Correlation filter for object tracking with temporal-spatial constraint

Junxu Wei\textsuperscript{1,3}, Lifeng Yang\textsuperscript{1,3}, Tian Pu\textsuperscript{2,3}, Jian Li\textsuperscript{2,3}, Zhenming Peng\textsuperscript{2,3}

\textsuperscript{1} School of Optoelectronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
\textsuperscript{2} School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
\textsuperscript{3} Laboratory of Imaging Detection and Intelligent Perception, University of Electronic Science and Technology of China, Chengdu 610054, China

E-mail: zmpeng@uestc.edu.cn

Abstract. In this work, inspired by the Passive-Aggressive learning (PA), we proposed a Temporal-Spatial Constraint Correlation Filter (TSCF) model to simultaneously constrain the spatial mask and the update direction of the filter. Firstly, the spatial regular term ensures that the background redundancy information does not interfere with the filter update during the tracking process. Secondly, the temporal regular term ensures that the spatial mask and the filter do not change dramatically. Thirdly, our proposed TSCF model can be effectively solved based on the alternate direction method of multiplier (ADMM), where each sub-problem has a closed solution. Finally, our experiments on the OTB100 benchmark shows that our tracker has efficient performance compare with many advanced algorithms, which get an AUC score of 0.599 and an accuracy of 0.794.

1. Introduction

Target tracking is a general area of computer vision, widely used in the fields of machine vision \cite{1,2,3}, missile tracking, and precise positioning. In the actual environment, the effects of illumination, occlusion, background clutter and other challenges make the tracking effect decrease. Therefore, it is necessary to design a tracker that can maintain high robustness under the rapid change of target.

The current tracking method is mainly generative method and discriminative method. The generative method uses the similarity between the tracking target and the candidate region to achieve the target tracking. The generative method only utilizes the appearance characteristics of the target itself, resulting in low robustness when the target appearance changes rapidly. The discriminative method treats the tracking as a two-category problem, and designs a classifier that can effectively distinguish the target and the background to achieve tracking. For the discriminant method is more robust, it's currently the mainstream method of target tracking.

In recent years, through the research work of many scholars, the correlation filter methods have achieved great development. Bolme et al. \cite{4} first proposed correlation filter to the target tracking domain. This method is based on the minimum output sum of squared error (MOSSE) training correlation filter of the output result. The correlation between the candidate region and the initialization target is judged by the value of the response image, and the algorithm uses Fourier for speeds up the calculations to ensure real-time performance. Henriques et al. \cite{5} proposed the circulant structure of tracking-by-detection with kernels algorithm (CSK), which used the cyclic matrix to
intensively sample search area in order to effectively utilize the characteristics of the whole picture. Henriques et al. [6] introduced a kernel mechanism based on CSK, and proposed kernelized correlation filter (KCF), which uses multiple histogram of oriented gradients (HOG) features in the algorithm to map the linear regression problem of the linear space to the nonlinear space. Danelljan et al. [7] proposed discriminative scale space tracker (DSST) based on MOSSE for scale change problem. The algorithm trains two filters respectively, scale filter and position filter. The scale estimation and position estimation are regarded as two problems. In the process of tracking, two filters work independently and predict target localization and scale respectively, which achieving excellent performance. Based on feature fusion, Bertinetto et al. [8] proposed a feature-complementary tracking method, which combines HOG features with color-name (CN) features, effectively solving the problem that single HOG features are not robust under illumination changes and single CN features are not robust under rapid deformation. However, it sacrifices huge efficiency and improves the algorithm time complexity, and only adds a little tracking accuracy. The artificially chosen kernel function cannot be guaranteed to be optimal, which will cause systematic errors, Tang et al. [9] proposed multiple KCF (MKCF), which introduces multiple kernel functions based on KCF and proposes a corresponding solution method. Zhang et al. [10] combine multiple features, introduce multiple kernel functions in the calculation, integrate multiple filters, and adjust the weight of the kernel through the final tracking result, so that the core has the ability to adaptively select.

Although the correlation filtering algorithm has been widely used due to its speed advantage and good precision, it also has some shortcomings. First, the correlation filtering algorithm belongs to the template matching method. It has poor tracking effect for fast deformation and fast motion, and it’s easy to generate model drift phenomenon, which cause tracking failure. Second, the current correlation filtering algorithm is processed by the ideal Gaussian window, which cannot be updated in real time according to the change of the target appearance.

In view of the above problems, in this paper, we designed a correlation filter tracker with high robustness in the case of rapid deformation through two main tasks: correlation filter based on temporal and spatial constraints. The contribution of this work can be summarized as follows. First, we propose an adaptive spatial regularization term, which can effectively estimate the actual contour of objects. Second, we propose a temporal regular term based on time difference, which can effectively suppress the change of the current filter and space terms dramatically. Third, because the proposed spatial regular terms and temporal regular terms do not change the strong convex properties of traditional ridge regression, we use ADMM to efficiently solve the results. Overall, our tracker is capable of high accuracy and robustness under fast deformation conditions.

![Frame 2](image1.png)  ![Frame 129](image2.png)  ![Frame 184](image3.png)

**Figure 1.** Schematic diagram of spatial constraints.
2. Related work
Because the tracking target is subject to external factors (such as occlusion and lighting changes) and internal factors (such as rotation and deformation), this causes a significant change in the appearance contours, and allows the tracker to gradually focus on the wrong target, which eventually leads to model drift problems. In response to this problem, Danelljan et al. [12] received the SRDCF’s [11] inspiration propose the SRDCFdecon, which formed a data set by continuously accumulating the tracking samples, and dynamically adjusted the weights of each sample in the data set to make the filter change in the direction of tracking success. However, this method requires calculations for all samples in each sample set, and requires constant update weights, which greatly increases the time complexity but only improve the small precision under rapid changes. Li et al. [13] also made improvements on SRDCF, in which the tracking information of the previous frame was introduced into the calculation process of the current frame. However, this method cannot accurately represent the century contour of the tracking target from beginning to end. In this case, it is directly used as prior knowledge to participate in the current filter update process, which causes the component to move the focus of the filter to the wrong target after the target is changed dynamically, eventually leading to tracking failure.

3. Correlation filter with temporal-spatial constraint (TSCF)
In this section, we introduce the proposed correlation filter based on space-time constraints, and finally use the ADMM method to solve the problem efficiently.

3.1. Spatially Regularized Correlation Filters (SRDCF)
SRDCF method to learn effective multichannel CFs, and its objective function is defined as follows
\[
E(H, w) = \frac{1}{2} \left| \left| y - \sum_{k=1}^{K} x_k \ast h_k \right| \right|^2 + \frac{\lambda_1}{2} \left| \left| w \right| \right|^2 + \frac{\lambda_2}{2} \left| \left| w - w' \right| \right|^2 + \frac{\lambda_3}{2} \sum_{k=1}^{K} \left| \left| h_k - h_k^{-1} \right| \right|^2 + \frac{\lambda_4}{2} \left| \left| w - w'^{-1} \right| \right|^2
\]

(1)

where \( X = \{x_1, x_2, \ldots x_K\} \) is train data, \( H = \{h_1, h_2, \ldots h_K\} \) is filter, \( w \) is adaptive spatial weight, \( \lambda_i \) is the regularization parameters.

### 3.2. Our Objective Function

Motivated by online Passive-Aggressive (PA) [11] learning, we introduce weight-temporal regularisation and filter-temporal regularisation, take the SRDCF as baseline, resulting in our temporal-space constraint correlation filter (TSCF). TSCF can be exploited for simultaneous DCF learning and model updating. Besides, the ADMM algorithm can also be directly used to solve TSCF. Thus, TSCF incorporates both spatial and temporal regularisation into DCF. Our objective function is defined as follow:

\[
E(H, w) = \frac{1}{2} \left| \left| y - \sum_{k=1}^{K} x_k \ast (P^T h_k) \right| \right|^2 + \frac{\lambda_1}{2} \left| \left| w \right| \right|^2 + \frac{\lambda_2}{2} \left| \left| w - w' \right| \right|^2 + \frac{\lambda_3}{2} \sum_{k=1}^{K} \left| \left| h_k - h_k^{-1} \right| \right|^2 + \frac{\lambda_4}{2} \left| \left| w - w'^{-1} \right| \right|^2
\]

(2)

In this equation, the first term is the ridge regression term, \( X = \{x_1, x_2, \ldots x_K\} \) is training data, \( H = \{h_1, h_2, \ldots h_K\} \) is the filter, convolving \( X \) and \( H \) to fit the Gaussian-distributed ground truth \( y \). The second term is a regularization term introducing an adaptive spatial regularization on the filter \( H \). The third term attempts to make the adaptive spatial weight \( w \) be similar to a reference weight \( w' \), this constraint introduces a priori information on \( w \) and avoids model degradation. The fourth term and fifth term are temporal regularizations, they attempt to avoid mutations in adjacent two-frame filters and spatial constraints. \( \lambda_1, \lambda_2, \lambda_3 \) and \( \lambda_4 \) are the regularization parameters. \( P \in R^{T \times T} \) is a diagonal binary matrix to make the correlation operator directly apply on the true foreground and background samples.

### 3.3. Optimization

Correlation filters are usually learned in the frequency domain for effectiveness.

\[
E(H, \hat{G}, w) = \frac{1}{2} \left| \left| y - \sum_{k=1}^{K} \hat{x}_k \ast \hat{g}_k \right| \right|^2 + \frac{\lambda_1}{2} \left| \left| w \right| \right|^2 + \frac{\lambda_2}{2} \left| \left| w - w' \right| \right|^2 + \frac{\lambda_3}{2} \sum_{k=1}^{K} \left| \left| h_k - h_k^{-1} \right| \right|^2 + \frac{\lambda_4}{2} \left| \left| w - w'^{-1} \right| \right|^2
\]

s.t., \( \hat{g}_k = \sqrt{T}F^P h_k, k = 1, \ldots, K \)

(3)

where \( \hat{G} = \{\hat{g}_1, \hat{g}_2, \ldots \hat{g}_K\} \) is an auxiliary variable matrix. \( F \) is the orthonormal \( T \times T \) matrix of complex basis vectors to map \( T \)-dimensional vectorized signal into Fourier domain. For ADMM to solve this equation, the augmented Lagrangian from eq.3 can be formulated as

\[
L(H, \hat{G}, w, \hat{S}) = \frac{1}{2} \left| \left| y - \sum_{k=1}^{K} \hat{x}_k \ast \hat{g}_k \right| \right|^2 + \frac{\lambda_1}{2} \left| \left| w \right| \right|^2 + \frac{\lambda_2}{2} \left| \left| w - w' \right| \right|^2 + \frac{\lambda_3}{2} \sum_{k=1}^{K} \left| \left| h_k - h_k^{-1} \right| \right|^2 + \frac{\lambda_4}{2} \left| \left| w - w'^{-1} \right| \right|^2 + \frac{\lambda_5}{2} \sum_{k=1}^{K} \left| \left| \hat{g}_k - \sqrt{T}F^P h + \hat{S}_k \right| \right|^2
\]

(4)
where \( \hat{S} = \{ \hat{s}_1, \hat{s}_2, \ldots, \hat{s}_k \} \), \( s_k = \frac{1}{\mu} v_k \), \( v_k \) is the Lagrange multiplier.

Then, the ADMM algorithm is adopted by alternately solving the following subproblems:

**Subproblem H:** if \( \hat{G}, w \) and \( \hat{S} \) are given, the optimal \( H^* \) can be obtained as

\[
h_k^* = \arg \min_{h_k} \left\{ \frac{\lambda_1}{2} \| w \odot h_k \|_2^2 + \frac{\lambda_2}{2} \| h_k - h_k^{-1} \|_2^2 + \frac{\mu}{2} \| g_k - \sqrt{T} F P^T h + s_k \|_2 \right\} \\
= \frac{u T p \odot (s_k + g_k) + \lambda_3 h_k^{-1}}{\lambda_1 (w \odot w) + \lambda_3 I + u T p}
\]

**Subproblem G:**

\[
\hat{G}^* = \arg \min_G \left\{ \frac{1}{2} \| y - \sum_{k=1}^{K} \hat{g}_k \odot \hat{g}_k \|_2^2 + \frac{\mu}{2} \sum_{k=1}^{K} \| g_k - \sqrt{T} F P^T h + s_k \|_2 \right\}
\]

However, this problem has high computation complexity. Thus, we consider processing on all channels of each pixel, could be reformulated as:

\[
V_j^*(\hat{G}) = \arg \min_{v_j(\hat{G})} \left\{ \frac{1}{2} \| y - V_j(\hat{X})^T V_j(\hat{G}) \|_2^2 + \frac{\mu}{2} \sum_{k=1}^{K} \| V_j(\hat{G}) + V_j(\hat{M}) \|_2^2 \right\} \\
= \frac{1}{u T} \left( \frac{V_j(\hat{X}) V_j(\hat{X})^T}{u T + V_j(\hat{X}) V_j(\hat{X})} \right) \| y - V_j(\hat{X}) + u V_j(\sqrt{T} F P^T H) - \mu V_j(\hat{S}) \|_2
\]

where \( V_j(\hat{M}) = V_j(\hat{S}) - V_j(\sqrt{T} F P^T H) \), \( V_j(\hat{g}) \in R^{K \times 1} \) denotes the value of all channels of filter \( \hat{g} \) on pixel \( j \).

**Subproblem w:**

\[
w^* = \arg \min_w \left\{ \frac{\lambda_1}{2} \sum_{k=1}^{K} \| w \odot h_k \|_2^2 + \frac{\lambda_2}{2} \| w - w' \|_2^2 + \frac{\lambda_4}{2} \| w - w^{-1} \|_2 \right\} \\
= (\lambda_1 \sum_{k=1}^{K} h_k^T \odot h_k + \lambda_2 I + \lambda_4 I)^{-1}(\lambda_2 w' + \lambda_4 w^{-1}) \\
= \frac{\lambda_2 w' + \lambda_4 w^{-1}}{\lambda_1 \sum_{k=1}^{K} h_k^T \odot h_k + \lambda_2 I + \lambda_4 I}
\]

**Lagrangian Multiplier Update:**

\[
\hat{S}^{i+1} = \frac{\hat{S}^i + \hat{G}^{i+1} - \hat{H}^{i+1}}{2}
\]

where \( \hat{S}^{i+1} \) denotes the Fourier transform of the Lagrangian in the previous state, \( \hat{G}^{i+1} \) and \( \hat{H}^{i+1} \) are the current solutions to the two subproblems above at \( i + 1 \)th. The regularization constant \( \mu \) is commonly set as \( \mu^{i+1} = \min(\mu_{\text{max}}, \beta \mu^i) \).
Algorithm 1 TSCF algorithm at time step $t$.

Input
- Image, $I$;
- Correlation tracker and it’s frequency domain expression $H_t$ and $G_t$
- Lagrange multiplier, $S_t$
- Reference weight, $w_t$

Output
- Estimated $H_{t+1}$ and $G_{t+1}$;
- Update $S_{t+1}$;

Initialize: $\lambda_1 = 0.2$, $\lambda_2 = 0.001$, $\lambda_3 = 1000$, $\lambda_4 = 500$

For $t=2$ to end do
- Extracting FHOG features of $I$ as $x_t$;
- Computing response map with $x_t$ and $g_t$;
- Updating $H_{t+1}$ using Eq.(5);
- Updating $G_{t+1}$ using Eq.(7);
- Updating $w_{t+1}$ using Eq.(8);
- Updating $S_{t+1}$ using Eq.(9);

End for

return

4. Experiment
We implemented our tracker on the Matlab2018a, and runs on a PC with an Intel 8700k CPU, 32GB RAM. Our tracker could achieve 24fps approximately, which could be regarded as real-time tracker. We merely use five-scale HOG features as input. The regularization parameters $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ are empirically chosen as $\lambda_1 = 0.2$, $\lambda_2 = 0.001$, $\lambda_3 = 1000$ and $\lambda_4 = 500$, respectively. The penalty factor $\mu$ of ADMM is initially set to 1 and then updated by $\mu^{i+1} = \min(\mu_{\text{max}}, \beta\mu^i)$, where $\mu_{\text{max}} = 1000$ and $\beta = 10$.

We employ the image sequence basketball, whose resolution is $576 \times 432$, and the total number of frames are 725. This test mainly focuses on the tracking ability of the algorithm in the case of occlusion and deformation. We compare the improved algorithm proposed in this paper, BACF and SRDCFdecon. The result is shown in figure 3.
We employ the image sequence soccer, whose resolution is 640 × 360, and the total number of frames are 392. This test mainly focuses on the tracking ability of the algorithm in the case of fast motion and Out-of-Plane Rotation. We compare the improved algorithm proposed in this paper, BACF and SRDCFdecon. The result is shown in figure 4.
In order to evaluate the performance of TSCF, the experiment selects the current popular benchmark OTB100 for testing. The OTB100 contains 100 video sequences covering 11 different visual challenges, such as illumination variation (IV), scale transformation (SV), motion blur (MB), etc., each sequence may have multiple attributes.

As shown in figure 5, in general, our proposed tracking algorithm yields almost optimal results with an AUC score of 0.599 and an accuracy of 0.794.

Figure 6 shows AUC images under different challenges, including deformation, occlusion and rotation. Compared with SRDCF, the proposed TSCF improves by 4.8%, 2.0% and 3.5% respectively. The main contribution is that our proposed time and space constraints can enable the filter to learn the correct information during the tracking process.

5. Conclusion
In this paper, we propose a Temporal-Spatial constraint correlation filter to solve the model drift problem caused by rapid deformation. Spatial template and filter changes dramatically are prevented by introducing time-regular terms and spatial regular terms into SRDCF. In addition, as an extension of the online PA, the TSCF can adaptively balance the trade-off between active model learning and passive model learning, thereby providing a more reliable model in the case of large changes in appearance. TSCF does not change the convexity of the original SRDCF and can be efficiently solved using the ADMM algorithm. We conducted experiments on the OTB100 benchmark. The results show that TSCF is the preferable compared to the well-known algorithms in the tracking field, both in terms of tracking accuracy and robustness.
6. References

[1] Peng, Z, Zhang, Q, and Guan, A. Extended target tracking using projection curves and matching pel count[J]. Optical Engineering, 2007, 46(6): 066401.

[2] Li, M, Peng, Z, Chen, Y, Wang, X, Peng, L, Wang, Z, Yuan G, and He, Y. A novel reverse sparse model utilizing the spatio-temporal relationship of target templates for object tracking[J]. Neurocomputing, 2019, 323: 319-334.

[3] Li, M, Peng, L, Chen, Y, Huang, S, Qin, F, & Peng, Z. Mask Sparse Representation Based on Semantic Features for Thermal Infrared Target Tracking[J]. Remote Sensing, 2019, 11(17): 1967.

[4] Bolme D S. Visual object tracking using adaptive correlation filters[C]. The IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, 2010: 2544-2550.

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

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

[7] Danelljan, M, Häger, G, Khan, F, and Felsberg, M. Accurate scale estimation for robust visual tracking[C]. British Machine Vision Conference, Nottingham, September 1-5, 2014. BMVA Press, 2014.

[8] Bertinetto, L, Valmadre, J, Golodetz, S, Miksik, O, and Torr, P. Staple: Complementary learners for real-time tracking[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1401-1409.

[9] Tang M and Feng J. Multi-kernel correlation filter for visual tracking[C]. Proceedings of the IEEE international conference on computer vision. 2015: 3038-3046.

[10] Zhang, K, Zhang, L, Liu, Q, Zhang, D, and Yang, M. H. Fast visual tracking via dense spatio-temporal context learning[C]. European conference on computer vision. Springer, Cham, 2014: 127-141.

[11] Danelljan, M, Hager, G, Shahbaz Khan, F, and Felsberg, M. Learning spatially regularized correlation filters for visual tracking[C]. Proceedings of the IEEE international conference on computer vision. 2015: 4310-4318.

[12] Danelljan, M, Robinson, A, Khan, F. S, and Felsberg, M. Beyond correlation filters: Learning continuous convolution operators for visual tracking[C]. European Conference on Computer Vision. Springer, Cham, 2016: 472-488.

[13] Li, F, Tian, C, Zuo, W, Zhang, L, and Yang, M. H. Learning spatial-temporal regularized correlation filters for visual tracking[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4904-4913.

[14] Crammer, K, Dekel, O, Keshet, J, Shalev-Shwartz, S, and Singer, Y. Online passive-aggressive algorithms[J]. Journal of Machine Learning Research, 2006, 7(Mar): 551-585.

Acknowledgments
This work was funded by National Natural Science Foundation of China (61775030,61571096), Sichuan Science and Technology Program (2019YJ0167 and 2019YFG0307), and Open Research Fund of Key Laboratory of Optical Engineering, Chinese Academy of Sciences (2017LBC003).