SCIENCE CHINA
Information Sciences

• MOOP •

May 2022, Vol. 65 154201:1–154201:3
https://doi.org/10.1007/s11432-020-3272-y

Target tracking algorithm based on a broad learning system

Dan ZHANG[1,2], Tieshan LI[1,4*], C. L. Philip CHEN[1,3] & He YANG[1]

[1]*Navigation College, Dalian Maritime University, Dalian* 116026*, China;*
[2]*Innovation and Entrepreneurship Education College, Dalian Minzu University, Dalian* 116600*, China;*
[3]*Computer Science and Engineering College, South China University of Technology, Guangzhou* 510641*, China;*
[4]*School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu* 611731*, China*

Received 27 October 2020/Revised 12 March 2021/Accepted 5 May 2021/Published online 7 September 2021

Citation Zhang D, Li T S, Chen C L P, et al. Target tracking algorithm based on a broad learning system. Sci China Inf Sci, 2022, 65(5): 154201, https://doi.org/10.1007/s11432-020-3272-y

Target tracking is one of the most important research topics in the field of computer vision. It has been widely applied in aerospace, automatic monitoring, navigation, human-computer interaction, and artificial intelligence.

Target tracking is a common and difficult task. During the tracking process, target variations are dramatic in terms of scale and position. Additionally, target signals are subject to interference such as occlusion, illumination changes, and background clutter. Therefore, it is of great theoretical and practical value to study target tracking algorithms that can handle large amounts of data, adapt to complex backgrounds, and guarantee real-time performance.

Deep learning can be used to solve the problem of deep information acquisition. Based on the successful application of convolutional neural networks (CNNs) in the image processing direction, CNNs were successfully applied to video tracking. Although deep learning performs well at target tracking, real-time tracking must be improved in terms of its computational cost. A broad learning system (BLS) was proposed in [1]. As an alternative to a deep network architecture, its calculation speed is very fast. Additionally, the BLS can extract sparse features from training data and sparse feature learning models are attractive for exploring essential characterization. Based on these advantages we propose a target tracking algorithm based on BLS using a candidate region search and SURF [2] feature matching of multiple clues. This represents an attempt at applying broad learning to target tracking.

The proposed method was evaluated through extensive experiments and compared to the CT [3], KCF [4], TLD [5], LeNet-5 [6], C-COT [7], and MDNet [8] methods on four datasets to verify its effectiveness. Although the accuracy of our method is not the best on all datasets, it exhibits good adaptability, which is critical for many applications. In terms of tracking time, the proposed method provides the minimal value on various datasets.

Target tracking based on BLS. The detailed process of the BLS tracking algorithm is defined as follows. (1) The tracking datasets are processed by BLS to train an evaluator. (2) In the case of multiple continuous frames containing a target, candidate regions are generated based on the target positions in previous frames using the search method. (3) In the case of consecutive frames with a missing target, the SURF algorithm [2] is adopted for full image feature matching. (4) In either case, the candidate regions are evaluated by a well-trained BLS evaluator.

The first step in this system is to train the BLS evaluator. The tracking dataset for training is X, where $X \in \mathbb{R}^{N\times M}$, N is the number of input samples, and M is the dimension of each sample. ϕ_i is the transformation, and the ith mapped feature is denoted by Z_i. Assume that the number of feature nodes generated at each time instance is N_1. Then,

$$Z_i = \phi_i(XW_{ei} + \beta_{ei}), \quad i = 1, 2, \ldots, n, \tag{1}$$

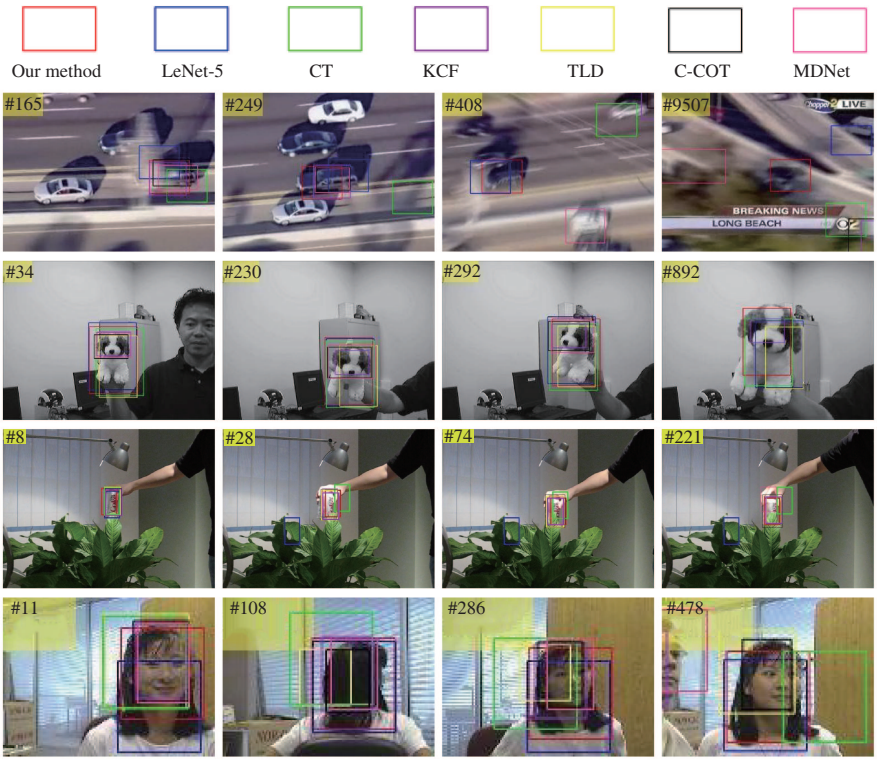
where W_{ei} are random weights and β_{ei} is a bias term. One can obtain the proper corresponding dimensions from input data, X, and N_1. W_{ei} and β_{ei} $(i = 1, \ldots, n)$ are sampled from a normal distribution in the interval $[-1, 1]$. The dimensions of Z_i are $N \times N_1$. Matrix $Z^n \equiv [Z_1, \ldots, Z_n]$ indicates n groups of feature nodes. Z^n is connected to a layer of enhancement nodes. To extract sparse features from input training data, the optimization function is defined as

$$\arg\min_{\hat{W}} : \|Z\hat{W} - X\|_2^2 + \lambda\|\hat{W}\|_1, \tag{2}$$

where $\hat{W}$ is a sparse autoencoder solution and Z is the desired output for the given linear equation. The alternating direction method of multipliers (ADMM) solution to the problem defined above can be written as follows:

$$\arg\min_{W} : f(w) + g(o) \quad \text{s.t. } w - o = 0, \tag{3}$$

* Corresponding author (email: zhangdan@dlnu.edu.cn)



(a)

Method	Carchase (s)		Coke (s)		Dog1 (s)		Girl (s)	
	Training time	Tracking time	Training time	Tracking time	Training time	Tracking time	Training time	Tracking time
Ours	0.29	918.55	0.22	49.21	1.09	197.98	0.96	69.73
LeNet-5	95.99	1063.90	12.75	64.04	104.25	307.40	31.61	96.27
C-COT	NAN	6787.8	NAN	199.43	NAN	920.14	NAN	308.85

(b)

Figure 1 (Color online) Tracking performance of different methods. (a) Tracking effectiveness on four datasets; (b) real-time performance of compared methods.

where $w = W_e$. The optimization iteration is defined as

$$\begin{cases} w_{k+1} := (Z^{\mathrm{T}}Z + \rho I)^{-1}(Z^{\mathrm{T}}x + \rho(o^k - u^k)), \\ o_{k+1} := S_{\frac{\lambda}{\rho}}(w_{k+1} + u_k), \\ u_{k+1} := u_k + (w_{k+1} - o_{k+1}), \end{cases} \tag{4}$$

where $\rho > 0$ and S is the following soft thresholding operator:

$$S_k(a) = \begin{cases} a - k, & a > k, \\ 0, & |a| \leqslant k, \\ a + k, & a < -k. \end{cases} \tag{5}$$

By optimizing the weights, the information contained the input data can be retained and sparsified.

The goal of generating enhancement nodes is to increase the nonlinearity of the network and complement the random feature nodes. The jth group of enhancement nodes is denoted by H_j, and it is assumed that the number of enhancement nodes is N_3. Then, we have

$$H_j = \xi_j(Z^n W_{hj} + \beta_{hj}), \quad j = 1, 2, \ldots, m, \tag{6}$$

where ξ_j is a nonlinear activation function, W_{hj} are random weights, and β_{hj} is a bias term. One can obtain the appropriate corresponding dimensions based on the input data Z^n and N_3. W_{hj} and β_{hj} $(j = 1, \ldots, m)$ are sampled from a normal distribution in the interval $[-1, 1]$. Matrix $H^m \equiv [H_1, \ldots, H_m]$ is used to denote group m of the enhancement nodes.

The output is set to Y $(Y \in \mathbb{R}^{N \times C})$, where C is the dimension of the corresponding outputs. Then, the broad learning system can be written as

$$\begin{aligned} Y =& [Z_1, \ldots, Z_n | \xi(Z^n W_{h1} + \beta_{h1}), \ldots, \xi(Z^n W_{hm} + \beta_{hm})] W_m \\ =& [Z_1, \ldots, Z_n | H_1, \ldots, H_m] W_m \\ =& [Z^n | H^m] W_m, \end{aligned} \tag{7}$$

where $W_m = [Z^n|H^m]^+ Y$ and W_m are the connection weights of the BLS. W_m are computed using the ridge regression approximation. After setting $[Z^n|H^m]$ to A, for the pseudoinverse we have that

$$A^+ = \lim_{\lambda \to 0}(\lambda I + AA^{\mathrm{T}})^{-1} A^{\mathrm{T}}. \tag{8}$$

The computation of W_m is extremely rapid through the approximation equation (8), where λ is a regularization parameter set to 10^{-8}. In this manner, the entire broad learning network is trained.

The second step in this system is to obtain effective candidate regions. According to different scenarios of object

loss and occlusion, the tracking process is conducted under the framework by selecting the candidate region search method or the SURF feature matching algorithm. Case 1: If the object is not occluded or lost for a long period of time, to consume less resources, the positions around the target object are selectively searched. The search boxes are set to different sizes for different tracking targets. At the target location and its surroundings, nine windows of equal size with a step length of four pixels are generated. After removing any windows intersecting the image border, the remaining windows are selected as candidate regions. Because the candidate regions for evaluation are only selected within a small range, the number of calculations is reduced. Case 2: If the target is occluded or lost for a long time (five frames), then the feature matching based on the SURF algorithm and a full-image target search are performed to ensure accuracy. If the number of identified points is greater than four, then the target is considered to be found and a weighted average is calculated to represent the center of the feature points. Therefore, the proposed method can adapt to target tracking with deformation, occlusion, and significant loss while maintaining acceptable tracking speed.

The third step in our method is to evaluate the candidate regions and select the window with the highest evaluation score as the position of the target. However, if the scores of all candidate regions are very low (less than a predetermined threshold), then our method judges that the target is lost and counts the number frames in which the target is lost.

Experiments and analysis. One tracking dataset "Carchase" from TLD [4] and three tracking datasets "Coke", "Dog1", and "Girl" from OTB [9] were used to evaluate the efficiency of our method. Target occlusion, loss, illumination, and scale variations exist in these video sequences. In our experiments, the structure of the BLS consisted of 10×6 feature nodes and 150 enhancement nodes. The tracking performance is presented in Figures 1(a) and (b). Because of a lack of relevant data support for C-COT, there were no data for comparison for this method at training time. The tracking effectiveness is shown in Figure 1(a). The training and the tracking times are presented in Figure 1(b).

As shown in Figure 1(a), the proposed algorithm provides good tracking performance in terms of accuracy and adaptability. In Figure 1(b), the training time and tracking time of our method for different datasets are significantly lower than those obtained by the other methods. This demonstrates that our method not only has good adaptability, but also good real-time tracking performance.

Conclusions and discussion. Based on our experimental results, the following two aspects will be discussed. (1) Tracking adaptability: First, we trained an accurate information evaluator. BLS is the process of acquiring sparse features. Sparse feature learning models are attractive for exploring the essential characteristics of tracking data. Based on statistical target occlusion and loss, we can adjust the candidate region search and SURF feature matching. By using such a tracking strategy, we can effectively enhance tracking adaptability. (2) Time consumption: Our method has a small time overhead because BLS has few parameters and is solved using ridge regression. Experimental results demonstrated the effectiveness of the proposed algorithm, but it still has some deficiencies. Taking tracking speed as an example, there is still room for further improvement. Robustness to long-term occlusion and loss, as well as scale variation, must be improved. Additionally, online tracking must be implemented. Follow-up research will focus on these issues and additional studies will be required to develop faster and more robust online target tracking systems.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 51939001, 61976033, U1813203, 61803064, 61751202), Natural Foundation Guidance Plan Project of Liaoning (Grant No. 2019-ZD-0151), Science & Technology Innovation Funds of Dalian (Grant No. 2018J11CY022), and Fundamental Research Funds for the Central Universities (Grant No. 3132019345).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Chen C L P, Liu Z L. Broad learning system: an effective and efficient incremental learning system without the need for deep architecture. IEEE Trans Neural Netw Learn Syst, 2018, 29: 10–24
2 Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF). Comput Vision Image Underst, 2008, 110: 346–359
3 Zhang K H, Yang M H. Real-time compressive tracking. In: Proceedings of the 12th European Conference on Computer Vision, Florence, 2012. 864–877
4 Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 583–596
5 Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 1409–1422
6 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324
7 Danelljan M, Robinson A, Khan F S, et al. Beyond correlation filters: learning continuous convolution operators for visual tracking. In: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, 2016. 472–488
8 Hyeonseob N, Bohyung H. Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 4293–4302
9 Wu Y, Lim J, Yang M H. Object tracking benchmark. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 1834–1848