Visual Target Tracking Based on Online Discriminant Feature Enhancement

Min Wu, Yufei Zha*, Bing Chen
School of Aerospace Engineering, Air Force Engineering University, Xi'an, China

*Corresponding author e-mail: 1820304877@qq.com

Abstract: Most of the current similarity algorithms adopt the idea of offline training and online tracking. Firstly, the network is trained offline through a large amount of data, and then the common features of the target are extracted according to the trained network to realize online tracking. Since the learning is a common feature of a large amount of data in the offline training process, different targets have different characteristics. When performing online tracking, the use of common features to express a specific target has its limitations. Therefore, on this basis, this paper increases the online update of the network, and updates the network parameters online, so that the network learns the specific characteristics of the current target and improves the ability of the network to express specific targets. We evaluate the proposed approach on the tracking benchmark: OTB50, OTB100 and VOT2015 dataset. A large number of experimental results show that our algorithm can improve the discriminating ability of deep features and achieve accurate tracking.

1. Introduction

Target tracking is one of the hotspots in the field of computer vision field. In the past few decades, the research on target tracking has made great progress. The current target tracking methods mainly include similarity matching, correlation filtering, SVM [1], and so on. With the promotion of deep learning in the field of target tracking, the target tracking algorithm based on similarity matching has been rapidly developed. YCNN [2] is the first algorithm that combines deep learning with similarity matching. It trains network parameters offline through a large amount of data, and then uses the trained network to achieve online tracking. Siamese fc is a similarity matching algorithm published in ECCV 2016. The algorithm trains the network offline through the VID [3] database, learns the common features between similar targets. When tracking online, the trained network is used to extract the discriminative features of the target, and the exact location of the target is obtained.

Figure 1 shows the heat diagram of our algorithm and Siamese fc. the left of the figure are sequences with similar distractors, the red frame is the target, and the right is the heat diagram of our algorithm and Siamese fc. It can be seen from the figure that when similar distractors appear around the target, Siamese fc can't distinguish the target and distractors well. Our algorithm is updated online, which improves the network's ability to express the target and can distinguish the target well.

In summary, the algorithm based on the similarity matching only learns the commonality, and cannot realize the problem of expressing the target. The article proposes the idea of train network parameters online in the online tracking process, let the network learn the individuality of each target and improve the ability of the network to discriminate target and background.
2. Visual Target tracking based on online discriminant feature enhancement

On the basis of Siamese fc, in order to improve the expressive ability of the network to the specific target, this article proposes a visual target tracking based on online discriminant feature enhancement. The basic concept is divided into two steps: offline training commonality and online updating individuality. Offline training aims to train network parameters through a large amount of data, and learn common features between similar targets; online update is to train the network according to the current target, and extract individual features of the current target. The commonality and individuality can maximize the response to the current target and improve the ability of the network to express the current target. Figure 2 is a block diagram of an online update network. Where x is determined by the target of the first frame, and the input template is 125×125×3; z is the search region, and is the positive sample saved for different frame, the size is 255×255×3. The target template X and the search region Z are input into the network, and the deep features are extracted, and the network parameters are updated together with the sample label.

2.1 Commonality

In this article, the network is an AlexNet network with five layers of convolutional layer. In the offline training process, this article uses the idea of Siamese fc to realize the training of network parameters by using VID database, and learns the common features between similar targets. Loss
function \( f \) is the logistic. During the tracking process, the parameters of the first three layers are kept unchanged, and the common features of the target are extracted.

\[
v_{\text{com}} = \text{corr}(f(x), f(z))
\]

In the formula, \( f(x) \) is common features of the target template \( x \), \( f(z) \) is the common features of the search region \( z \), \( v_{\text{com}} \) is the response of target template and search region. As can be seen from the figure 2, the common feature can extract similar parts of the target template \( x \) and the search region \( z \), and achieves a distinction between the target and the background.

### 2.2 Individuality

As can be seen from Figure 2, offline training learns the common features between similar targets, but the common features can produce a large response to targets and similar distractors at the same time, and cannot be distinguished. This article is based on offline training networks, and online training parameters of C4 and C5 according to the sample. Through offline training and online update, the network can simultaneously express the commonality and individuality of the target, and achieve a better expression of the current target.

As shown in Figure 2, the input template is \( X \), which is the target frame of the first frame, and \( z \) is the search region. First, enter the template \( x \) and the search area \( z \) into the network, and extract the common features from the first three layers. After the common features are passed through the last two layers, the individuality of the current target is learned.

\[
v_{\text{par}} = \text{corr}(g(x), g(z))
\]

In the formula, \( v_{\text{par}} \) is the response of the template and the search region. \( g(x) \) is the individual features of the target template \( x \), \( g(z) \) is the individual features of the search region \( z \). The response value and the label are simultaneously brought into the logistic function to update the parameters of the last two layers.

\[
\arg\min \frac{1}{N} \sum_{i=1}^{N} L(v_{\text{par}}, y_i)
\]

As can be seen from Figure 2, the feature obtained by the last layer of the convolution layer can distinguish the target from the background, and the template only responds to the part of the search region that belongs to the target.

### 3. Experiment

#### 3.1 Experimental Analysis

**Model update:** In order to prove the improvement of the algorithm after adding the network update, this article conducts experiments in the OTB100 database, and explains the algorithm separately for different update conditions: first frame update, interval update (interval), failure update (failure) and overall performance (overall_performance). The first frame update uses the template of the first frame and the search region to train the network parameters. The interval update is to update the network parameters by using the saved samples every 30 frames, and the failure update is the network parameter when the tracking fails. The update is performed, and the overall performance is the result of the first frame update, the interval update, and the failed update.
As can be seen from Figure 3, the experimental results after only the first frame update are lower, and can only reach 0.669 and 0.491. After the interval update is added, the experimental results are greatly improved by updating the network parameters every 30 frames, indicating that using the time information to update the network parameters can make the network model more adaptable to the target time domain change. The failure update is added on the basis of the first frame update. When the tracking failure occurs, the network model is updated by using the saved positive samples to improve the discriminating ability of the network model. When the first frame update, interval update and failure update are used at the same time, the accuracy and success rate of the algorithm are greatly improved, which are 0.833 and 0.609 respectively, and is increased by 8.5% and 5% based on the baseline.

**Qualitative results comparison.** Offline training learns the common features between similar targets, and updates the individuality of the current target online. By combining the common features and individual features of the target, the ability to discriminate against the current target can be improved. In order to prove this decision-making ability, this article selects six video sequences with various tracking difficulties from the OTB100 database, and compares them with DCFNet, HCF [4], CNN-SVM [5], siamfc3s, siamese_yuan to verify.

As can be seen from Figure 4, in the face of similar targets, scale transformation and occlusion, the algorithm can distinguish the target from the background and adapt to the time domain of the target. Compared with each algorithm, our algorithm is robust to each tracking difficulty and can achieve accurate tracking.

**3.2. Overall performance comparison**

**OTB50 database.** In the OTB50 database, this article selects DCFNet, MUSTer [6], FCNT [7], CNN-SVM, SINT, SRDCF [8], MEEM [9], siamfc3s, siamese_yuan as the comparison algorithm with our algorithm. DCFNet, SINT, FCNT is the algorithm based on two-way network. siamfc3s is the experimental result published by the Siamese fc. siamese_yuan is the experimental result obtained by the code published by the author, and is also the basic algorithm of this article. As can be seen from the
figure, compared with the basic algorithm, the algorithm of this article achieves an accuracy of 11% and a success rate of 7.1%.

**OTB100 database.** The OTB100 database is an improved database on the OTB50 database. In order to prove the effectiveness of the algorithm, this article also conducted comparative experiments on the OTB100 database. The comparison algorithms are DeepSRDCF [10], HDT [11], HCF, DCFNet, CNN-SVM, SRDCF, MEEM, Siamese_yuan.

As shown in the figure, the accuracy of the algorithm in this chapter is 83.3%, the success rate is 60.9%, and the increase is 8.5% and 5% respectively on the basis of baseline. The experimental results show that the performance of the tracking algorithm can be improved by increasing the online update operation during the online tracking phase and learning the commonality and personality of the target.

**VOT2016 database.** VOT2016 is one of the most popular databases in the current target tracking area. Therefore, this chapter uses the vot2016 database to compare with the current advanced algorithms to prove the advanced nature of the algorithm. The comparison algorithms are mainly ACT, MLDF, STC, DAT, DFT, BDF, DFST.

**Table 1** Comparison of VOT2016 database results

| Algorithm | A-Rank | Failures | EAO | AUC |
|-----------|--------|----------|-----|-----|
| baseline  | 0.52   | 23.97    | 0.27| 0.40|
| unsupervised | 0.43   | 42.60    | 0.17| 0.28|
| ACT       | 0.50   | 15.04    | 0.32| 0.44|
| MLDF      | 0.37   | 66.18    | 0.11| 0.15|
| STC       | 0.45   | 28.35    | 0.21| 0.30|
| DAT       | 0.44   | 59.61    | 0.14| 0.21|
| DFT       | 0.37   | 51.45    | 0.14| 0.18|
| BDF       | 0.46   | 50.39    | 0.15| 0.32|
In the table, red represents the performance-first algorithm, blue is the performance-ranked algorithm, and green is the third-performance algorithm. By comparing the results, the algorithm in this chapter achieved the first place on the overlap and the second place on the Failures and EAO [12]. Experiments show that the overall performance of the tracking algorithm can be improved by increasing the online update operation.

4. Conclusion

At present, the algorithm based on similarity matching only uses the offline training network parameters to learn the general similarity between similar targets. In the online tracking process, because the tracking target is a specific target, the general feature cannot achieve the expression of the current specific target. Based on such problems, this chapter adds online update of the network based on the similarity algorithm. By training the network parameters online, the network learns the specific features of the current specific target based on the general characteristics, and combines the general features with the specific features. Achieve expression of the target. Through a large number of experiments, it is shown that the combination of general features and specific features can achieve better expression of current target.

References

[1] Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression[J]. Neural Networks, 2004, 17(1):113-126.
[2] Chen K, Tao W. Once for All: a Two-flow Convolutional Neural Network for Visual Tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017:1-1.
[3] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-Convolutional Siamese Networks for Object Tracking[C]. European Conference on Computer Vision. 2016.
[4] Ma C, Huang J B, Yang X, et al. Hierarchical Convolutional Features for Visual Tracking[C]/ 2015 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society, 2015.
[5] Hong S, You T, Kwak S, et al. Online tracking by learning discriminative saliency map with convolutional neural network[C]. International Conference on International Conference on Machine Learning, 2015.
[6] Zhibin H, Chen Z, Wang C, et al. MULTI-STORE Tracker (MUSTer): a Cognitive Psychology Inspired Approach to Object Tracking[C]/ IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2015.
[7] Wang L, Ouyang W, Wang X, et al. Visual Tracking with Fully Convolutional Networks[C]. IEEE International Conference on Computer Vision. 2016.
[8] Danelljan M, Gustav Häger, Khan F S, et al. Learning Spatially Regularized Correlation Filters for Visual Tracking[C]. 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2016.
[9] Zhang J, Ma S, Sclaroff S. MEEM: Robust Tracking via Multiple Experts Using Entropy Minimization[J]. 2014.
[10] Danelljan M, Gustav Häger, Khan F S, et al. Convolutional Features for Correlation Filter Based Visual Tracking[C]. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). IEEE, 2016.
[11] Qi Y, Zhang S, Qin L, et al. Hedged Deep Tracking[C]. Computer Vision & Pattern Recognition. 2016.
[12] Henriques J F, Rui C, Martins P, et al. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels[M]/Computer Vision-ECCV 2012. Springer Berlin Heidelberg, 2012:702-715.