IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Infrared target tracking algorithm combining deep features and gradient features

To cite this article: Min Wu *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **490** 042009

Infrared target tracking algorithm combining deep features and gradient features

Min Wu[1], Yufei Zha[1,*], Bin Chen[1]

School of Aerospace Engineering, Air Force Engineering University, Xi'an, China

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

Abstract: Due to the low contrast between the target and the background of the infrared sequence image, the edge of the image is blurred and the dynamic range of the gray level is small, what features is used to describe the target becomes the key to tracking. Deep features and gradient features are the main features of most current tracking algorithms. However, the target semantic of deep feature extraction pays attention to intra-class classification, ignores intra-class differences, and is easily interfered by similar background (distractor); gradient features as a local area feature, it is not susceptible to background interference, but it cannot adapt to the dramatic deformation of the target. Based on the complementarity of these two features, this paper proposes an infrared target tracking algorithm that combines deep features and gradient features. In this paper, deep features and gradient features are used to represent the semantic and local structure of the target respectively, which enhances the ability to represent arbitrary targets. Next, the tracking model established by different features further improves the robustness of tracking. Finally, this paper establishes a model mutual aid mechanism, and uses the complementarity between the deep feature tracking model and the gradient feature tracking model to accurately target. In the experiment, this paper selects the latest infrared video tracking database (VOT-TIR2016) to verify the effectiveness of the proposed algorithm. The results show that compared with the current mainstream tracking algorithm, the algorithm achieves a 3.8% improvement in accuracy and a 4.3% improvement in success rate, it can effectively handle the effects of similar background and deformation in tracking.

1. Introduction

Infrared target tracking is a key technology in the military field such as infrared warning system, monitoring of low-altitude and ground targets by visual systems under no-load and infrared homing guidance [1]. Since the infrared sensor does not radiate energy into the air, it only detects and tracks the target by receiving the heat of the target radiation, so it is not easy to be reconnaissance or positioning, and has strong anti-interference ability; at the same time, since the target inevitably radiates heat, It also creates conditions for target detection and tracking using infrared sensors [2].

The contrast between the target and the background of the infrared image is low, the edge of the image is blurred and the dynamic range of the gray level is small [3], making infrared image tracking a challenging task [4]. The classical KCF [5] uses the gradient feature as the input feature, when the target is deformed, it is easy to cause the loss of the tracking target. Along with the promotion of deep learning, the HCF [6] based on deep features is widely used. The HCF algorithm extracts the convolutional layer features of the CNN network. The high-level features reflect the semantic information of the target, and the low-level features reflect the spatial characteristics of the target. CNN networks pay more attention to inter-class classification objects, ignoring intra-class differences, and thus have certain limitations in tracking specific infrared targets.

Aiming at the advantages and disadvantages of the two algorithms, this paper combines a large number of infrared tracking algorithms, and refers to the method of combining two features in staple [7], and proposes an infrared target tracking algorithm that combines deep features and gradient features. Figure 1 is a flow chart of the tracking algorithm. The algorithm extracts the feature from the deep feature and the gradient feature respectively. The deep feature and the gradient feature of the infrared target local region are used to minimize the regularization. The infrared multi-objective method is used to establish the infrared target tracking model. According to the complementary nature of the two features, the two features are used to determine the respective weights of the target's response values, and the mutual aid mechanism is used to combine the different prediction results to achieve the tracking of the infrared target. Compared with KCF and HCF, our algorithm improves the accuracy and robustness of tracking.

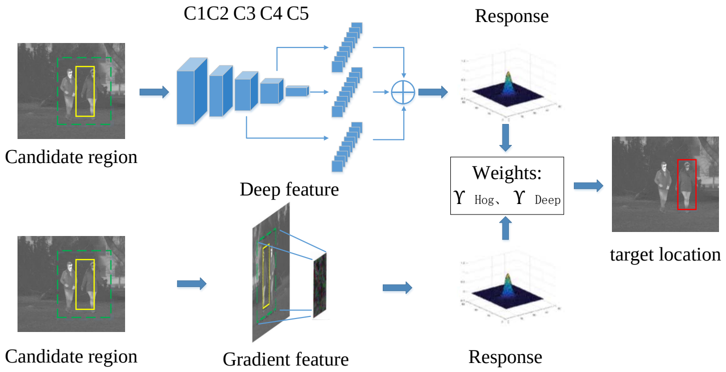

Figure 1 The Flow chart of tracking algorithm

2 Infrared target tracking algorithm combining deep features and gradient features

2.1 Model construction

The tracking algorithm in this paper adopts the strategy of detecting first and tracking, that is, firstly using the position of the target of the initial frame, training a target detector, and then using the target detector to detect whether there is a target in the predicted position of the next frame, and then use the new test results to update the training set and update the target detector. In the t-frame image when training the target detector, in the search box x_t, you may get multiple target frames for the same target S_t, select the target box with the highest response value as the position p_t of the target in the image：

$$p_t = \arg\max_{p\in s_t} f(\mathrm{T}(x_t, p); h, \rho) \qquad （1）$$

Which, $f(\mathrm{T}(x_t, p); h, \rho)$ is the function that calculates the response value of multiple target frames in the search box x_t corresponding to the real target frame p, according to the model parameters h and ρ ; $\mathrm{T}(x_t, p)$ is the function that extract gradient and deep features of the search box x_t and target box p ; $f(\mathrm{T}(x_t, p); h, \rho)$ can obtain the match between the predicted target and the detected target, and the higher the response value, the closer to the real target.

In this paper, by combining the deep feature and the gradient feature, according to the loss value l_{Deep} of the deep feature and the loss value l_{Hog} of the gradient feature, the loss function $L(h, \rho)$ is established by the regularized least squares method, the formula is expressed as:

$$L(h,\rho) = \min_{h,\rho} \gamma_{Hog} l_{Hog} + \gamma_{Deep} l_{Deep} + \lambda \|h\|^2 + \beta \|\rho\|^2 \qquad （2）$$

Which, γ_{Hog} is the weight of the gradient feature, γ_{Deep} is the weight of the deep feature; λ, β is a regularization parameter, Over-fitting is controlled by the adjusted value λ, β .

Extract the gradient features of the target frame, the target box is converted to its response value corresponding to the real target frame by filters h , and get the loss value of the gradient feature:

$$l_{Hog} = \sum_i (h^T \phi_i - y_i)^2 \qquad (3)$$

Where T is the transpose of the vector, ϕ_i is the gradient feature of the *i*-th target box, y_i is the label of the *i*-th target box.

The deep feature extraction is a feature of the three-layer convolutional layer in the CNN network, and the loss value of the deep feature is obtained from the coarse-grained to the fine-grained method:

$$l_{Deep} = \sum_i (\rho^T \sum_{l=3-i, i=0}^{i=1} \frac{1}{1+\mu_{l,l-1}} (\psi_i^{\ l} + \mu_{l,l-1} \psi_i^{\ l-1}) - \mathrm{y}_i)^2 \qquad (4)$$

Which, ρ^l is the related filter of the l -th layer, $\psi_i^{\ l}$ is layer 2 convolutional features of the i -th frame, $\mu_{l,l-1}$ is the constraint value of the feature of the l -th layer on the feature of the $l-1$ -th layer.

2.2 Gradient Feature Module

The image of the target frame is divided into several cell units, and each cell unit is combined into one block, and all the blocks are connected in parallel to obtain a gradient feature of the target frame.

$$\phi_{m+1} = par(\alpha_m, \alpha_{m+1}) \quad m = 1 \cdots n \qquad (5)$$

Which, α_m is the represents the gradient feature of the *m*-th block, n is the number of blocks in the target box, $par(\alpha_m, \alpha_{m+1})$ is the gradient feature of paralleling the first block and the second block, ϕ_{m+1} is the gradient feature of $m+1$ block in parallel. Each block represents local gradient information. The gradient feature of the target frame is composed of several local gradient features, reflecting the local area information of the target frame.

$L(h, \rho)$ is the partial derivative of x through the loss function, you can get the filter about the gradient feature h . Loop through the input gradient features, The characteristics after the loop are $\phi = F^H diag(\phi) \mathrm{F}$. According to the fact that the circulant matrix can be diagonalized by the discrete Fourier matrix, so that the matrix is inversely transformed into the property of the eigenvalue inversion, Ability to convert h to frequency domain for calculation, which greatly reduces the computational complexity, and increases the number of samples, improving the performance of the model. At this time, the model parameter function h is:

$$h = \frac{\gamma_{Hog} y \otimes \phi}{\gamma_{Hog} \phi \otimes \overline{\phi} + \lambda} \qquad (6)$$

Reference a non-linearly mapped column vector $\alpha(\phi)$ to make the mapped samples linearly separable in the new space. $\alpha(\phi)\alpha(\phi)^T$ is the covariance matrix similar to kernel space variables, Let K denote the kernel matrix of the kernel space. Since the Gaussian kernel uses the weighted mean of the pixel neighborhood to replace the pixel value of the point, and the weight of each neighborhood pixel is monotonically decreasing with the distance from the center, this applies to the image of the gradient feature. Constructing a ridge regression problem in two dimensional spaces by using gaussian kernel $k^{\phi_t \phi_{t-1}}$ in new space, you can get:

$$h^T \phi_t = \hat{\partial} * \hat{k}^{\phi_t \phi_{t-1}} \qquad (7)$$

Which,

$$\partial = \frac{\gamma_{Hog} y}{\phi \otimes \bar{\phi} + \lambda} \qquad (8)$$

$$k^{\phi_t \phi_{t-1}} = \exp(-\frac{1}{\sigma^2}(\|\phi_t\|^2 + \|\phi_{t-1}\|^2 - 2F^{-1}(\sum \hat{\phi}_t * \hat{\phi}_{t-1}))) \qquad (9)$$

ϕ_t is the gradient feature of the target frame of frame t, σ is the gaussian kernel variance.

2.3 deep feature module

Using CNN network to extract the characteristics of the target frame, according to the richer semantic information of the image in the lower layer of the neural network, in order to fully extract the semantic information of the target and retain the accurate spatial information, this paper extracts Conv3_4, Conv4_4 in CNN network. The feature of the three-layer convolution layer of Conv5_4 is used as the deep feature of the target frame.

Partially guided by ρ through the loss function $L(h,\rho)$, a function of the filter for the t-th frame of the deep feature can be obtained:

$$\rho(\mathrm{t}) = \frac{\gamma_{Deep} A(\mathrm{t})}{\gamma_{Deep} B(\mathrm{t}) \otimes \bar{B}(\mathrm{t}) + \beta} \qquad (10)$$

Which:

$$A(\mathrm{t}) = y \otimes \sum_{l=3-i,i=0}^{i=1} \frac{1}{1+\mu_{l,l-1}} (\psi_i^{\,l} + \mu_{l,l-1} \psi_i^{\,l-1}) \qquad (11)$$

$$B(\mathrm{t}) = \sum_{l=3-i,i=0}^{i=1} \frac{1}{1+\mu_{l,l-1}} (\psi^{l} + \mu_{l,l-1} \psi^{l-1}) \qquad (12)$$

The optimal filter for the t-th frame can be described by updating the minimum output error among all the tracking results, however this involves solving the problem of the linear equations. In order to obtain a robust approximation, this paper uses the moving average. The method updates the numerator and denominator of the filter to achieve the purpose of module update.

$$\hat{A}(\mathrm{t}) = (1-\eta_{Hog})\hat{A}(\mathrm{t}-1) + \eta_{Hog}\hat{A}(\mathrm{t})' \qquad (13)$$

$$\hat{B}(\mathrm{t}) = (1-\eta_{Hog})\hat{B}(\mathrm{t}-1) + \eta_{Hog}\hat{B}(\mathrm{t})' \qquad (14)$$

Which, η_{Hog} is the learning rate of each frame.

2.4 Module mutual help

Since the deep feature module and the gradient feature module have errors in predicting the target position, the larger the error, the more inaccurate the position prediction result for the target. In order to obtain a final model in combination with the trained modules, the weight of each module is designed in this paper, and the weight is based on the error of each module. The weight of the module is inversely proportional to the error. By using modular mutual aid to combine different estimates to reduce inaccurate predictions, the tracking accuracy of the final model can be improved to some extent. Since the response value of the module reflects the matching between the real target and the candidate target, the higher the similarity of the candidate frame, the more the semantic information and the local area information reflect the real target. Therefore, the response value of each module is used in this paper. To evaluate the module and use it to define the error:

$$\varepsilon = 1 - \max_{p \in s_t} f(\mathrm{T}(x_t, p); h, \rho) \qquad (15)$$

By using the difference between the maximum response value of each module and the difference of the real target tag to reflect the error of the module, the weight of each module can be obtained by combining the errors of the two modules:

$$\gamma_{Hog}=1-\frac{\varepsilon_{Hog}+\tau}{\varepsilon_{Hog}+\varepsilon_{Deep}+\tau} \tag{16}$$

$$\gamma_{Deep}=1-\frac{\varepsilon_{Deep}+\tau}{\varepsilon_{Hog}+\varepsilon_{Deep}+\tau} \tag{17}$$

Which, τ is a constant close to 0, its role is to avoid the denominator is 0 when ε_{Hog} and ε_{Deep} is 0 , and the weight of the module cannot be obtained.

3. Experimental results and analysis

3.1 Qualitative analysis

Figure 2 is a schematic diagram of the infrared tracking results. The target of the first video tracking is a drone, and the three different algorithms are used to test the tracking effect of non-living objects. The experimental results are shown in the figure. In the 61st to 91st frames of the image, the target passes through the window. Since the brightness of the window and the color of the target are similar, the edge information of the target is not obvious enough. At this time, KCF, KCFDP, HCF algorithm is difficult to track the target, and the algorithm can effectively track the target, which shows that the tracking accuracy of this algorithm is better than the other three algorithms when the edge of the target and background is not obvious.

The second video is about animal tracking, which uses three different algorithms to track non-human creatures. The experimental results are shown in the figure., we can see that the tracking accuracy of this algorithm is better than KCF and KCFDP when the object is deformed.

The third video is about human tracking, using three different algorithms to track the human body. The experimental results are shown in the figure, in the tracking of the entire video, because the target of the tracking is a specific People, and there are many other people in the video, this similar goal will. It has a great impact on the target tracking. It can be seen from the figure. The tracking accuracy of the algorithm is better than the other three algorithms.

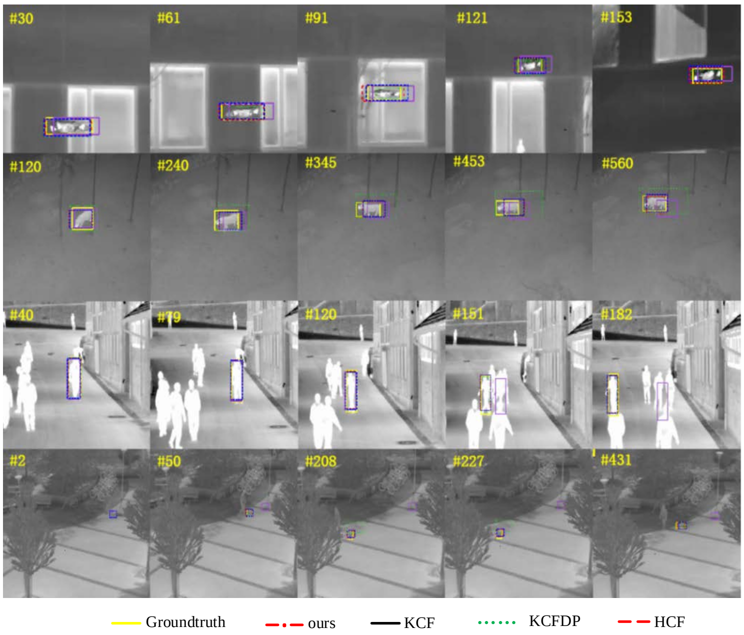

Figure 2 Schematic diagram of the tracking results

The fourth video is about small target tracking, using three different algorithms to track small targets. The experimental results are shown in the figure. In the tracking of the entire video, the tracked object is small, which causes great trouble for extracting detailed feature information. The algorithm combines the advantages of deep features and traditional features. It can not only extract the gradient information of the target, but also extract the semantic information of the target, which improves the accuracy and success rate of the tracking.

3.2 Quantitative analysis

In this paper, the method combines the deep feature with the HOG feature, and utilizes the advantages of the two features in different video tracking to complement each other and improve the tracking effect. In this experiment, in order to better demonstrate the superiority of the algorithm, this paper will target the three algorithms in a video library and compare the tracking results. The results are shown in Figure 3:

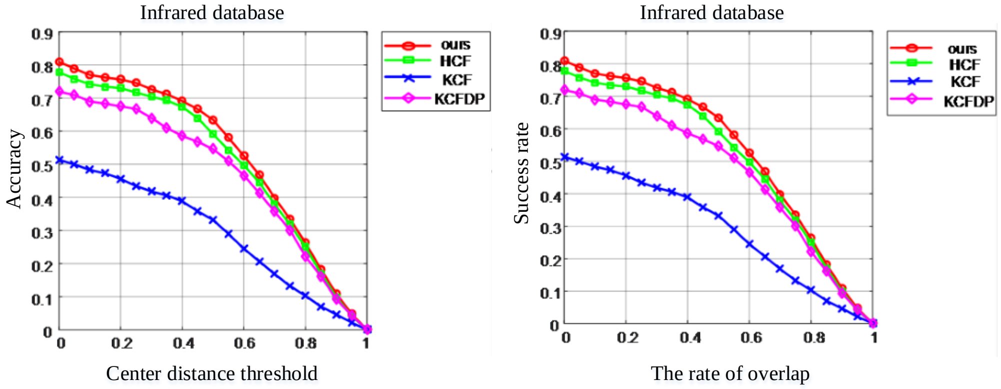

Figure 3 The comparison diagram of the accuracy and success rate

In this paper, the accuracy and success rate of tracking [9, 10] is used as the evaluation index of the tracking algorithm. The left graph in Figure is the accuracy of the tracking. The abscissa is the center distance threshold and the ordinate is the accuracy, that is, the prediction. The distance from the center of the target to the center of the real target is lower than the number of frames of the threshold in the entire video library; the right picture shows the success rate of the tracking, the abscissa is the threshold of coverage O_f . The ordinate is the success rate, that is, the proportion of the frame with the coverage greater than the threshold in the entire video library [18]. It can be seen from the figure that the algorithm has better tracking performance.

4 Conclusion

In the infrared target tracking, the deep feature ignores the intra-class difference and the gradient feature does not adapt to the target deformation. This paper proposes a method of fusing the deep feature and the gradient feature, using the deep feature and infrared representation of the infrared target semantic information. The gradient feature of the local area information of the target is established by the regularized least squares method [11], and the mutual prediction method is used to combine different prediction results to determine the position of the target. Through experiments, compared with HCF algorithm, KCF algorithm and KCFDP algorithm, the effectiveness and superiority of the proposed algorithm are fully proved.

Acknowledgments

This research has been supported by National Natural Science Foundation of China (No.61701524, No.61773397, No.61472442, No.61473309, No.61703423).

References

[1] Fan G, Venkataraman V, Tang L, et al.ON Boosted and Adaptive Particle Filters for Affie-Invariant Target Tracking in Infrared Imagery[J]. Augmented Vision Perception in Infrared,2009,3889(s 2-3):441-466.
[2] Lian Jie, Han Chuanjiu. Automatic Tracking Method of Infrared Target Based on Mean Shift[J]. Journal of Microcomputer, 2008, 24(4): 285-287.
[3] Yilmaz A, Shafique K, Shah M. Target tracking in airborneforward looking infrared imagery[J].

Image&Vision Computing, 2003,21(7):623-635.

[4] Danelijan M, Khan F S, Felsberg M, et al. Adaptive Color Attributes for Real-Time Visual Tracking[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014:1090-1097.

[5] Henriques J F，Rui C, Martins P, et al. High-Speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2014,37(3):583-596.

[6] Ma C, Huang J B, Yang X, et al. Hierarchical Convolutional Features for Visual Tracking[C]//IEEE International Conference on Computer Vision.IEEE,2015:3074-3082

[7] Danelljan M, Hager G, Khan F S, et al. Learning Spatially Regularized Correlation Filters for Visual Tracking[C]//IEEE International Conference on Computer Vision.IEEE,2015:4310-4318.

[8] Chang O, Constante P, Gordon A et al. A Novel Deep Neural Network that Uses Space-Time Features for Tracking and Recognizing a Moving Object [J]. Journal of Artificial Intelligence and Soft Computing Researh, 2017, 7.

[9] Y Wu, J. Lim, and M-H. Yang. Online object tracking: A benchmark[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2013:2411-2418.

[10] Wu Y, Yang M H. Object Tracking Benchmark[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015,37(9):1834-1848.

[11] Vedaldi A, Zisserman A. Efficient Additive Kernels via Explicit Feature Maps[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(3):480-492.