analyzing and improving the performance of the VGGNet network based on ImageNet training, three networks were constructed for tracking, and the classification task data set and tracking task were well combined, and a good tracking effect was obtained [3] . Now, more scholars are studying the neural network extraction and tracking target convolution feature, which opens up a new path for the follow-up research based on deep learning TT algorithm.

This paper first introduces the concept of feature fusion of TT and the component modules of TT; then proposes target feature extraction and TT algorithms based on convolutional neural networks; then through experiments to verify the magnitude of the dataset and different ResNet backbone networks for CNN algorithms The influence of the tracking effect of the model, and finally compared the tracking accuracy and average success rate of the CNN algorithm.

2. TT Algorithm

2.1. Feature Fusion of TT

In the TT framework, the input at the beginning of the frame is a frame by frame of images, and then in the TT process, features are extracted from the search area, and the features are used as the input of the tracker. Currently widely used manual features such as HOG features, color features, etc. can only deal with some challenging scenarios [4]. In order to improve the adaptability of the algorithm in various scenarios, many algorithms fuse features. Although the scene adaptability of the algorithm has been greatly improved after feature fusion, and the tracking accuracy has also been improved, the feature fusion will also generate additional time overhead, reduce the speed of the algorithm, and the HOG feature can perform well on the target. Introducing color features in the case of tracking can sometimes make tracking inaccurate [5-6].

In addition to being able to better help the algorithm track targets, multi-feature fusion algorithms sometimes have negative effects. For example, when the HOG feature tracking result is accurate, introducing other features may affect the HOG feature tracking, making the tracking inaccurate [7]. When the HOG feature can accurately track the target, the use of the feature fusion mechanism and the calculation of the color feature or the gray feature will increase unnecessary overhead [8].

Considering the accuracy of the algorithm, it is a research problem to reduce the negative impact on HOG feature tracking, thereby improving the tracking accuracy in some scenarios. Considering the real-time performance of the algorithm, it is necessary to reduce the computational overhead as much as possible while ensuring the tracking accuracy of the algorithm and increase the tracking speed of the algorithm [9].

2.2. TT Components

From the vertical point of view, the TT algorithm mainly includes 5 modules. The overall schematic and the role of each module are shown in Figure 1:

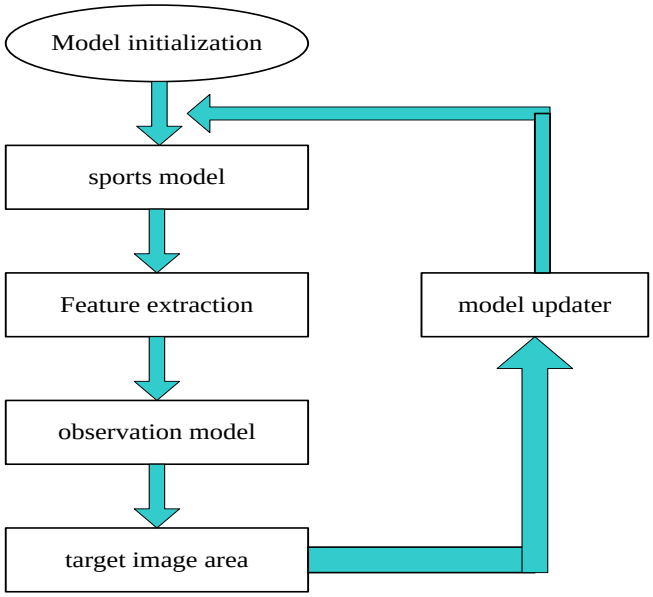

Figure 1. Schematic diagram of the tracking framework

(1) Motion model

Motion model based on the estimation of the previous frame, the motion model generates a set of candidate regions or bounding boxes for the target, and the target is determined in the candidate frame. The motion model mainly establishes the relationship between the frame before and after the motion state of the tracking target in the whole video sequence, and directly or indirectly predicts the target in the candidate frame [10].

(2) Feature extraction

HOG (Histogram of Oriented Gradients) combines the original color map features. As the name suggests, the HOG feature fuses the original color map features. But now the mainstream tracking algorithms are based on convolutional neural network features. Usually, we model the tracking system as a binary classification problem. When extracting features from convolutional neural networks, small networks such as VGG-M are mostly used to extract features. The tracker only obtains the information of the first frame of the video sequence, and the prior knowledge of the target is limited. Therefore, offline pre-training and online fine-tuning can be used to extract more effective prior information [11-12].

(3) Observation model

The observation model makes a confidence judgment on the candidate area of the current frame, and calculates the probability that the candidate frame is the target. Usually, the observation model is considered to be the key part of the tracking system. Visual features are extracted from images, and the features are input into the observation model, decision-making or matching, and the precise location of the target is determined according to the final result [13]. Among the several modules of tracking, the robustness of the observation model is the key to the success of the algorithm.

(4) Model updater

Usually model updaters include online classifiers, incremental subspace learning algorithms, and real-time template change updates. Ensuring that the appearance update of the target and the background can be accurately described without causing the model's ability to describe the target is also a major problem in computer vision TT [14].

(5) Integrated processor

The results of a single tracker are sometimes unstable, and the performance varies greatly even with small perturbations in the parameters. At present, some scholars propose an integrated tracker

for this problem, which is composed of multiple trackers.

3. Algorithm of Feature TT Based on Neural Network

3.1. Image Target Feature Extraction Based on CNN

In this paper, CNN is used to extract the features of the input image. The input image goes through five convolution blocks, five pooling layers and three fully connected layers to obtain the classification result. For the CNN, there are five parts in total. After the input image passes through the five parts in the CNN in turn, multi-layer convolution features and classification results of different resolutions will be obtained.

(1) Input layer: The resolution of the input image in CNN is specified as 224*224.

(2) Convolutional layer: The five convolutional blocks in CNN contain a total of thirteen convolutional layers. Each convolutional layer contains a convolutional kernel with different number of channels but the size is 3*3. The input and volume the convolution kernel of a certain number of channels in the convolution layer is convolutional to obtain the convolution features of the current convolution layer [15].

(3) Non-linear layer: The activation function is used in the CNN to change the value that plays a lesser role in the convolution feature to 0, thereby highlighting the value that plays a greater role in the convolution feature. The Sigmoid function will limit each value in the convolution feature to 0-1 so that each feature value is similar, and the Relu function will change the negative value in the convolution feature to 0 to highlight the positive value in the convolution feature. and the larger the positive value, the greater the effect [16-17].

(4) Pooling layer: There are five pooling layers in the CNN and they are located behind each convolutional block to reduce the resolution of the convolutional feature map. The pooling operation not only reduces the computational complexity but also makes the representation of convolutional features more robust. We use max-pooling because the convolutional features obtained after max-pooling are basically unchanged when the image is shifted a little.

(5) Fully connected layer: There are three fully connected layers in CNN, and the fully connected layer is to obtain the classification result according to the multi-layer convolution features obtained in the previous four parts. In TT, we do not need to classify the target, we It is only necessary to extract multi-layer convolutional features that are more capable of representing the tracking target and select appropriate convolutional features to train the appearance representation model and related filters.

3.2. TT Algorithm Based on Siamese CNN

Siamese CNN refers to inputting two inputs into two identical neural networks respectively, and these two neural networks respectively map the input to a new space to form a representation of the input in the new space. The parameters are shared between the two networks, and the similarity of the two inputs is evaluated through the calculation of the loss, as shown in Figure 2. Due to the good performance of the Siamese neural network, the Siamese network has been widely used and developed rapidly, and a series of high-performance network architectures have been derived. The Siamese neural network model needs to be run twice to get the loss. For video TT, the distance between the template frame and the current frame is measured [18].

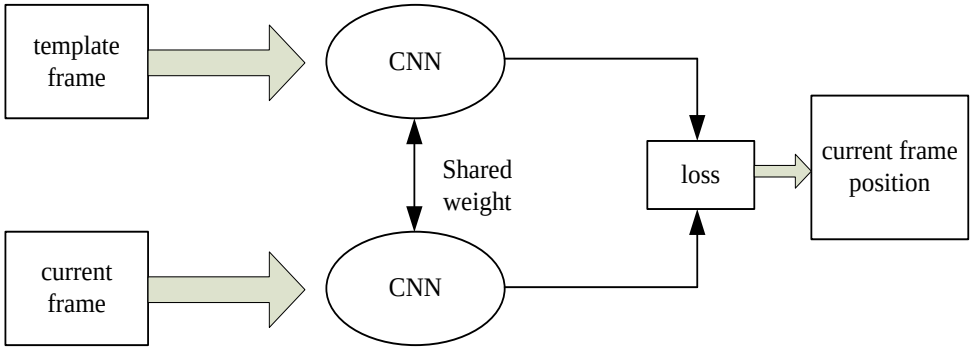

Figure 2. Object tracking of Siamese CNN

The essence of the SiamFC algorithm is to measure the similarity between the search area and the known target template, and to achieve TT according to the calculation results. The SiamFC algorithm makes full use of the advantages of end-to-end learning. It only needs to be trained offline on large-scale datasets without online updates, and can achieve a good coordination between accuracy and real-time performance, so it has become a similarity-based algorithm. Learn the classic tracking algorithm.

In the SiamFC algorithm, the deep Siamese neural network structure is used to learn a similarity matching function f(k; x). The calculation method of the similarity matching function is shown in formula (1):

$$f(k;x) = g(\varphi(k), \varphi(x)) \tag{1}$$

Among them, the function φ is equivalent to a feature extractor. The function g is used to measure the similarity matching degree between the template image k and the search image x, allowing the k and x dimensions to be different.

The specific loss function formula used by the SiamFC algorithm is as follows:

$$L(y,v) = \frac{1}{|D|}\sum_{v\in D} f(y[u], v[u]) \tag{2}$$

$$y[u] = \begin{cases} +1, if k\|u-c\| \le R \\ -1, otherwise \end{cases} \tag{3}$$

where y[u]∈{+1,-l} is the label corresponding to the image pair, +1 represents the positive sample, -1 represents the negative sample, v[u] is the response map, and u∈D represents the value in the score map. For each position, f is the logistic loss function, k is the total step size of the network, c is the center of the target, and R is the defined radius.

4. Algorithm Application

4.1. Performance Evaluation of Feature TT Algorithm Based on CNN

The most used dataset is the MOT Challenge. It includes MOT15, MOT16, MOT17, MOT20 and other series of data sets. The tracking performance of the target needs to be measured by standard evaluation indicators. At present, for the multi-TT problem, it is mainly based on the CLEAR-MOT indicator. For the realization of an ideal multi-TT, it should be ensured that the emerging targets can be identified and tracked in time, followed by tracking. The target position is

consistent with the real position, and finally, the target ID switch is avoided and the ID of the same target is kept consistent. The evaluation index of multi-TT is mainly the multi-TT accuracy MOTA (Multiple Object Tracking Accuracy), FN, FP and IDSw are the number of loss, false detection and mismatch. IDP is the recognition precision; IDR is the recognition recall; IDF1 is the recognition F value.

The experiment uses the MOT17 data set, because the performance test of this data set in the field of TT is often used as a public data set, and because this data set is more objective than the MOT16 detector, it is newer than the MOT20 data, and the experimental test uses Not much, lack of contrast. Therefore, the public data set of MOT17 is selected as the model verification in this paper.

For the magnitude of the dataset, the MOTA accuracy metric of the experiment can be improved. In order to verify the effect of the magnitude of the data set on the effect of the CNN algorithm model, the MOT17 data set was divided into a quarter (MOT17-quarter), a half of the MOT17 data set (MOT17-half) and the complete MOT17 data set for comparative experiments . The experimental results are shown in Figure 2.

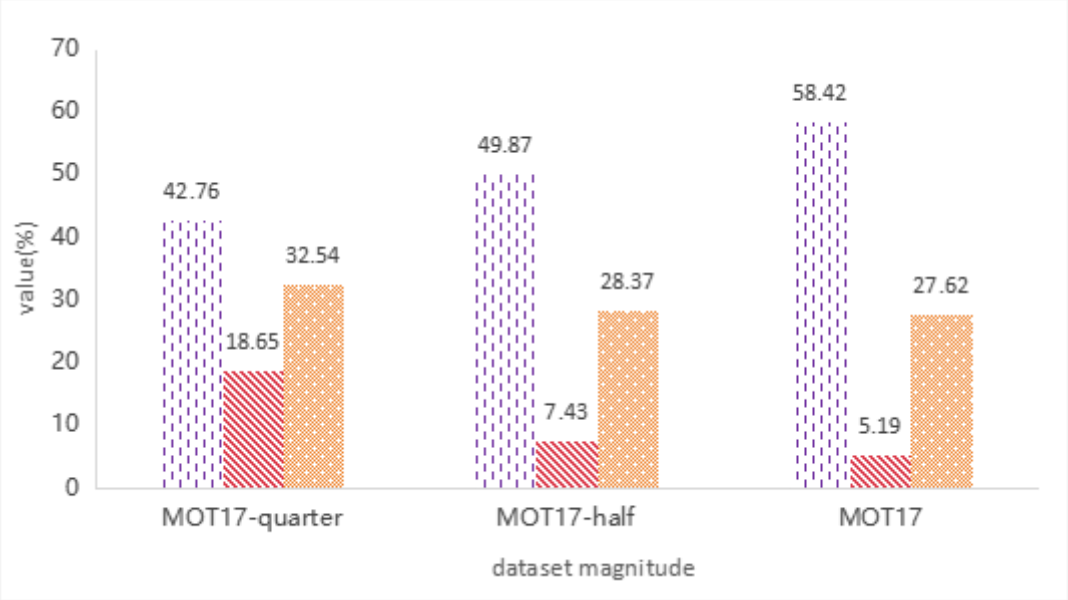

Figure 3. The magnitude of the dataset affects the effect of the model (%)

It can be seen from the experimental results in Figure 3 that the magnitude of the dataset has a serious impact on the model. Especially the indicator of MOTA accuracy, of course, this indicator is also an important indicator to measure the pros and cons of multi-target trackers, and this indicator increases with the increase of the magnitude of the data set. It shows that if you want to improve the accuracy of the model, it can be greatly improved by increasing the data set.

Table 1. Effects of different ResNet backbone networks on the model

	MOTA	IDF1	IDSw
ResNet-34	58.34%	63.51%	4456
ResNet-50	60.67%	64.03%	4392
ResNet-101	61.59%	64.78%	4013

In order to verify the effect of different ResNet backbone networks on the CNN algorithm model, different CNN backbone networks were tested on the MOT17 dataset. The experimental results are shown in Table 1, mainly for ResNet-34, ResNet-50 and ResNet-101. Test, the experimental results can see that ResNet-50 is better than ResNet-34, and the best effect is ResNet-101 network. The deeper the ResNet-101 network design is, the output of the next layer is the linear combination of

the previous layer plus activation, which can be The fusion gets more and more flexible features, so the performance is better.

4.2. Analysis of Tracking Effect of CNN Algorithm

Table 2. Comparison results between CNN algorithm and mainstream algorithms (%)

	CNN	KCF	CN	SAMF	HCF
Precision	94.7	76.8	62.5	84.3	91.2
Success rate	88.5	71.4	57.9	77.6	80.7

Table 2 shows the comparison of the experimental results between the CNN algorithm and the KCF, CN, SAMF and HCF algorithms. It can be seen from the table that the HCF algorithm using convolution features has better tracking performance than other algorithms, and the tracking accuracy and average success rate are respectively the tracking accuracy and average success rate of the CNN algorithm are 94.7% and 88.5%, respectively. Compared with the HCF algorithm, the tracking accuracy is increased by 3.5% and the success rate is increased by 7.8%.

5. Conclusion

Convolutional networks are an example of deep learning deeply influenced by neuroscience principles, which have achieved great success in the field of feature object tracking. The TT algorithm has experienced a long development process before it has mature and stable technical achievements at this stage. Therefore, analyzing and learning the pioneering template tracking algorithms that have appeared in the long history can not only help us master the core ideas and pioneering ideas, but also it can promote our deeper innovation based on it.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Rohan A, Rabah M, Kim S H. Convolutional Neural Network-based Real-Time Object Detection and Tracking for Parrot AR Drone 2. IEEE Access, 2019, PP(99):1-1. https://doi.org/10.1109/ACCESS.2019.2919332

[2] Jondhale S R, Wakchaure M A, Agarkar B S, et al. Improved Generalized Regression Neural Network for Target Localization. Wireless Personal Communications, 2021, 125(2):1677-1693. https://doi.org/10.1007/s11277-022-09627-9

[3] Sebkhi N, Sahadat N, Hersek S, et al. A Deep Neural Network-Based Permanent Magnet Localization for Tongue Tracking. IEEE Sensors Journal, 2019, PP(99):1-1.

[4] Bohush R, Zakharava I. 109-116. - Citation: Bohush RP, Zakharava IY. Person Tracking Algorithm Based on Convolutional Neural Network for Indoor Video Surveillance. Computer

Optics, 2020, 44(1):109-116. https://doi.org/10.18287/2412-6179-CO-565

[5] Issaadi S, Issaadi W, Khireddine A. New Intelligent Control Strategy by Robust Neural Network Algorithm for Real Time Detection Of An Optimized Maximum Power Tracking Control In Photovoltaic Systems. Energy, 2019, 187(Nov.15):115881.1-115881.21. https://doi.org/10.1016/j.energy.2019.115881

[6] Ham D, Cho H C, Yoon Y J, et al. Feature Based Extended TT Using Automotive 2D LIDAR. Transactions of the Korean Institute of Electrical Engineers, 2021, 70(1):224-235. https://doi.org/10.5370/KIEE.2021.70.1.224

[7] Mayorca-Torres D, Guerrero-Chapal H, J Mej á-Manzano, et al. Multi-TT for Sperm Motility Measurement Using the Kalman Filter and JPDAF: Preliminary Results. RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao, 2019, 1(E22):282-294.

[8] Shiozuka M, Yotsumoto T, Takahashi K, et al. Agent-based Tracking Method Addressing Target Recognition Errors. IEEJ Transactions on Electronics Information and Systems, 2020, 140(4):484-491. https://doi.org/10.1541/ieejeiss.140.484

[9] Moorthy S, Joo Y H. Multi-Expert Visual Tracking Using Hierarchical Convolutional Feature Fusion Via Contextual Information. Information Sciences, 2021, 546(2021):996-1013. https://doi.org/10.1016/j.ins.2020.09.060

[10] Bhat P G, Subudhi B N, Veerakumar T, et al. Multi-Feature Fusion in Particle Filter Framework for Visual Tracking. IEEE Sensors Journal, 2020, 20(5):2405-2415. https://doi.org/10.1109/JSEN.2019.2954331

[11] Wulff D, Kuhlemann I, Ernst F, et al. Robust Motion Tracking of Deformable Targets in the Liver Using Binary Feature Libraries in 4D Ultrasound. Current Directions in Biomedical Engineering, 2019, 5(1):601-604. https://doi.org/10.1515/cdbme-2019-0151

[12] Pichlmeier S, Pfeiffer T. Adaptive Target Enhancement Determines Levels of Guidability in Multiple Object Tracking. Vision Research, 2021, 183(13):61-72. https://doi.org/10.1016/j.visres.2021.02.001

[13] Hsu H M, Cai J, Wang Y, et al. Multi-Target Multi-Camera Tracking of Vehicles Using Metadata-Aided Re-ID and Trajectory-Based Camera Link Model. IEEE Transactions on Image Processing, 2021, PP(99):1-1. https://doi.org/10.1109/TIP.2021.3078124

[14] Walia G, Ahuja H, Kumar A, et al. Unified Graph-Based Multicue Feature Fusion for Robust Visual Tracking. IEEE Transactions on Cybernetics, 2020, 50(6):2357-2368. https://doi.org/10.1109/TCYB.2019.2920289

[15] Chu V C, D'Zmura M. Tracking Feature-Based Attention. Journal of Neural Engineering, 2019, 16(1):308-321. https://doi.org/10.1088/1741-2552/aaed17

[16] Dash P P, Patra D. Efficient Visual Tracking Using Multi-Feature Regularized Robust Sparse Coding and Quantum Particle Filter Based Localization. Journal of Ambient Intelligence and Humanized Computing, 2019, 10(2):449-462. https://doi.org/10.1007/s12652-017-0663-5

[17] Nodehi H, Shahbahrami A. Multi-Metric Re-identification for Online Multi-Person Tracking. IEEE Transactions on Circuits and Systems for Video Technology, 2021, PP(99):1-1. https://doi.org/10.1109/TCSVT.2021.3059250

[18] Sliti O, Hamam H. Efficient Visual Tracking Via Sparse Representation and Back-Projection Histogram. Multimedia Tools and Applications, 2019, 78(15):21759-21783. https://doi.org/10.1007/s11042-019-7439-1