Single Target Tracking Using Reliability Evaluation and Feature Selection

Li Wei
Jincheng College
Nanjing University of Aeronautics and Astronautics
Nanjing, China
e-mail: carry_wei@nuaa.edu.cn

Meng Ding and Xu Zhang
School of Civil Aviation
Nanjing University of Aeronautics and Astronautics
Nanjing, China
e-mail: nuaa_dm@nuaa.edu.cn

***Abstract*—This paper proposes a visual tracking algorithm using reliability evaluation and feature selection mechanism in the framework of Correlation Filter (CF). Three additional modules are used to improve the current CF-based tracker. Firstly, a module of reliability evaluation is used to determine whether the current tracking result is reliable. Secondly, an updating module is used to determine whether to update the target model by comparing the reliability of current tracking result with historical average. Thirdly, a feature selection module is presented to select hand-crafted feature or deep convolutional feature according to the current tracking state. Experimental results on a benchmark dataset of fifty challenging test sequences show that the proposed method can reduce the interference of complex factors effectively.**

Keywords- Visual tracking; feature selection; reliability evaluation; convolutional neural network; threshold segmentation.

I. INTRODUCTION

Visual tracking is a remarkable research issue in computer vision and image understanding, with wide-ranging applications including intelligent surveillance, robot and unmanned aerial vehicle. Generally, existing tracking algorithms can be divided into two types as generative algorithm and discriminative algorithm. The former describes target by learning an appearance model and searches for the most similar area in a new frame as the result. The latter extracts the distinguishing feature of target and then uses a discriminative classifier to detect the target from backgrounds. Recently, as one of classic discriminative algorithms, the trackers based on correlation filter (CF) have achieved outstanding performance in different benchmarks. Furthermore, features from pre-training convolution neural network (CNN) are substituted for the hand-crafted features to further improve the performance of correlation filter-based methods [1,2].

This paper proposes a visual tracking method which uses reliability evaluation and feature selection in the framework of CF [3,4]. The main contributions of this paper can be summarized as follows: Firstly, this paper introduces an effective mechanism to evaluate the current tracking state by analyzing the response map obtained from the test sample and the target model. Secondly, the tracker determines whether to update the target model by comparing the reliability of current tracking result with historical average. Thirdly, a feature selection method is used to integrate the performance advantages of deep features and the speed advantages of traditional features. Extensive experimental results demonstrate that the proposed tracking algorithm achieves a higher accuracy and robustness and performs favorably against the other comparison algorithms.

II. THE PROPOSED ALGORITHM

A. Discriminant Correlation Tracking with Multichannel

Firstly, according to the initial location and size of the target, the proposed tracking algorithm extracts the feature map of corresponding image patch containing the target. The feature map f is composed of d-channel features and f^l denotes a feature channel of feature map f. A combination of correlation filters h which contains d filters corresponding to d feature channels can be achieved by minimizing the L2 error ε between the expected output g and the correlation response [5],

$$\varepsilon = \left\| g - \sum_{l=1}^{d} h^l \bigstar f^l \right\|^2 + \lambda \sum_{l=1}^{d} \left\| h^l \right\|^2 \qquad (1)$$

where $\bigstar$ means the correlation operation. g is the desired correlation output and a 2D Gaussian function centered in the target. λ is regularization parameter to alleviate the over fitting. As a linear least square problem, Eq.(1) can be quickly solved in the frequency domain by using the Parseval's theorem. The filter is given as

$$H^l = \frac{\overline{G}F^l}{\sum_{k=1}^{d} \overline{F^k} F^k + \lambda}, \; l = 1,...,d \; . \qquad (2)$$

Here, the capital letters represent the discrete Fourier transform (DFT) of the corresponding variables. The bar $\overline{G}$ is used to denote the complex conjugate of the variable. Since all operations of Eq. (2) are pointwise operations, the computation in frequency domain is significantly reduced compared with in spatial domain. The solution of Eq. (2) is the optimal filter h. However, the feature map f_t from different frames t should be taken into account to obtain a robust correlation filter. Therefore, we can obtain the optimal model by using the average of correlation errors of Eq.(1) over the feature maps $f_1,\ldots,f_t$ at all frames. However, the cost of computation is too large for online tracking tasks. A better method is to obtain a robust approximate solution using the precise solution of a single feature map from Eq. (2). Specifically, the numerator A_t^l and denominator B_t^l of the filter is updated by using a new feature map f_t as

$$A_t^l = (1-\eta) A_{t-1}^l + \eta \overline{G} F_t^l, \; l = 1,...,d \qquad (3)$$

978-1-7281-4653-9/19/$31.00 ©2019 IEEE
DOI 10.1109/ISCID.2019.00059

$$B_t = (1-\eta)B_{t-1} + \eta\sum_{k=1}^{d}\overline{F_t^k}F_t^k \qquad (4)$$

Where η is a learning rate. For the following frame t, the feature map z_t is extracted from the image patch with the center of the prediction location in the previous frame. In the frequency domain, the DFT of y_t is calculated by

$$Y_t = \frac{\sum_{l=1}^{d}\overline{A_{t-1}^l Z_t^l}}{B_{t-1}+\lambda} \qquad (5)$$

The location of target in current frame can be estimated based on the maximum of the response y_t which is the inverse DFT to the result of Eq. (5) $y_t = \mathcal{F}^{-1}\{Y_t\}$.

B. Tracking State Evaluation

Most trackers update the model directly without considering the reliability of current tracking result. In fact, if the estimated position is inaccurate, updating the model continuously is likely to lead to tracking failure. To solve this problem, the proposed method evaluates the confidence of the current detection result based on the feedback of the response. Based on the evaluation results, the tracker determines whether updating the model and whether using the deep convolutional feature. To some extent, the maximum response and the shape of response map reflect the confidence of the tracking results. When the features extracted from the test sample match the learned model, the correlation response map should have a sharp and single peak, remaining smooth in other regions. The steeper the single peak is, the more reliable the tracking result will be. On the contrary, if there are obviously fluctuations in the response map, it means that the confidence of the tracking results is very low. If we continue using the currently incorrect sample to update the target model, the model will learn the feature from the background or something cover the target in subsequent frames. In this situation, the response from the background or some cover will be greater than the response from the real target.

Our tracker evaluates the confidence of the tracking result by two evaluation indexes. The first one is the maximum response F_{max} in the response map $F_{response}$, defined as

$$F_{\max} = \max F_{response} \qquad (6)$$

Usually, the higher the maximum response is, the more reliable the tracking result is. However, the single index of maximum response does not reflect the fluctuation of the response map. For this reason, we use a second index called Area Ratio of Response (*ARR*), which is defined as

$$ARR = \frac{numel\left(find\left(otsu\left(F_{response}\right)\right)==1\right)}{area\left(F_{response}\right)}. \qquad (7)$$

where, $otsu(F_{reponse})$ denotes the binary image of the response map by using Otsu method with an adaptive threshold. The function $numel(find(.)==1)$ denotes the number of pixels whose value is equal to one in the binary image, and $area(.)$ denotes the area of the binary image.

C. Model Updating

This paper implements a selective update strategy according to the current tracking state. The tracker updates the model continuously in the first T frames of the video sequence to retain some historical information of the evaluation indexes. After the first T frame, the tracker evaluates the confidence of current tracking result. If the following two conditions are met simultaneously, current tracking result will be trusted to have a high confidence and then the tracker updates the target model.

$$F_{\max,t} > \theta mean(F_{\max,2},...,F_{\max,t}) \qquad (8)$$

$$ARR_t < \delta mean(ARR_2,...,ARR_t) \qquad (9)$$

The former means the maximum response in the current frame should be higher than the historical average, and the latter means the *ARR* index in the current frame should also be lower than the historical average. If these indexes do not satisfy the above conditions, the tracking result will be too unreliable to update the target model.

D. Deep Convolutional Feature

The proposed method integrates the performance advantages of deep features and the speed advantages of traditional features. When tracking in a simple scene without complex interference, hand-crafted feature like HOG is good enough to estimate the target position accurately and efficiently. When tracking a target in complex scenes, the deep feature is more discriminative to track the target than traditional feature. Therefore, we set up a feature selection mechanism to combine these two kind of feature. The hand-crafted feature is the main feature and deep convolutional feature is an auxiliary feature. When the tracking result is unreliable, the tracker uses convolution neural network, such as AlexNet22 and VGG-Net23 network to exact feature and estimate the target location. Furthermore, we evaluate the tracking result by the response map of deep features too. If the *ARR* index of the deep feature response satisfies Eq. (10), the tracking result is adopted and the target model of the handcraft feature is updated, where μ is a peak area ratio set manually. Otherwise, the tracking result obtained by the deep feature model is not adopted because the tracking result is still unreliable. The target may be under the interference of occlusion or other interference. The tracker continues to follow the results of hand-crafted features. In addition, considering the advantage of semantic discriminative information and real-time performance, we only learn and update the deep feature model only in the first k frame of the video sequence.

$$ARP_{deep} < \mu \qquad (10)$$

On the other hand, as the increase of the number of convolution layers, the spatial resolution of the outputs of convolution layers gradually decreases. This low resolution is difficult to satisfy the request of locating the target accurately. Consequently, this paper uses the outputs of the last convolution layer in the network structure as the deep feature, and we expand the size of the feature by bilinear interpolation to locate the target more accurately.

E. Algorithm implementation

According to the above discussion, the proposed tracking method using reliability evaluation and feature selection is summarized as follows:

The proposed tracking algorithm

Algorithm 1. The proposed tracking algorithm

Input: initial target bounding box x_0

Output: estimated object state $x_t = (\hat{x}_t, \hat{y}_t, \hat{s}_t)$, handcraft feature model H_h and deep feature model H_d

1. **repeat**
2. Crop out the searching window in frame t according to $(\hat{x}_{t-1}, \hat{y}_{t-1})$ and extract the handcraft features;
   // Translation estimation
3. Compute the correlation response map $F_{response}$ using H_h and Eq. (5) to estimate the new position (x_t, y_t) , using Eq. (6) and Eq. (7) compute $F_{\max}$ and ARR ;
   //Re-detection
4. **if** the tracking result is not reliable
5. Crop out the searching window in frame t according to $(\hat{x}_{t-1}, \hat{y}_{t-1})$ and extract the deep features;
6. Compute the correlation response $F_{response}$ using H_d and Eq. (5) to estimate the location (x_{t_cnn}, y_{t_cnn}) , using Eq. (7) compute ARR;
7. **if** the tracking result is reliable $(x_t, y_t) = (x_{t\_cnn}, y_{t\_cnn})$;
8. **end**
9. **end**
   // Scale estimation
10. Estimate the optimal scale $\hat{s}_t$;
    // Model update
11. **if** the first frame
12. Using Eq. (4) learning model H_h both and H_d ;
13. **else if** the tracking result is not reliable
14. Update the hand-crafted model H_h ;
15. **end**
16. **until** End of video sequences

III. Experimental Results

In our experiment, PCA-HOG [6] is used as the handcrafted feature for target representation. The feature extracted by VGG-16 network trained on ImageNet is used as the deep feature, which can be obtained from the Matconvnet toolkit. The proposed algorithm is implemented in MATLAB 2016a and all the evaluation algorithms run on a 3.40GHz PC with 16GB RAM. The specific parameters in this paper are set as follows: The regularization parameter of Eq. (1) equals to $\lambda=10^{-2}$, and the learning rate is set to $\eta=0.025$. The size of the search window is set to 2 times of the target size for handcraft model as 2.5 times for deep model generally. The standard deviation of the desired Gaussian function output is set to 1/16 of the target size for handcraft model as 1/5 of the target size for deep model. The number of frames to update the deep feature model is set to $k=3$. The related parameters in the tracking result discriminant mechanism are set to $\theta=0.4$, $\delta=3$, $\mu=0.2$. In the experiments, the parameters of the tracker are fixed.

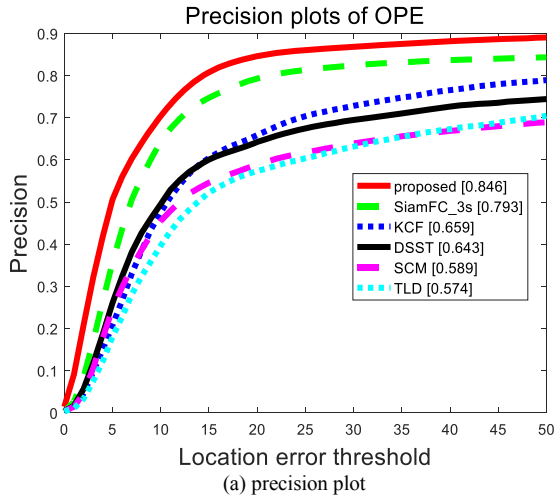
We assess the proposed method on a large benchmark dataset OTB-2013 [7] that contains 50 test sequences. The test method is used to track the target by frame by frame after initializing the initial frame in the sequence. The contrast algorithms used in the experiment are set according to the open source code in the database.

Three criteria are used for quantitative performance evaluation [7]: (1) Center location error (CLE) indicates the average Euclidean distance between the ground-truth and the estimated center location. (2) Precision rate (PR) demonstrates the percentage of frames whose estimated location is within the given threshold distance (20 pixel generally) of the ground truth. (3) Success rate (SR), which is defined as the percentage of frames where the bounding box overlap surpasses a threshold (50% generally).

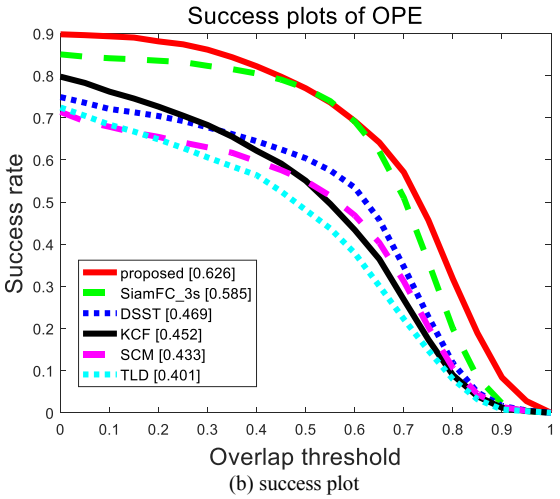
We evaluate the proposed algorithm using deep convolutional feature and discriminant mechanism (DFDM) on the benchmark with comparisons to five state-of-the-art trackers. These five trackers come from three typical categories of tracking algorithms: (1) tracking with correlation filter (DSST[5], KCF[3]), (2) tracking with multiple online classifiers (TLD[8], SCM[9]) (iii) tracking with convolution neural network (SiamFC[10]). We show the results in one-pass evaluation (OPE) using the precision rate and success rate as shown in Figure. 1, where the legend contains the AUC score for each tracker. The results shown in Figure.1 illustrate that the proposed approach performs well against the existing methods in OPE. Moreover, we present the quantitative comparisons of precision rate where the given threshold distance is equal to 20 pixels, AUC score and tracking speed in Table 1. The speeds are from the original paper as the default processor is CPU. The first and second maximum values are highlighted by bold and underline. The proposed method performs favorably against existing methods in precision rate (PR) and AUC score with a better real-time performance. Figure.2 illustrates the tracking results in several test sequences and shows the proposed algorithm performs favorably against the other five algorithms.

IV. Copyright Forms and Reprint Orders

In this paper, we propose an effective algorithm based on correlation filter. The proposed method can evaluate the current tracking state effectively by analyze the response map. According to the current tracking state, the tracker determines whether the samples are reliable and whether to update model. We further use the deep convolutional feature to track targets in case of unreliable tracking result. Both the quantitative and qualitative experimental results show that the proposed algorithm is superior to the existing methods in terms of accuracy, efficiency and robustness.



(a) precision plot



(b) success plot

Figure. 1 The overall performance of OPE on OTB-2013

Figure. 2 Tracking results of all the six algorithms on five challenging sequences. (From top to down are Jogging-2, Shaking, Soccer, Singer2 and Fleetface.)

TABLE I. COMPARISONS WITH STATE-OF-THE-ART TRACKERS ON OTB-2013.

Tracker	*PR(%)*	*AUC(%)*	*Speed(FPS)*
KCF	65.9	40.1	**172**
DSST	64.3	46.9	24
SCM	58.9	43.3	0.5
TLD	57.4	42.9	21
SiamFC_3s	79.3	58.5	86(GPU)
Proposed	**84.6**	**62.6**	19

ACKNOWLEDGMENT

This research has been supported by the National Natural Science Foundation of China (No.61673211, No.U1633105, No.61203170), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No.18KJB590002), and Aeronautical Science Foundation of China (No.20155152041).

REFERENCES

[1] C. Ma, J. B. Huang, X. Yang, et al., "Hierarchical convolutional features for visual tracking," In: *IEEE international conference on computer vision*, Santiago, Chile, 7-13 Dec. 2015. pp. 3074-3082.

[2] M. Danelljan, G. Hager, K. F. Shahbaz, et al., "Convolutional features for correlation filter based visual tracking," In: *IEEE International Conference on Computer Vision Workshops*, Santiago, Chile, 7-13 Dec. 2015. pp.58-66.

[3] J F Henriques, R Caseiro, P. Martins, et al., "High-speed tracking with kernelized correlation filters", *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp.583-596, Mar. 2014.

[4] M. Danelljan, K. F. Shahbaz, M. Felsberg, et al., "Adaptive color attributes for real-time visual tracking," In: *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, 23-28 Jun. 2014. pp. 1090-1097.

[5] M. Danelljan, G. Hager, F. S. Khan, et al., "Discriminative scale space tracking", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561-1575, Aug. 2017.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, et al., "Object Detection with Discriminatively Trained Part-Based Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, Sep. 2010.

[7] Y. Wu, J. Lim, and M. H. Yang, "Online Object Tracking: A Benchmark," In: *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, USA, 23-28 Jun. 2013. pp. 2411-2418.

[8] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-Learning-Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409-1422, Jul. 2012.

[9] W. Zhong, H. Lu, and M. H. Yang, "Robust object tracking via sparsity-based collaborative model," In: *IEEE Conference on Computer Vision and Pattern Recognition*, Rhode Island, 18-20 Jun. 2012. pp. 1838-1845.

[10] L. Bertinetto, J. Valmadre, J. F. Henriques, et al., "Fully-Convolutional Siamese Networks for Object Tracking," In:*European Conference on Computer Vision*, Amsterdam, The Netherlands, 8-16, Oct. 2016. pp. 850-865.

1. Single Target Tracking Using Reliability Evaluation and Feature Selection

Accession number: 20203409061886
Authors: Wei, Li (1); DIng, Meng (2); Zhang, Xu (2)
Author affiliation: (1) Nanjing University of Aeronautics and Astronautics, Jincheng College, Nanjing, China; (2) Nanjing University of Aeronautics and Astronautics, School of Civil Aviation, Nanjing, China
Source title: Proceedings - 2019 12th International Symposium on Computational Intelligence and Design, ISCID 2019
Abbreviated source title: Proc. - Int. Symp. Comput. Intell. Des., ISCID
Volume: 1
Part number: 1 of 2
Issue title: Proceedings - 2019 12th International Symposium on Computational Intelligence and Design, ISCID 2019
Issue date: December 2019
Publication year: 2019
Pages: 228-231
Article number: 9098305
Language: English
ISBN-13: 9781728146522
Document type: Conference article (CA)
Conference name: 12th International Symposium on Computational Intelligence and Design, ISCID 2019
Conference date: December 14, 2019 - December 15, 2019
Conference location: Hangzhou, China
Conference code: 159910
Publisher: Institute of Electrical and Electronics Engineers Inc.
Abstract: This paper proposes a visual tracking algorithm using reliability evaluation and feature selection mechanism in the framework of Correlation Filter (CF). Three additional modules are used to improve the current CF-based tracker. Firstly, a module of reliability evaluation is used to determine whether the current tracking result is reliable. Secondly, an updating module is used to determine whether to update the target model by comparing the reliability of current tracking result with historical average. Thirdly, a feature selection module is presented to select hand-crafted feature or deep convolutional feature according to the current tracking state. Experimental results on a benchmark dataset of fifty challenging test sequences show that the proposed method can reduce the interference of complex factors effectively.

Number of references: 10
Main heading: Target tracking
Controlled terms: Electric current control - Feature extraction - Intelligent computing - Reliability - Statistical tests
Uncontrolled terms: Benchmark datasets - Complex factors - Correlation filters - Current tracking - Reliability Evaluation - Selection mechanism - Single target tracking - Visual tracking algorithm
Classification code: 723.4 Artificial Intelligence - 731.3 Specific Variables Control - 922.2 Mathematical Statistics
DOI: 10.1109/ISCID.2019.00059
Funding Details: Number: 20155152041, Acronym: -, Sponsor: Aeronautical Science Foundation of China; Number: 18KJB590002, Acronym: -, Sponsor: Natural Science Research of Jiangsu Higher Education Institutions of China; Number: 61203170, Acronym: NSFC, Sponsor: National Natural Science Foundation of China;
Funding text: ACKNOWLEDGMENT This research has been supported by the National Natural Science Foundation of China (No.61673211, No.U1633105, No.61203170), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No.18KJB590002), and Aeronautical Science Foundation of China (No.20155152041).
Compendex references: YES
Database: Compendex

Data Provider: Engineering Village