Background Spatial Correlation Filter in Multi-Channel Object Tracking

Jiayi Lyu¹, Ning He²*, Xin Sun² and Jian Xue³

¹ Capital Normal University, Beijing, 100048, China
² Beijing Union University, Beijing, 100101, China
³ University of Chinese Academy of Sciences, Beijing, 100049, China
* email: xxthening@buu.edu.cn

Abstract. Traditional correlation filters have high tracking efficiency, but because of the negative samples in the correlation filter that are obtained by moving the object patch itself, the real background information is not included in the sample composition. In this paper, the method for obtaining and detecting the background spatial information in multiple channel is given. A method that combines the background spatial information is integrated into the framework of correlation filter, and an experiment was carried out on existing publicly available datasets. The experimental results show that the tracking effect in some complicated scenes (e.g., illumination changes, fast motion, scale changes, and occlusion) is improved. This approach can effectively improve the success rate and accuracy of object tracking in complicated scenes without affecting the frame rate, obtaining a performance that is superior to that of other correlation filters.

1. Introduction

Object tracking is a key research topic in the field of computer vision. Its applications include real-time monitoring and human-machine interaction. To solve the problem of object tracking, a specific object is usually given in the first frame of the video sequence, so it is necessary to achieve real-time and accurate tracking in the next frame of that sequence. This requires an algorithm to predict the location of the object in the next frame using the object and background information. However, because of the rapid movement of the object and changes in the background, tracking often drifts, making it a challenging problem.

In recent years, the application of correlation filtering based on the discriminative algorithm has achieved great improvements. The high-speed tracking of the correlation filter is the main advantage of this approach. The first algorithm to use a correlation filter was MOSSE[1]. The MOSSE filter tracks an object online using the least-squares-sum error, and its adaptive correlation filter mode outputs a response so that the object position can be determined. Using the MOSSE filter framework as a basis, many extended algorithms have appeared. The CSK algorithm[2] uses cyclic structure to detect the correlation of adjacent frames in the frequency domain, which reduces the amount of computation and greatly improves the detection speed. The CN algorithm[3] uses an 11-dimensional colour space to expand CSK colour and uses PCA technology to reduce the dimensions to two, which reduces the runtime of the operation and improves performance. Henriques et al. employed a Gaussian kernel in MOSSE and proposed the KCF tracking algorithm[4] to solve the problem of scale changes. This algorithm uses HOG features to define the connections among multi-channel features, extending...
To solve this limitation, this paper proposes an improvement to the correlation filter by taking the background information around the object into account and integrates it into the learning filter. Finally, the frame output can be integrated with most correlation filter trackers to improve their performance while maintaining a high frame rate.

2. Our method

2.1. Correlation filter tracker
Correlation filters usually use a discriminant form to learn update filters that can infer the region of interest (i.e., the object location) in subsequent frames. It enables dense sampling around the object, and all possible translation positions of the object are modelled as cyclic shifts over a fixed-size search window, forming a circulant matrix $A_0$. Using the property of circular matrix diagonalization in the Fourier domain, the matrix transposition can be realized quickly and the following equation can be obtained.

$$\hat{w} = \left( \hat{a}_0 \odot \hat{y} \right) / \hat{a}_0 \odot \hat{a}_0 + \lambda_i$$

(1)

Here, * denotes the conjugate. When detecting an object, filter $w$, which is updating, is convoluted with search window $z$ in the next frame, where $Z$ represents the patch circulant matrix. The position with the largest response value is the predicted position of the object in the next frame in that search window, that is, $r_p (w, Z) = Zw$, which is expressed as follows in the Fourier domain.

$$\hat{r}_p = \hat{Z} \odot \hat{w}$$

(2)

Equation (2) is the detection formula in the original domain.

2.2. Background spatial tracking algorithm
The circulant matrix corresponding to the object area patch $a_0$ of interest is $A_0$, and the circulant matrix $A_i$ corresponding to the object area patch $a_i$ is obtained by the cyclic shifting of $k$ negative sampling areas around it. Here, $a_i \in \mathbb{R}^n$ and $A_i \in \mathbb{R}^{nm}$. We need to train filter $w \in \mathbb{R}^n$ so that the response value in the object area is much higher than that in the background area. By adding a background image patch to the objective function $\min \left\| A_0 w - y \right\|_2^2 + \lambda_1 \left\| w \right\|_2^2$ (Here, vector $w$ represents the trained correlation filter. The square matrix $A_0$ represents the circulant matrix of the image patch $a_0$, translation transformation and objective function $y$ is a two-dimensional Gaussian image vector.) and further adding parameter $\lambda_2$ to control the response intensity to the background region, the response of the filter to the background region is far lower than it is to the object region. The final proposed convex objective function is as follows.

$$\min \left\{ \left\| A_0 w - y \right\|_2^2 + \lambda_1 \left\| w \right\|_2^2 + \lambda_2 \left\| w - \hat{w} \right\|_2^2 \right\}$$

Here, $\hat{w}$ is the updated filter.
\[
\text{min } \| A_i w - y \|_2^2 + \lambda_1 \| w \|_2^2 + \lambda_2 \sum_{i=1}^k \| A_i w \|_2^2
\]  

(3)

2.3. Multi-channel characteristics

In the original domain, the circulant matrix \( A_i \) for \( k \) background information patches is added to the circulant matrix \( A_0 \) of object region \( a_0 \) to form a new matrix \( B \). Because in a single channel, \( B \) is composed of the circulant matrix of the object and background image patches, \( B \in \mathbb{R}^{(k+1) \times n} \). For multi-channel features in the original domain, a matrix \( B \) of size \((k+1) \times nm \) is used instead of \( B \), where \( m \) denotes the number of channels of the feature. At this point, the objective function becomes

\[
f_t(\mathbf{w}) = \|\mathbf{B} \mathbf{w} - \mathbf{y}\|_2^2 + \lambda_1 \| \mathbf{w} \|_2^2
\]

(4)

The filter for multi-channel features can be obtained by setting the gradient of the objective function to zero as follows.

\[
\mathbf{w} = (\mathbf{B}^T \mathbf{B} + \lambda_1 I)^{-1} \mathbf{B}^T \mathbf{y}
\]

(5)

Using the diagonalization properties of circulant matrices, we also obtain

\[
\hat{\mathbf{w}} = \left[ \tilde{C}_{i1} \ldots \tilde{C}_{i,m} \right]^{-1} \begin{bmatrix} \text{diag} (\hat{a}_{0,0} \odot \hat{y}) \\ \vdots \\ \text{diag} (\hat{a}_{0,m} \odot \hat{y}) \end{bmatrix}
\]

(6)

where,

\[
\tilde{C}_{ji} = \text{diag} \left( \hat{a}_{0,j} \odot \hat{a}_{0,i} + \lambda_2 \sum_{i=1}^k \hat{a}_{ij} \odot \hat{a}_{ij} \right) + \lambda_1 I
\]

\[
\hat{a}_{ij} = \text{diag} \left( \hat{d}_{ij} \odot \hat{a}_{ij} \right), \quad j \neq i
\]

In the dual domain, the closed form solution of multi-channel features is given by \( \hat{\alpha} = (\mathbf{B}^T \mathbf{B} + \lambda_1 I)^{-1} \mathbf{y}, \quad \hat{\alpha} \in \mathbb{R}^{km} \). Using the circulant matrix, we obtain

\[
\hat{\alpha} = \begin{bmatrix} \text{diag} \left( \tilde{d}_{0,0} \right) \ldots \text{diag} \left( \tilde{d}_{0,k} \right) \\ \vdots \\ \text{diag} \left( \tilde{d}_{k,0} \right) \ldots \text{diag} \left( \tilde{d}_{kk} \right) \end{bmatrix}^{-1} \begin{bmatrix} \hat{y} \\ \vdots \\ 0 \end{bmatrix}
\]

(7)

where, \( \tilde{d}_{ij} \) is given by

\[
\hat{d}_{ij} = \lambda_2 \sum_{i=1}^m \left( \hat{a}_{ij} \odot \hat{a}_{ij} \right) + \lambda_1
\]

\[
\hat{d}_{ii} = \sqrt{\lambda_2} \sum_{i=1}^m \left( \hat{a}_{ii} \odot \hat{a}_{ii} \right), \quad j \neq i
\]

Detection formula \( r_d(\hat{\alpha}, \mathbf{B}, \mathbf{Z}) = \mathbf{Z}^T \hat{\alpha} \) is diagonalized in the Fourier domain as follows.

\[
\hat{r}_d = \sum_{i=1}^m \hat{z}_i \odot \hat{a}_{0,i} \odot \hat{a}_0 + \sqrt{\lambda_2} \sum_{j=1}^k \sum_{i=1}^m \hat{z}_i \odot \hat{a}_{ij} \odot \hat{a}_j
\]

(8)
3. Experiments

The improved correlation filter method was added to several tracking algorithms based on classic correlation filters: SAMF\[5\], STAPLE\[11\], and MOSSE\[1\]. The experimental results are compared with those of Struck, SCM, TLD, DFT, CXT, LSK, OAB, and all algorithms are compared with each other on the standard data sets OTB-50 and OTB-100\[12,13\]. The experimental results are compared in different scenarios. All algorithm trackers were run on the same operating platform (Intel (R) Core (TM) i7-4790 CPU@3.60 GHz) using MATLAB.

3.1. Quantitative evaluation metrics

Performance was evaluated using the success rate and accuracy metrics defined in OTB-50 and OTB-100. The main method for measuring precision is to calculate the error between the boundary frame and the center of the ground truth. That is, the distance between them should be less than a threshold, which is a percentage of the video sequence frame. Generally, a threshold of 20 pixels is used, which gives different percentages for different frame sizes. The success rate is mainly measured by rate of overlap between the tracking and annotation boxes. When the overlap rate of a frame is greater than a given threshold, the tracking is considered to be successful, and the success rate is the percentage of frames that are successful. In general, the overlapping threshold is set to 0.5. The area under the curve (AUC) for the success rate is used to rank the tracking algorithms. Robustness is evaluated using time-robustness evaluation (TRE) and space-robustness evaluation (SRE) measurements. The OPE metric refers to the average accuracy and success rate when the object location is initialized according to the exact location in the first frame before the algorithm is run on the test sequence.

3.2. Parameters setting

For fair comparison, the correlation filters for the background spatial information were given the same parameters setting. The number of negative sampling areas (background) \( k \) was set to 4 because when \( k \) is greater than 4, the runtime of the algorithm increases significantly but the results are not substantially improved. The improved SAMF, STAPLE, and MOSSE algorithms are referred to as BS-SAMF, BS-STAPLE, and BS-MOSSE, respectively. The learning rate was set to 0.005, 0.015, and 0.025, and parameter \( \lambda_2 \) was set to 0.4, 0.5, and 20 for the three algorithms, respectively.

3.3. Quantitative Results

Figures 1(a)-(f) show the success rates for particular scenarios: illumination variation, scale variation, background clutter, occlusion, fast motion, and motion blur. The results show that the algorithms using the proposed method perform well in most situations and do not drift. In particular, if the appearance of the object (occlusion) or the surrounding background (background clutter) changes dramatically in successive frames, larger search and matching areas based on the background spatial information gives better results than other methods. Figures 2(a)-(d) presents the TRE precision in other scenarios.
Figure 1. SRE results (AUC). BS_SAMF (green curve), BS_STAPLE (red curve), and BS_MOSSE are compared with other methods.

Figure 2. TRE results for deformation, low resolution, in-plane rotation, and out-of-view scenes. BS_SAMF (green curve), BS_STAPLE (red curve), and BS_MOSSE are compared with other methods.

4. Conclusion
This paper proposed a method for improving correlation filter trackers by sampling the background information around the object while overlapping the background region with the object region. This approach led to good tracking results with low computational cost, especially for some specific scenarios such as illumination changes, occlusion, scale changes, background clutter, and rapid motion. Trackers using the approach can accurately detect and track the object while avoiding drift. The experimental results show that introducing negative sample information improves the tracking performance of the correlation filter compared with algorithms that do not use this information.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (61872042, 61572077), and the Key Projects of the Beijing Education Commission (KZ201911417048).
References

[1] Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui,Y.M. (2010) Visual object tracking using adaptive correlation filters. Computer Vision and Pattern Recognition (CVPR), 2010:2544–2550.

[2] Henriques, J.F., Rui, C., Martins, P. (2012) Exploiting the Circulant Structure of Tracking-by-Detection with Kernels[M]. Computer Vision – ECCV, Springer Berlin Heidelberg, 2012:702–715.

[3] Danelljan, M., Khan, F.S., Felsberg, M., Weijer, J.V.D. (2014) Adaptive colour attributes for real-time visual tracking. Computer Vision and Pattern Recognition (CVPR), 2014:1090–1097.

[4] Henriques, J.F., Rui, C., Martins, P. (2014) High-Speed Tracking with Kernelized Correlation Filters. IEEE Transactions on Pattern Analysis & Machine Intelligence, 37:583–596.

[5] Li, Y., Zhu, J. (2014) A scale adaptive kernel correlation filter tracker with feature integration. Computer Vision-ECCV 2014 Workshops, 2014:254–265.

[6] Danelljan, M., Hager, G., Khan, F.S., Felsberg, M. (2014) Accurate scale estimation for robust visual tracking. Proceedings of the British Machine Vision Conference BMVC, 2014:1–11.

[7] Hong, Z., Chen, Z., Wang, C., Mei, X., Prokhorov, D., Tao, D. (2015) Multistore tracker (muster): A cognitive psychology inspired approach to object tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:749–758.

[8] Ma, C., Yang, X., Zhang, C., Yang, M.H. (2015) Long-term correlation tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:5388–5396.

[9] Liu, T., Wang, G., Yang, Q. (2015) Real-time part-based visual tracking via adaptive correlation filters. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:4902–4912.

[10] Li, Y., Zhu, J., Hoi, S.C. (2015) Reliable patch trackers: Robust visual tracking by exploiting reliable patches. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:353–361.

[11] Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.S. (2016) Staple: Complementary learners for real-time tracking. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016:1401–1409.

[12] Wu, Y., Lim, J., Yang, M.H. (2013) Online object tracking: A benchmark. The IEEE Conference on Computer vision and pattern recognition (CVPR), 2013:2411–2418.

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