Construction and Simulation of Deep Learning Algorithm for Robot Vision Tracking

Siping Xu¹ and Lan Chen²

¹College of Mathematics and Physics, Hubei Polytechnic University, Huangshi 435000, China
²Operation Department, Wuhan Jianxing Urban Resources Operation Management Co., Ltd., Wuhan 430000, China

Correspondence should be addressed to Siping Xu; 2010800405@hbut.edu.cn

Received 5 June 2022; Revised 30 June 2022; Accepted 11 July 2022; Published 20 July 2022

As one of the indispensable basic branches of computer vision, visual object tracking has very important research value. Therefore, a deep learning based on robot vision tracking is evaluated. Based on the basic principles of target tracking and search principle, a deep learning algorithm for visual tracking is constructed, and finally, evaluated, and simulated. The results showed that the accuracy rate increased from 90.9% to 90.13% after the addition of channel attention mechanism module. Variance was reduced from 3.78% to 1.27%, with better stability. The EAO, accuracy, and robustness of the algorithm are better than those without significant region weighting strategy. The strategy of using the improved residual network SE-ResNet network to extract multiresolution features from the correlation filtering framework is effective and helpful to improve the tracking performance.

1. Introduction

As one of the indispensable basic branches of computer vision, visual object tracking has very important research value. The problem is specifically defined as that in a video sequence, an object of any category at any position in the initial frame is designated as the target, and the target tracking algorithm can frame the target quickly and accurately in the subsequent frames by means of image processing and machine learning. Target tracking technology, which can realize the above functions and has both real-time and robust performance, is the core of artificial intelligence-related applications. For example, in the autonomous driving system, target tracking can estimate and predict the position trajectory of pedestrians and vehicles in front of the vehicle, which can make decisions for the vehicles’ next direction and speed. In the road navigation system, target tracking can avoid the dynamic obstacles on the road ahead. In the urban surveillance system, target tracking saves a lot of manpower for searching and tracking. Target tracking can also be embedded in the UAV equipment to achieve autonomous obstacle avoidance and follow the designated target. Review the development of tracking algorithm, and its milestone innovation is usually established on the breakthrough of some theory or method, roughly experienced four stages: the machine learning methods represented by support vector machine Bayesian classifier sparse can realize simple and complete scene tracking. The discriminant model based on particle filter can distinguish the complex background well, but the sampling process is time-consuming and random. Target tracking under the framework of correlation filtering mainly includes two aspects: cyclic shift sampling and ridge regression objective function optimization. With complete mathematical theory and high stability, it is the preferred method to try in landing applications. A lot of work is devoted to theoretical improvement and precision improvement of such algorithms, as well as tracking algorithms based on deep learning. Feature expression has good robustness and implicit nonlinear fitting ability. However, the exploration
time of such methods is short, and there is still a large space for development compared with mature algorithms in terms of accuracy and speed, as shown in Figure 1.

2. Literature Review

A large number of studies show that the target tracking algorithm has always been the breakthrough of machine vision learning, and new target tracking algorithms and ideas continue to emerge. However, it is difficult to have an algorithm to deal with all kinds of complex scenes, and the difficulties in improving the performance of target tracking are mainly as follows: structural changes to the nonrigid target itself, perspective transformation of rigid targets, similarity in scene and target characteristics, similarity between multiple objects, a change in illumination, occlusion of a target, a sudden change in its direction of motion, changes in target scale and resolution, and limitations in computational time and space complexity. Existing algorithms have solved one or more of these problems to a certain extent, but there is still a lot of room for performance improvement.

As early as in the 1960s, Ding et al. proposed a method to obtain three-dimensional shape information of objects from two-dimensional images [1]. This method based on computer theory requires physical photography to achieve [2]. At this point, the machine vision theory and practice research for the purpose of analyzing object 3D scene is like a fire like tea [3]. At the beginning of 1970, Sami et al. established a systematic computer vision theory, which laid the foundation for some researches on machine vision theory and was a milestone progress. Its core content was to recover three-dimensional geometric shapes of objects based on two-dimensional images [4]. Since the early 1980s, Zieliński and Markowska-Kaczmar’s research on machine vision has been a hot research field in modern high-tech research and has become more and more mature in practical application [5]. Wu et al. first proposed the concept of mean-drift vector in 1975 [6]. Jahanbakt. applied the iterative procedure of calculating mean-drift vector to image segmentation and target tracking [7].

On the basis of current research, deep learning of the robot vision tracking algorithm is proposed, and most target tracking algorithms are based on candidate to find the target. Therefore, how to effectively generate screening candidate samples based on the location of the last frame is also a key link in target tracking. Most target tracking algorithms default that the movement of the object between consecutive frames is not too violent. Therefore, the motion model can be used to generate candidates around the location of the object in a frame. At present, there are mainly two ways to generate candidate frames: particle filter and sliding window particle filter, which use the predicted position of the above frame as the center to transform the radiation parameters of six candidate frames, so as to obtain a series of candidate samples [8]. The six parameters include horizontal displacement, vertical displacement, rotation angle, aspect ratio, stretch ratio, and dimension particle filter. Most of the methods use reconstruction error as the benchmark for target screening. However, this method also has its disadvantages; that is, when generating candidate samples, there will be a lot of redundant calculation due to the overlap between samples [9]. In addition, the number of samples is also difficult to control: using too many samples will cause redundant calculation and tight computer memory, resulting in very slow tracking. If too few samples are used, the tracking speed will be improved to some extent, but the area where the real target is often cannot be proposed as a candidate, thus reducing the tracking accuracy. Therefore, many parameters of particle filter have many problems of manual adjustment in practical application. Sliding window is a method of exhaustive extraction in theory, which uses the horizontal and vertical displacement of the target frame to extract candidates [10]. Because a circular sample is put forward, and can be fast calculation, in fu, Dhiman et al. makes this method also has been widely used, but the nature of the edge effect, because of its circulation sample window need to add after the gauss window to calculate, led to the fast moving object is extremely easy to produce trace drift phenomenon. At present, there are many algorithms to solve this problem [11].

3. Appearance Modeling of Target Tracking

3.1. Basic Principles of Target Tracking. Although target tracking algorithms have different ideas based on point, line and plane, or generative and discriminant, they all revolve around the basic flow chart of target tracking based on four basic modules, including target feature extraction target construction model target search strategy and target model update. Firstly, it is necessary to initialize the tracking target and determine its initial position, then extract effective feature information in the target area to describe the target accurately, and then establish the target model [12]. Finally, according to the target model, an appropriate search strategy is designed to estimate the optimal target area according to the interference encountered in the target tracking process, and the target model is reasonably updated to adapt to the change of the target appearance [13], see Figure 2.

3.2. Search Strategy for Target Tracking. Target search strategy is to find the best method of similarity measure in the current image search which is most similar with the target area; usually by some distance to calculate the target tracking algorithm, the similarity measure reflects the similarity between the target template with the candidate; so, the selection of similarity measurement strategy is critical and has a direct effect on the result of the target tracking. Appropriate similarity measurement method can objectively reflect the relationship between candidate target and template [14].

Euclidean distance is the most common definition of distance, which represents the real distance between two points. The Euclidean distance formula between two points on a two-dimensional plane is shown in Formula (1).

\[
\text{dist}_{ab} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{1}
\]

The Euclidean distance formula between two points in n
-dimensional space is shown in Formula (2):

$$\text{dist}_{ab} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$  \hfill (2)

3.3. Algorithm Framework. Offline training twin network of the two branches of learning in a 255 × 255 search area locates 127 × 127 goals, and it studied the estimation function, a similarity to the target and the search area of every position compared to predict a confidence figure; in the image, the target area high degree of confidence, confidence is low background area [15]. In particular, this algorithm proposes a crosscorrelation layer to calculate the similarity between each position and the target in the search area at one time, as shown in Formula (3).

$$F(z, x) = \phi(z) \ast \phi(x) + v.$$  \hfill (3)

The two branches of the network adopt the same structure and parameters and are composed of three parts, namely, local pattern detection module background, modeling module, and integration module [16]. The details of these modules will be elaborated in the following sections. The last crosscorrelation operation is performed on the output of the integrated modules. The algorithm employs logical loss function to train the network, as shown in Formula (4).

$$L = \log (1 + e^{-yv}).$$  \hfill (4)

In this paper, the target location task is described as a conditional probability modeling task; so, this paper first uses standard conditional probability learning to explain the algorithm [17]. Its purpose is to find a parameter matrix $W$ for each video, which can minimize the loss of the prediction function $F$, where $L$ is the average hundred of the loss of $N$ target templates and sample pairs in the local search area, as shown in Formula (5).

$$\min_w \frac{1}{N} \sum_{i}^N L(F(x, Wz)).$$  \hfill (5)

3.4. Quantitative Analysis. Compared with THE FCT and ODFS algorithms, this algorithm has better performance on background anti-interference and occlusion problem, is not easy to lose and drift due to environmental interference and occlusion, and has strong robustness to the influence of background clutter scale change occlusion light. For David I and FaceocC2 sequences, although the tracking success rate of this algorithm is lower than that of the HCF algorithm, the real-time performance of this algorithm is very good on the basis of extracting image features with compressed sensing, while the real-time performance of the HCF algorithm is not ideal due to extracting image features with convolutional neural network [18]. Compared with the KCF algorithm, although the tracking success rate for David and FaceooC2 is slightly lower, it is
4. Experiments and Analysis

4.1. Comparison of Different Models. In the experiment, we first establish an independent MDNet vehicle tracking model and then add the channel attention mechanism module to see whether the tracking result is optimal. Finally, two attention mechanism modules are added, combined with case segmentation, to improve our tracking efficiency and solve the problems of vehicle occlusion in the tracking process [21]. We analyzed the differences between these models, calculated these indicators by the cross-validation method, and evaluated the robustness of the algorithm, where F1 represents the harmonic average of accuracy and recall rate, A. It can be seen from the comparison that the accuracy is improved from 90.9% to 90.13% after the addition of channel attention mechanism module. Variance was reduced from 3.78% to 1.27%, with better stability [22]. After the addition of the two attention mechanisms, combined with the image segmentation algorithm, tracking accuracy is higher, and the algorithm is more stable, as shown in Table 2.

Table 1: Track success rate and frame rate.

| Method       | Shaking | David | Face  | Sylvesier | Shaking | David | Faceooc | Sylvesier |
|--------------|---------|-------|-------|-----------|---------|-------|--------|-----------|
| Ours         | 89.5    | 93.4  | 88.1  | 91.2      | 27.8    | 29.1  | 28.3   | 29.5      |
| FCT          | 84.3    | 91.1  | 84.2  | 86.1      | 31.7    | 38.3  | 38.6   | 36.1      |
| ODFS         | 78.9    | 81.2  | 83.2  | 85.4      | 29.6    | 35.1  | 37.4   | 27.6      |
| KCF          | 9.6     | 95.6  | 96.5  | 97.4      | 23.5    | 39.2  | 132.1  |           |
| HCF          | 77.2    | 96.3  | 94.2  | 95.3      | 0.1     | 0.6   | 0.5    |           |

Table 2: Comparison of different models.

| Method                        | Accuracy | Variance | F   | G   | A   |
|-------------------------------|----------|----------|-----|-----|-----|
| MDNet                         | 89.99    | 2.99     | 81.22 | 84.33 | 66.55 |
| Ad channel attention module   | 90.33    | 1.28     | 83.22 | 86.99 | 65.77 |
| The algorithm in this chapter  | 90.66    | 0.65     | 87.66 | 87.33 | 66.21 |

Furthermore, we explore the robustness and accuracy of different time series video data and truncate the test video into units with length ratios of 1, 2, 3, and 4 to form three groups of test data. We input these different data into the model to measure the tracking effect of the model on different video sequence lengths and compare the results. The results show that the shorter the video sequence is, the worse the tracking effect is, and the longer the video sequence is, the better the tracking effect is [23], see Table 3.

4.2. ResNet Network Model. Residual neural network, which for the first time, made it possible to train ultradeep neural networks. It is the first time to train neural network successfully and get good results in computer vision competition. Although the expression ability of convolutional neural network is enhanced with the deepening of layers, the performance degradation of neural network may occur. One reason for this is that the deeper the neural network, the greater the gradient disappearance or explosion. In order to solve the phenomenon of gradient disappearing after neural network deepening, a residual element is proposed, as shown in Figure 3.

4.3. Ablation Experiments. In order to verify the effectiveness of the significant regional weighting strategy in this chapter, ablation experiments were conducted in this chapter, and the algorithm in this chapter was compared with the algorithm without significant regional weighting strategy in VOT2016 and VOT2017. Ours represents no significant region weighting strategy; otherwise, it is the same as the algorithm in this chapter. It can be seen from the two tables that the algorithm in this chapter is superior to the algorithm without significant region weighting strategy in EAO accuracy and robustness. Experimental results show that significant region weighting strategy can improve tracking performance effectively. In addition, compared with ECO, the benchmark tracking algorithm in this chapter, the algorithm...
in this chapter still outperforms ECO in the accuracy and robustness of EAO even without significant regional weighting strategy. The results show that the strategy of using the improved residual network SE-ResNet network to extract multiresolution features from the correlation filtering framework is effective and helpful to improve tracking performance, as shown in Tables 4 and 5.

| Video sequence | Accuracy | Variance | F   | G   | A    |
|----------------|----------|----------|-----|-----|------|
| 1              | 76.5     | 3.20     | 66.32 | 74.21 | 45.23 |
| 2              | 87.23    | 2.50     | 76.35 | 79.66 | 56.33 |
| 3              | 88.54    | 0.89     | 82.54 | 85.44 | 61.55 |
| 4              | 89.34    | 0.67     | 89.64 | 87.55 | 66.89 |

Table 3: Comparison of results of different video sequences.

4.4. Contrast Pooling of Ideas with Other Approaches. In order to extract appearance features from each 3D bounding box candidate constructed from the diagram, we propose a point-attention pooling method to abstract the interactions of internal points. In this part, ablation research is conducted on the use of point attention pooling, set abstraction layer, feature average, or feature maximum. Among them, SA layer is the same as the paper, feature averaging and feature maximization methods are connected behind MLP, and their outputs are one-dimensional features of the same size as our proposed method. We use these four methods to extract proposal capabilities and then use them as appearance features to build diagrams. We use these four methods to extract proposal capabilities and then use them as appearance features to build the diagram. Other settings of the framework we index are consistent with the original network. Below, a 3D bounding box candidate usually contains parts from different objects. Therefore, it is necessary to completely collect points on the surface of the same object and learn the semantic and geometric information associations between them when extracting index region features. We explore the effect of directional features on point-attention pooling, and we can see that it results in a gain of 0.3% compared to the learning index region features using only semantic features and 3D coordinates. Here is our interpretation: direction vector causes points belonging to the same object to attract each other, belonging to different objects mutually exclusive.

5. Conclusion

Aiming at the problem of unreliable spatial information association in complex scenes such as frequent occlusion of visual target interaction, this chapter focuses on the similarity of the same individual features rather than different individual features. The discriminant apparent information can be obtained by training two kinds of classification-based network and discriminant and generative learning network on large scale rerecognition data sets, which can provide reliable matching clues for subsequent data association in multitarget tracking. In this chapter, based on the extended multitarget tracking of spatial information association, the apparent information is transferred to the multitarget tracking process, and the apparent feature measurement method and the matching mechanism of multilayer cue association are constructed. Experiments verify that the multitarget tracking algorithm with the fusion of the apparent information reduces the number of mistaken identity transformation between the targets and improves the stability of the trajectory from the representation ability of the two apparent features, the statistical m-index tracking speed and other aspects. In addition, apparent features based on discriminant and generative learning networks are more reliable in association. There are still many challenges in practical application and popularization of multitarget tracking technology.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.
Acknowledgments

This work was supported by Key Project of Hubei Polytechnic University in 2021, Research on Enterprise Financial Warning Algorithm Based on Big Data (Project No. 21XJZ04A).

References

[1] Y. Ding, L. Hua, and S. Li, “Research on computer vision enhancement in intelligent robot based on machine learning and deep learning,” *Neural Computing and Applications*, vol. 18, no. 1, pp. 778–790, 2020.

[2] O. Abdel Wahab, A. Mourad, H. Otrok, and T. Taleb, “Federated machine learning: survey, multi-level classification, desirable criteria and future directions in communication and networking systems,” in *The IEEE communications surveys & tutorials*, IEEE, 2020.

[3] A. Sharma and R. Kumar, “Performance comparison and detailed study of AODV, DSDV, DSR, TORA and OLSR routing protocols in ad hoc networks,” in *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, IEEE, 2016.

[4] H. Sami, A. Mourad, and W. El Haj, “Vehicular-OBUs-as-on-demand-fogs: resource and context aware deployment of containerized micro-services,” in *IEEE/ACM transactions on networking*, vol. 28, no. 2, pp. 956–979, 2019.

[5] P. Zieliński and U. Markowska-Kaczmar, “3d robotic navigation using a vision-based deep reinforcement learning model,” *Applied Soft Computing*, vol. 110, no. 9, article 107602, 2021.

[6] X. Wu, X. Ling, and J. Liu, “Location recognition algorithm for vision-based industrial sorting robot via deep learning,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 7, p. 1955009.1-1955009.18, 2019.

[7] M. Jahanbakht, W. Xiang, L. Hanzo, and M. R. Azghadi, “Internet of underwater things and big marine data analytics—a comprehensive survey,” *IEEE Communication Surveys and Tutorials*, vol. 23, no. 2, pp. 904–956, 2021, Second quarter.

[8] C. Feng, G. Nie, Q. N. Naveed et al., “Optimization of sorting robot control system based on deep learning and machine vision,” *Mathematical Problems in Engineering*, vol. 2022, 7 pages, 2022.

[9] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering (TNSE)*, vol. 7, no. 2, pp. 766–775, 2020.

[10] S. Zhu, Y. Zhang, S. Coleman, S. Wang, R. Li, and S. Liu, “Semi-supervised learning for person re-identification based on style-transfer-generated data by cyclegans,” *Machine Vision and Applications*, vol. 32, no. 6, 2021.

[11] A. Dhiman, N. Shah, P. Adhikari, S. Kumbar, I. S. Dhanjal, and N. Mehendale, “Firefighting robot with deep learning and machine vision,” *Neural Computing and Applications*, vol. 34, no. 4, pp. 2831–2839, 2022.

[12] H. X. Huynh, N. D. Trung, T. Nguyen, B. H. Le, and T. H. Le, “Maneuverable autonomy of a six-legged walking robot: design and implementation using deep neural networks and hexapod locomotion,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, p. 2021, 2021.

[13] X. Zhou, W. Liang, K. Wang, R. Huang, and Q. Jin, “Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data,” *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 1, pp. 246–257, 2021.

[14] Y. Li, W. Zheng, X. Liu, Y. Mou, L. Yin, and B. Yang, “Research and improvement of feature detection algorithm based on fast,” *Rendiconti Lincei. Scienze Fisiche e Naturali*, vol. 32, no. 4, pp. 775–789, 2021.

[15] W. Shang, Z. Gao, D. Nicolo et al., “Benchmark analysis for robustness of multi-scale urban road networks under global disruptions,” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2022.

[16] R. Stolkin, A. Greig, M. Hodgetts, and J. Gilby, “An em-e-mrf algorithm for adaptive model based tracking in extremely poor visibility,” *Image and Vision Computing*, vol. 26, no. 4, pp. 480–495, 2008.

[17] X. Cheng, B. Yang, A. Hedman, T. Olofsson, H. Li, and L. Van Gool, “NIDL: a pilot study of contactless measurement of skin temperature for intelligent buildings,” *Energy and Buildings*, vol. 198, pp. 340–352, 2019.

[18] S. Yang, B. Deng, J. Wang et al., “Scalable digital neuromorphic architecture for large-scale biophysically meaningful neural network with multi-compartment neurons,” *IEEE transactions on neural networks and learning systems*, vol. 31, no. 1, pp. 148–162, 2020.

[19] H. Wu, “Research and implementation of remote sensing image vectorization,” *Key Engineering Materials*, vol. 439-440, no. 1, pp. 220–224, 2010.

[20] P. Ajay, B. Nagaraj, R. Arun Kumar, R. Huang, and P. Ananthi, “Unsupervised hyperspectral microscopic image segmentation using deep embedded clustering algorithm,” *Scanning*, vol. 2022, Article ID 1200860, 2022.

[21] Z. Lv, L. Qiao, J. Li, and H. Song, “Deep-learning-enabled security issues in the internet of things,” *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9531–9538, 2021.

[22] X. Zhao, X. Liu, J. Liu, J. Chen, F. Shihong, and F. Zhong, “The effect of ionization energy and hydrogen weight fraction on the non-thermal plasma volatile organic compounds removal efficiency,” *Journal of Physics D: Applied Physics*, vol. 52, no. 14, article 145201, 2019.

[23] G. Veselov, A. Tselykh, A. Sharma, and R. Huang, “Applications of artificial intelligence in evolution of smart cities and societies,” *Informatica (Slovenia)*, vol. 45, no. 5, p. 603.