Multi-scale Correlation Filter Tracking Algorithm Based on Feature Fusion

Xiaokang Ren, Hongxiang Wang*, Yongye Wang, Xingzhen Li, Xingxing Liu
School of computer science and Engineering, Northwest Normal University, Lanzhou Gansu, 730070, China
1509952451@qq.com
renxk@nwnu.edu.cn

Abstract. For the Kernel Correlation Filter (KCF) algorithm using a single feature can not fully describe the tracking target, and is susceptible to the change of the target scale, lacking the estimation of the target scale size, an improved feature fusion and multi-scale correlation filter tracking algorithm based on KCF is proposed. Firstly, in the stage of location prediction, color features and directional gradient histogram hog features are fused by feature series, and feature dimensions are added to realize target location prediction. Then the scale filter is introduced into the predicted target position to estimate the optimal target scale as the tracking result. Experimental results show that the accuracy and success rate of the improved algorithm are significantly improved compared with other classical tracking algorithms, and it can deal with scale change, occlusion, deformation and other complex situations robustly.

1. Introduction
Target tracking is one of the basic problems in the field of computer vision[1], which is to predict the size and position of the target in the subsequent frames given the target size and position in the initial frame of a video sequence. It has been widely used in automatic driving, intelligent transportation, human-computer interaction, behavior analysis and other fields, but it still faces the challenges of complex environment such as illumination, occlusion, scale change, target rotation.

Researchers have proposed a large number of discriminant-based target tracking algorithms. In recent years, the algorithm based on correlation filter (CF) has made great progress in tracking accuracy and real-time, and has attracted extensive attention and research in the field of tracking. Bolme et al. first proposed the minimum output error and filter (MOSSE)[2] tracking algorithm, which greatly improved the speed of target tracking. However, due to the use of single-channel grayscale features, it is not suitable for complex scenes. Henriques et al. proposed Circulant Structure with Kernels(CSK) [3] tracking algorithm, which used single-channel grayscale features and added cyclic matrix and kernel function. On the basis of CSK, Henriques et al. proposed a high-speed kernel-related (KCF) [4] filter, using the directional gradient histogram HOG feature instead of the original grayscale feature, and using The Gaussian kernel function is optimized to improve the tracking accuracy. However, like MOSSE and CSK algorithms, the issue of scale changes is not considered. On this basis, Danelljan et al. proposed an adaptive scale (DSST) [5] tracking algorithm using HOG features. The algorithm designed two independent filters for separate training and position filtering. The device is used for target positioning, the scale filter is used for scale estimation, and the robustness of the algorithm is improved on the basis of the original algorithm MOSSE.
Aiming at the problems of scale change, occlusion and deformation in KCF tracking algorithm, a multi-scale correlation filtering tracking algorithm based on feature fusion is proposed. The algorithm fuses hog features and color features to achieve target position prediction. Then a 33 scale filter is introduced to adjust the target scaling according to the predicted position, and the adaptive maximum scaling ratio is adjusted as a result of the trace. Compared with the experimental results, the algorithm can better cope with complex environment and achieve robust tracking.

2. KCF tracking algorithm

The algorithm in this paper is improved based on the KCF tracker. KCF introduces a cyclic matrix, and all training samples are obtained by cyclically moving the target samples. The discrete Fourier transform is used to diagonalize the cyclic matrix to improve efficiency and reduce algorithm complexity. The introduction of kernel function can map a linear problem to a non-linear space, so that linearly inseparable samples become linearly separable, so that the algorithm is more general.

2.1. Classifier training

The training process of the classifier is a ridge regression problem. Let the training sample set be \( \mathbf{x}_i, y_i \) and the response function \( f(z) = w^T z \), find a linear regression function so that the error function is:

\[
\min_w \sum_i (f(x_i) - y_i)^2 + \lambda \|w\|^2
\]  
(1)

Obtain the minimum value, that is, to find a set of optimal weights \( w \).

Let the derivative be 0. Since the inversion of the matrix is involved, the complexity is too high, so it is written in the form of the complex number field:

\[
w = (X^H X + \lambda I)^{-1} X^H y
\]  
(2)

Reference [4] introduces the kernel function \( \varphi(x_i) \) to map the linear problem to the nonlinear space.

The filter \( w \) is represented by

\[
w = \sum_i \alpha_i \varphi(x_i)
\]

Convert the previous problem of seeking \( w \) into the problem of seeking \( \alpha \), the solution of ridge regression based on kernel function is obtained:

\[
\alpha = (K + \lambda I)^{-1} y
\]  
(3)

K denote the circulant matrix, Convert K into a diagonal matrix through Fourier transform, so the solution of \( \alpha \) into the Fourier domain:

\[
\hat{\alpha} = \frac{\hat{y}}{(K^* + \lambda)}
\]  
(4)

Here \( K^* \) represents the first row of the \( K \) matrix.

2.2. fast detection

The kernel matrix between the training sample \( \mathbf{x} \) and all candidate image blocks \( \mathbf{z} \) is

\[
K^z = C (k^x)
\]

Then the response of the test sample is

\[
\hat{f}(z) = \hat{k}^x \varpi \hat{\alpha}
\]  
(5)

The position corresponding to the maximum value \( \hat{f}(z) \) is the predicted position.
3. Improved KCF target tracking algorithm

3.1. Adaptive feature fusion

The traditional KCF tracking algorithm uses a single HOG gradient histogram feature to extract the features of the sample. When the target encounters slight deformation or light changes, the HOG feature has a good tracking effect; however, in complex scenes such as severe deformation and fast motion, the HOG feature will fail to track. As a global feature, the color CN feature focuses more on characterizing the color information of the sample, and is insensitive to target deformation and rapid motion. This paper is based on the invariance and globality of CN features when characterizing samples, and is not affected by changes in image rotation and translation. It is proposed to integrate HOG features and color features to increase the dimension of feature representation and improve the robustness of the algorithm.

In this paper, feature fusion is performed in the form of feature concatenation. The color feature vector of the image is \( c = [c_1, c_2, \ldots, c_m] \), and the HOG feature vector is \( h = [h_1, h_2, \ldots, h_n] \). The fused feature vector is \( x = [x_1, x_2, \ldots, x_n] \) represents different features,

\[
K(x) = \exp \left( -\frac{1}{\sigma^2} \left( x' x + x' R x - 2 F^{-1} \left( \sum_{c} x_c' x_c \right) \right) \right)
\]  

(6)

\( x_c \) represents a mixture of two features, that is, a simple vector superposition of the two features.

In summary, multi-channel feature fusion expands a single feature into multiple features, and extracts the HOG feature and the color CN feature of the directional gradient histogram, respectively. The HOG feature is composed of the directional histogram of the local grid cells of the image. In this paper, the 31-dimensional feature channel. As a color feature, the CN feature divides the color into: black, blue, brown, gray, green, orange, pink, purple, red, white, yellow 11 colors, that is, 11-dimensional color space, better The target is extracted and recognized, and it performs better when the color video sequence with color features encounters severe deformation and fast movement.

3.2. Multiscale filter

The KCF tracking algorithm does not consider the scale change of the target. In the actual video tracking process, the scale of the tracking target often changes. When the target scale becomes smaller, the filter will learn a large number of negative samples; when the target scale becomes larger, the filter tracks some local information, which is not adaptable to the scale change, resulting in tracking offset or tracking failure. Reference [5] is based on the MOOSE filter, which introduces a scale filter separately to quickly and accurately respond to changes in the target scale. In this paper, the target is scaled on the KCF tracking algorithm according to this method.

This article divides the tracking process into two independent processes: position prediction and scale prediction. In KCF, the position filter is used to obtain the target position according to equation(5) and then the current target position is used as the center point to estimate with the newly added scale filter. The scale of the target, the scale is constructed as

\[
P \times R, \quad n \in \left\{ \left\lfloor \frac{S-1}{2} \right\rfloor, \ldots, \left\lfloor \frac{S-1}{2} \right\rfloor \right\},
\]

The scale of the target, the scale is constructed as \( P \times R \) represents the target size of the frame, \( a \) is the scale factor, and \( S \) is the number of scales.

For the input sample \( f \) has a d-dimensional feature map representation, the optimal correlation filter \( h \) is constructed by establishing a minimum cost function, the frequency domain filter \( h \) is:
In order to reduce the amount of calculation, the above formula is split into numerator and denominator forms and iteratively updated:

\[ A_i^t = (1 - \eta) A_{i-1}^t + \eta \bar{G} F_i^t \]  
\[ B_i = (1 - \eta) B_{i-1} + \eta \sum_{k=1}^{d} F_i^k + \lambda \]  

\( \eta \) is the learning rate, and \( t \) and \( t-1 \) represent the current frame and the previous frame. In the new frame \( z \), the response value of the scale filter is obtained, and the scale corresponding to the maximum value is the target scale of the new frame

\[ y = \mathbf{r}^{-1} \left\{ \frac{\sum_{i=1}^{d} \bar{A} Z_i^t}{B + \lambda} \right\} \]  

4. Experimental results

4.1. Experimental environment and parameter settings

All experiments in this article were completed on a PC with Matlab R2016b, Intel(R) Core(TM) i5-4210U CPU, clocked at 2.4GHz, and 4GB memory. In the experiment, the parameters of the traditional KCF algorithm remain unchanged, the regularization parameter \( \lambda = 0.0001 \), the padding window is 2.5 times the target area, and the Gaussian kernel bandwidth \( \sigma = 0.5 \). The feature of combining HOG feature and CN color feature is adopted, in which the number of HOG feature directions is 9, and the size of HOG feature is 4×4. Set the scale filter series \( S \) to 33 scales, the scale increment factor \( \alpha = 1.02 \), and the learning rate set to \( \eta = 0.025 \).

4.2. Performance evaluation method

In order to conduct a comprehensive evaluation of the proposed algorithm, the experiment used two evaluation methods[6]: precision plot and success plot. The center position error (CLE) is used to evaluate the tracking accuracy. CLE represents the average Euclidean distance between the center position of the tracking result and the calibrated center position. The smaller the value, the more accurate the tracking result. When the CLE value is less than a certain threshold (20 pixels set in this article), the tracking is considered successful. In the success rate, given the current tracking frame and the calibrated target frame, the overlap rate

\[ S = \frac{\text{area} \left( B_r \cap B_c \right)}{\text{area} \left( B_r \cup B_c \right)} \], where \( B_r \) is the current tracking frame and \( B_c \) is the calibrated target frame, overlapping. The larger the rate \( S \) value, the more accurate the result. When the \( S \) value is greater than the given threshold (generally 0.5), the tracking is considered successful.

4.3. Quantitative analysis

In this experiment, a one-pass evaluation (OPE) method based on the accuracy map and the success rate map was used to test the accuracy and success rate of 50 video sequences (including 11 attributes) in the OTB-2013 data set, and to compare with other such as KCF, DSST, CSK, Struck, TLD and other classic tracking algorithms are compared.

Figure 1 is an OPE accuracy graph and success rate graph for the overall performance of the six tracking algorithms. As shown in the figure, the accuracy of the algorithm in this paper is 80.2%, which is 5.6% higher than the DSST algorithm with the same multi-scale filter, and 5.2% higher than the classical KCF correlation filter using hog features. In terms of success rate, it is second only to DSST algorithm (0.7%) and better than KCF algorithm (3.8%). The results show that the improved
algorithm can effectively improve the tracking accuracy and ensure the real-time and robustness of the algorithm by fusing multiple features and adding adaptive multi-scale filter.

Figure 2 shows the accuracy graph and success rate graph of some attributes of the six algorithms. The improved algorithm ranks high on the accuracy graph and success rate graph. Compared with the traditional KCF algorithm, the improved algorithm has improved in all aspects. Figure 2a and 2d, when dealing with scale transformation, improved the algorithm by 7.7% and 5.9% respectively compared with the KCF algorithm, and the effect is significant after adding the scale filter. In terms of success rate, it is slightly lower than the DSST algorithm that also has scale adaptation. When dealing with deformation, as shown in Figure 2b and 2e, the proposed improved algorithm improves the accuracy and success rate by 6.5% and 3.9%, respectively, indicating that the CN and HOG features have strong robustness after fusion. Figure 2c and 2f shows that when facing occlusion, it has also improved. So that combines multiple features and adaptive scale can better adapt to complex environments such as deformation, scale change, occlusion, and background mixing, which is better than other Tracking algorithms.
4.4. Qualitative analysis
In order to facilitate a more intuitive comparison algorithm effect, this article compares the DSST, KCF and CSK algorithms test the occlusion and scale changes respectively.

4.4.1 Scale changes
In the Singer1 data set shown in Figure 3a, the singer has a significant scale change when switching the lead shot. For example, at frames 56, 129, and 350, the lens distance changes from near to far, especially at frame 129. It becomes smaller and suffers from serious light problems. The improved algorithm accurately describes the change in scale, and responds well to changes in lighting. The tracking effect of the KCF and CSK algorithm has not reached expectations. In the carScale data set shown in Figure 3b, the vehicle is from far to near, and the scale is from small to large. At 169 frames, the branches block the vehicle, and the CSK algorithm that uses only gray features loses the target. At 189 frames, when the vehicle appears again in the field of view and the scale becomes larger, the improved algorithm accurately reflects the scale change of the target. This is because the scale filter is introduced in this paper to better cope with the change in the size of the target.

![Figure 3](image1)

4.4.2 Occlusion
In the football data set shown in Figure 4c, similar target occlusion occurs. In frames 301 and 361, due to the extremely similar characteristics of the target and the background, other algorithms incorrectly tracked the No. 37 and No. 38 athletes. Ours algorithm accurately tracked the No. 24th, there was no departure from the goal. In the freeman data set shown in Figure 4d, the target appears continuous rotation, occlusion, attitude change and other disturbances, which makes other algorithms track errors, and the improved algorithm achieves the target tracking robustly. It can be seen that compared with other algorithms, the improved algorithm has better robustness in the case of occlusion. Due to the improved algorithm fusing color features and HOG features, which can better deal with occlusion problems.
5. Conclusion

This paper proposes a feature fusion multi-scale correlation filtering tracking algorithm based on KCF tracking algorithm. Combining CN features and HOG features, increasing feature dimensions, reducing interference in complex scenes such as severe occlusion and deformation, and improving the accuracy of target position prediction. In the scale prediction stage, scale filters are introduced to establish scale pyramids to adjust the scale of the target and enhance the adaptability of the scale. The experiment compares 6 mainstream algorithms. Through quantitative and qualitative analysis, it is shown that the improved KCF tracking algorithm can effectively cope with complex environments and achieve accurate tracking of targets.

References

[1] Smeulders A W M, Chu D M, Cucchiara R, et al. Visual tracking: An experimental survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(7): 1442-1468.

[2] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 2544-2550.

[3] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2012: 702-715.

[4] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 37(3): 583-596.

[5] Danelljan M, Häger G, Khan F, et al. Accurate scale estimation for robust visual tracking[C]//British Machine Vision Coference, Nottingham, September 1-5, 2014. BMVA Press, 2014.

[6] Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2411-2418.