Adaptive context-aware correlation filter target tracking

Saijun Zhou, Chengwang Zhang, Xuying Xiong, Ran He, Jingang Qiu
Shijiazhuang Campus, Army Engineering University, Shijiazhuang, Hebei, 050001, China
saijun_zhou@foxmail.com

Abstract: Aiming at the problem that the traditional correlation filter target tracking algorithm has low tracking accuracy under the conditions of fast motion, occlusion and complex background, an adaptive context-aware correlation filter target tracking algorithm is proposed in this paper. On the basis of the relevant filtering algorithm, the boundary effect and fixed learning rate brought by cyclic displacement are improved as the main purpose. Firstly, an adaptive sampling strategy based on the extreme value of the response graph is added to the context information in the training stage of the classifier. Then, A piecewise learning rate adjustment strategy is utilized to make the algorithm better adapt to the target change. Finally, the performance of the algorithm is verified by the standard data set. The experimental results show that the proposed algorithm improves the tracking accuracy of DCF and SAMF algorithm respectively. It not only has good robustness in the case of fast motion, occlusion, complex background, etc., but also can be integrated into most relevant filtering algorithms as a framework.

1. Introduction
Serving as an important research direction in the field of computer vision, target tracking technology possesses broad application prospects in many fields. In recent years, despite remarkable development achieved, target tracking technology still faces various challenges like scale changes, illumination changes, occlusion, target deformation and rapid motion [1]. The method based on correlation filtering has become one of the current research hotspots due to its advantages of high speed, high precision and strong robustness. Literature [2] proposed a kernel correlation filter tracker (KCF), improving its real-time performance by introducing cyclic shift and fast Fourier transform (FFT). In [3], a kernel-correlation filter tracker (SAMF) with scale adaptation and feature fusion is proposed, which combines hog features with colour name features (CN) and uses scale-based methods to estimate target scales. The traditional kernel-related filtering algorithm usually utilizes a cosine window to reduce the boundary effect, however, it reduces the negative sample information simultaneously, resulting in a lower discriminative ability of the classifier. To solve the problem, literature [4] proposed a spatial regularization correlation filter (SRDCF). By adding spatial regularization, the spatial weight of the sample in the target position region is improved, while the weight of the boundary region is punished. Although accuracy of the algorithm is improved, as no closed solution exists, iterative optimization using Gauss-Seidel method inevitably results in worse real-time performance. Literature [5] proposed a finite boundary correlation filter (CFLB), which introduces more real samples by increasing the detection area to reduce the filter size. With no closed solution as well, it is iterated by the alternating direction multiplier algorithm. It is reported that the tracking speed is improved, while the accuracy sacrificed to some extent. On the basis of literature [5], literature [6] proposed a background-aware correlation filter tracker (BACF), improving the sample quality based
on the cyclic shift, in this way the speed and accuracy of the algorithm are significantly raised. In [7], a context-aware correlation filter tracker (CACF) is proposed, which improves the accuracy of the correlation filtering algorithm with a small speed loss. Whereas the context information is sampled at a fixed position, and the background information cannot be well mined. As a result, the accuracy will be severely influenced when the background is complicated and the target motion is blurred. To this end, this paper modifies the sampling strategy and learning rate adjustment method, and proposes an adaptive context-aware correlation filter which outperforms in tracking accuracy compared with the traditional relevant filtering algorithm.

2. adaptive context-aware correlation filter

The tracking process of the traditional nuclear correlation filter can be described as: the initial state of the target is determined by manual frame selection in the initial frame, and then the regularized least squares classifier is trained by the padding at the target position, with the training utilized in the next frame. The completed classifier detects the target location and updates the classifier. The performance improvement of the tracking algorithm mainly results from the cyclic shift of the training samples to approximate the equivalent of dense sampling and reduce the computational overhead by FFT transformation. However, the precondition for the cyclic shift equivalent to the actual displacement requires uniform background and large-scale target displacement. Usually used to limit the boundary effect of cyclic displacement, cosine window also reduces the context information around the target, leading to worse discriminating ability of the classifier. When the target is in complex conditions including fast motion, motion blur, background interference, etc., drift or tracking failure will occur. Aiming at the above problems, the proposed algorithm introduces the adaptively selected context information into the training phase of the traditional correlation filtering algorithm, and derives the closed solution.

2.1 context-aware correlation filter

In the current frame, the traditional correlation filter class tracker uses the ridge regression training classifier, and the ridge regression of the objective function is:

$$\min_w \|f - y\|^2_2 + \lambda \|w\|^2_2$$

where, $f$ is the classification function which can be represented by a linear combination of base samples: $f = X_0 w$, $\lambda$ is regularization parameters preventing over-fitting, $w$ is the relevant filter coefficient. $X_0$ is the cyclic shift matrix for the base sample.

On the basis of equation (1), with the correlation filter response graph of the previous frame, $n$ background samples is adaptively obtained surrounding the target sample $x_0$ (containing context information), represented by $x_i$, whose corresponding cyclic shift matrix is $X_i$. The $n$ background samples are treated as the negative samples to train the classifier, and the classifier is restricted with a higher response value at the target sample and a response value close to zero at the background sample. So the ridge regression of the objective function after adding context information is displayed as:

$$\min_w \|X_0 w - y\|^2_2 + \lambda \|w\|^2_2 + \gamma \sum_{i=1}^n \|X_i w\|^2_2$$

where $\gamma$ is the background sample returning to zero for regularization parameters.

Combining the background sample in equation (2) with the cyclic matrix of the base sample, the following equation can be obtained

$$\min_w \|Aw - \hat{y}\|^2_2 + \lambda \|w\|^2_2$$

The ridge regression represented by the formula (3) is consistent with the proof of the standard ridge, and it also has a closed solution:
\[ w = (A^T A + \lambda I)^{-1} A^T \tilde{y} \] (4)

where \( A \) is a block circular matrix, which is set in Fu-space

\[
A = \begin{bmatrix}
X_0 \\
\sqrt{\gamma} X_1 \\
\vdots \\
\sqrt{\gamma} X_n
\end{bmatrix}
\] (5)

\[
\tilde{y} = \begin{bmatrix}
y \\
0 \\
\vdots \\
0
\end{bmatrix}
\] (6)

The discrete Fourier transform (DFT) matrix can be used for diagonalization. \( A \) is shown in formula (8), among them:

\[
\hat{x}_n = F(x_n) = \sqrt{n} Fx_n
\] (7)

\[
A = F^H \begin{bmatrix}
\text{diag}(\hat{x}_0) \\
\text{diag}(\hat{x}_1) \\
\vdots \\
\text{diag}(\hat{x}_n)
\end{bmatrix} F
\] (8)

where \( F \) is the DFT matrix, \( \hat{x}_n \) is the value after the Fourier transform.

From the above derivation, we can conclude that the closed solution in the Fourier Domain is:

\[
\hat{w} = \frac{\hat{x}_0 \odot \hat{y}}{\hat{x}_0 \odot \hat{x}_0 + \lambda + \gamma \sum_{i=1}^n \hat{x}_i \odot \hat{x}_i}
\] (9)

In the case of nonlinear ridge regression, the discriminative ability of the classifier is enhanced by using nuclear techniques:

\[
f = w^T x_0 = \sum_{i=1}^n a_i K(x_0, x_i)
\] (10)

Therefore, the closed solution is shown as:

\[
\alpha = (AA^T + \lambda I)^{-1} \hat{y}
\] (11)

Utilizing the properties of the periodic matrix, the closed solution in the Fourier domain can be obtained:

\[
\hat{\alpha} = \begin{bmatrix}
\text{diag}(a_{00}) & \cdots & \text{diag}(a_{0n}) \\
\vdots & \ddots & \vdots \\
\text{diag}(a_{n0}) & \cdots & \text{diag}(a_{nn})
\end{bmatrix}^{-1} \begin{bmatrix}
\hat{y} \\
\vdots \\
0
\end{bmatrix}
\] (12)

where vector \( a_{ij} \) \( (i, j \in \{1, 2, \ldots, n\}) \) can be exhibited as follows:

\[
a_{00} = \hat{x}_0 \odot \hat{x}_0 + \lambda
\] (13)

\[
a_{0i} = \gamma (\hat{x}_i \odot \hat{x}_i) + \lambda, i \neq 0
\] (14)

\[
a_{ij} = \sqrt{\gamma} (\hat{x}_i \odot \hat{x}_j), i \neq j
\] (15)
In the next frame, the trained classifier is used to detect the target position, and the maximum response value of the target sample \( x_0 \) locating at the same position can be calculated as follows:

\[
\max f = \max F^{-1}\left(\hat{x}_0 \odot \hat{x}_0 \odot \hat{\alpha}_0 + \sqrt{\sum_{j=1}^{N} \hat{x}_j \odot \hat{x}_j \odot \hat{\alpha}_j}\right)
\]  

(16)

The position where the maximum response exists is regarded as the predicted position of the target in the current frame.

2.2 Adaptive sampling strategy based on response graph extremum

The selection of context information (background samples) can exert significant influence on tracking performance [7]. Even though it is versatile uniformly sampling of background information surrounding the target sample, this method may reduce the tracking accuracy or cause tracking failure if the target has partial occlusion or background interference. Therefore, to solve the problem, this paper proposes a sampling strategy based on the extreme value of the response graph.

If the correlation filter response graph of the current frame is \( R \), whose peak value is \( P = \max R \). In the response graph, the local maximum of the two-dimensional response graph in a 5x5 region can be computed:

\[
\text{localMaximum} = \text{localMaximum} R
\]  

(17)

where \( \text{localMaximum} \) represents a function used to calculate a local maximum. The position of Local maximum \( p_i \) is \( l_i \), and the position of peak value \( P \) is \( l_p \). During the sampling process, the size of the target sample is denoted by \( \text{size}_0 \). The sampled background sample is the same size as the target sample. If the position of the response graph peak value \( P \) in the current frame is \( L_p \), the position distance threshold \( L_{\text{threshold}} \) and the response threshold \( R_{\text{threshold}} \) can be set respectively, and the sampling result can be displayed as:

\[
L_i = \begin{cases} 
  l_i, & \text{if } l_i > L_{\text{threshold}} \text{ and } p_i > R_{\text{threshold}} \\
  l_p \pm \text{size}_0, & \text{others}
\end{cases}
\]  

(18)

where \( L_i \) denotes the actual location of the context information sample, \( \text{size}_0 \) denotes the offset in the horizontal and vertical directions. In the training process of the classifier, according to the filtered response map of the previous frame, the actual position samples can be calculated by equation (18) and then the background sample can be substituted into the equation (2) to compute the ridge regression. The sampling method is shown in Fig. 1.

![Fig.1 Adaptive sampling strategy](image)

During the tracking process, when large interference exists in the background information, several pseudo peaks may appear in the response graph. Utilizing the adaptive sampling strategy based on the extreme values of the response graph to sample the regions which satisfies the criteria, the context information containing large interference can be suppressed, which can guarantee robustness to areas with strong interference.
3. Segmentation adjustment algorithm of learning rate

Since the target and background are constantly changing during the tracking process, the traditional correlation filtering algorithm uses the fixed learning rate for linear interpolation to update the classifier model and the target appearance model. The method can adapt to the target appearance model or background changes to some extent. However, when the target has occlusion, in-plane rotation, motion blur, etc., the target appearance model may vary significantly, which will lead to errors in the updated model and result in tracking drift or tracking failure eventually.

The learning rate reflects the learning ability of an algorithm responding to the target appearance model change. When the learning rate becomes larger, better tracking effect of the target appearance model will be obtained. And if the learning rate gets smaller, the algorithm will better suit for the case that little change exists in the target appearance model with large change in background [8]. In response to the above problems, the proposed algorithm uses a segmentation learning rate adjustment method. The response graph extreme calculation results in the previous section is also used to adjust the learning rate in stages.

The first step is to calculate all extreme points \( p_i (i=1,2,\ldots,k) \) in the specified neighborhood except the peak \( P_t \). Then, the following judgement is conducted according to formula (19). If the response extreme point is greater than a half of the peak size, it is regarded that the target is severely occluded, exhibits in-plane rotation or deformation, which will result in poor tracking quality. Other weighting factors are selected based on experience.

The second step is to adjust the learning rate \( \eta \) and update the model according to the calculation result of the previous step as follows.

\[
\text{for } i = 1 : k \\
\quad \text{if } p_i < P \& p_i \geq 0.5\ast P \\
\quad \quad s_1 = s_1 + 1 \\
\quad \quad \text{else if } p_i < 0.5\ast P \& p_i \geq 0.2\ast P \\
\quad \quad \quad s_2 = s_2 + 1 \\
\quad \quad \quad \text{else} \\
\quad \quad \quad \quad s_3 = s_3 + 1 \\
\quad \quad \quad \text{if end} \\
\quad s = s_1 \ast 0.8 + s_2 \ast 0.1 + s_3 \ast 0.001 \\
\text{end} \\
\]

(19)

When \( s \geq 1.6 \), it indicates that the target appearance model changes greatly and it is better to use a higher learning rate. When \( 0.5 \leq s < 1.6 \), it means that the target appearance model does not change much and the background has certain changes, so the moderate learning rate is adopted. When \( s < 0.5 \), it means that the changes are relatively small, and the learning rate is small.

\[
\eta = \begin{cases} 
0.005 & s < 0.5 \\
0.015 & 0.5 \leq s < 1.6 \\
0.1 & s \geq 1.6 
\end{cases} \\
\]

(20)

The above learning rate segmentation adjustment algorithm is used to update the model:

\[
\hat{\alpha}^t = (1 - \eta)\hat{\alpha}^{t-1} + \eta \hat{\alpha} \\
\hat{x}^t = (1 - \eta)\hat{x}^{t-1} + \eta \hat{x} \\
\]

(21)  (22)

where \( t \) is the number of frames, \( \hat{\alpha} \) is the classifier model, \( \hat{x} \) is the target appearance model.
4. Experimental results and analysis

The experimental is conducted in Matlab2015b, and the configuration of the computer is displayed as follows: Inter Core i5-3230M 2.60GHz processor, 4GB memory. The experimental parameters are set as follows: training detection area padding = 2.0; regularization parameters \( \lambda = 10^{-4} \), \( \gamma = 25 \); position distance threshold \( L_{\text{threshold}} = \sqrt{\text{size}.x^2 + \text{size}.y^2} \); response threshold \( R_{\text{threshold}} = 0.3*P \); the number of background samples \( n = 4 \); HOG characteristic parameter is consistent with Literature [2].

In order to evaluate the performance of the tracking algorithm, the proposed algorithm is tested on the standard data set OTB-50[9]. The evaluation index adopted refers to Literature [9]: Center Location Error (CLE), Distance Accuracy (DP), overlap precision (OP). The evaluation method uses OPE (One-Pass Evaluation). CLE is the Euclidean distance between the tracked target center and the marked target center. DP is the percentage of CLE less than the preset threshold. The threshold is set to 20 pixels. OP is the percentage of the number of frames whose overlap rate is greater than the preset threshold, and the threshold is set to 0.5. If the obtained value is greater than the preset one, the tracking is considered successful. The overlap rate VOC is calculated as:

\[
VOC = \frac{\text{area}(B_r \cap B_G)}{\text{area}(B_r \cup B_G)}
\]

where \( B_r \) represents the actual tracking box, \( B_G \) represents the true tag tracking box.

Since the proposed algorithm can be regarded as the modification of the related filtering algorithm, it can be directly applied to all related filtering algorithms. Therefore, DCF and SAMF, as the representative of the relevant filtering algorithm, are selected to conduct two sets of comparison experiments. All algorithms adopted linear kernel function. The first group is compared by quantitative analysis on 50 sets of video sequences, and the second group selected video sequences with occlusion, motion blur, fast motion and in-plane rotation characteristics for qualitative analysis and comparison.

4.1 Quantitative analysis

Fig. 2 compares the DCF algorithm and the SAMF algorithm after adding adaptive context awareness (ACA).

It can be seen that the average DP and the average OP of the DCF algorithm after adding ACA increase 6.9% and 11.2% respectively. The SAMF algorithm increased 2.2% and 3.1% respectively. The quantitative analysis results show that the proposed algorithm has better tracking performance.
Fig. 3 selects part of the video possessing attributions like in-plane rotation, fast motion, and motion blur to conduct algorithm comparison. The DCF algorithm is used to compare with the context perception (CA) in Literature [7] and the ACA algorithm proposed in this paper. The results indicate the proposed algorithm has better tracking performance when the above attributes in the target.

4.2 Qualitative analysis

Qualitative analysis selects two typical sequences, namely, Soccer and Human7. The Soccer sequence mainly has infra-plane rotation, fast motion, partial occlusion and other interference. And the Human7 sequence mainly has fast motion and motion blur interference. Fig. 4 shows the tracking results on two sets of sequences respectively from DCF and DCF (CA) and DCF (ACA).

In the Human7 sequence, in the first 100 frames, and the target motion is slow. It can be seen that all three algorithms can accurately track the target. From 136 frames to 224 frames, the DCF algorithm quickly loses the target due to the rapid motion of the target caused by the sharp camera shake. On the contrary, DCF adding CA and ACA always track targets accurately.

In the Soccer sequence, before the first 40 frames, the target has certain motion blur, and all three algorithms can accurately track the target. Near the 77th frame, the target has certain in-plane rotation, the DCF algorithm produces drift, and the DCF (CA) algorithm loses the target. At the 125th frame, the target has partial occlusion with both in-plane rotation and severe motion blur, and DCF and DCF(CA) lose the target. However, DCF (ACA) always tracks the target accurately. Qualitative analysis of the sequence shows that the proposed algorithm possesses some unique advantages compared to the original DCF and the context-aware DCF when in-plane rotation and motion blur due to adaptive sampling to the target.

The context information that generates the interference and the segmentation adjustment learning rate make the classifier have stronger discriminating ability and can better adapt to appearance changes of the target, thus effectively improving the robustness to the above situations.
According to the above qualitative analysis, the proposed algorithm can effectively solve the problems of fast motion, in-plane rotation and partial occlusion, and has better tracking performance compared with traditional algorithms.

5. Conclusion
Based on the related filtering algorithm, an adaptive context-aware algorithm is proposed in this paper, aiming to solve the problem of cyclic shift by adaptively sampling the context information according to the extreme value of the response graph in the training phase of the classifier. With the segmentation learning rate adjusted, the algorithm can better adapt to changes in the appearance of the target. The experimental results indicate that the proposed algorithm is more robust in the case of fast motion, in-plane rotation and occlusion. More importantly, the proposed algorithm can be integrated into most relevant filtering algorithms as a framework, helping to improve their tracking performance.

References
[1] Zhang Wei, Kang Baosheng. Overview of Correlation Filter Target Tracking[J]. Journal of Image and Graphics, 2017, 22(8): 1017-1033.
[2] Henrique J F, Caseiro R, Martines P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transaction on Pattern Analysis and Machine Intelligent, 2015, 37(3): 583-596.
[3] Li Y, Zhu J K. A scale adaptive kernel correlation filter tracker with feature integration[C]//Proceedings of 2014 European Conference on Computer Vision.
[4] Martin Danelljan, Gustav Häger, Fahad Khan, Michael Feldberg. "Learning Spatially Regularized Correlation Filters for Visual Tracking"[C]. IEEE International Conference on Computer Vision (ICCV), IEEE, 2015: 4310-4318.
[5] H. Kiana Galoogahi, T. Sim, and S. Lucey. Correlation filters with limited boundaries. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 4630–4638.
[6] Hamed Kiani Galoogahi, Ashton Fag, Simon Lucey. Learning Background-Aware Correlation Filters for Visual Tracking. IEEE International Conference on Computer Vision (ICCV).
[7] Matthias Mueller, Neil Smith, Bernard Ghanem. Context-Aware Correlation Filter Tracking. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[8] Xiong Changzhen, Zhao Wei, Guo Fenhong. A Kernel Correlation Filtering Tracking Algorithm Based on Adaptive Feature Fusion[J]. Journal of Computer-Aided Design & Computer Graphics, 2017, 29(6): 1068-1074.
[9] Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
[10] Wang Yanchuan, Huang Hai, Li Shaomei, Gao Chao. Correlation Filter Target Tracking Based on Online Detection and Scale Adaptive[J]. Acta Optica Sinica, 2018, 38(02): 0215002.
[11] Hu Xiaowei, Chen Juan. Target tracking of kernel correlation filters with color features[j]. Electro-optical and Control, 2017, 24(06): 43-46.

[12] Chen Qianru, Liu Risheng, Fan Xin, Li Haojie. Robust target tracking with multi-correlation filtering adaptive fusion[J]. Journal of Image and Graphics, 2018, 23(02): 0269-0276.