Visual Tracking using Spatial-Temporal Regularized Support Correlation Filters

Binshan Li1, Chaorong Liu2, Jie Liu3, Huiling Gao1, Xuhui Song1 and Weirong Liu1,*

1 College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China
2 Key Laboratory of Gansu Advanced Control for Industrial Processes, Lanzhou University of Technology, Lanzhou 730050, China
3 National Demonstration Center for Experimental Electrical and Control Engineering Education, Lanzhou University of Technology, Lanzhou 730050, China

*Corresponding author’s e-mail: liu_weirong@163.com

Abstract. Support Correlation Filters (SCFs) have recently shown great potentials in real-time visual tracking. However, most of existing SCF trackers learn appearance models using the information of current frame, and completely neglect inter-frame information. Besides, they still suffer from unwanted boundary effects. In this paper, we proposed a novel Spatial-Temporal Regularized Support Correlation Filter (STRSCF) model, which introduces the spatial weight and temporal regularization term into SCF model. In order to improve the tracking performances, we extend STRSCF to multi-dimensional feature space. In addition, an effective optimization algorithm is developed to solve our STRSCF model in closed form solution. The experimental results on OTB-13 demonstrate that the STRSCF tracker performs superiorly against several state-of-the-art trackers in terms of accuracy and speed.

1. Introduction

Recent years have witnessed the rapid advances of discriminative methods in visual tracking. Discriminative methods regard the tracking problem as a regression or classification problem, which exploits classifiers to separate the tracking target from the background. Algorithms based on discriminative framework have demonstrated superior tracking accuracy, these algorithms include Support Vector Machines (SVMs) [1-4], Correlation Filters (CFs) [5-10], multiple instance learning [11] and so on.

The CF-based approaches. Benefited from the periodic assumption of training samples, the CF models can be learned very efficiently in the frequency domain via Fast Fourier Transform (FFT). The first CF-based method runs with the speed about 700 Frames Per Second (FPS) in [5]. The CF-based methods improve the tracking performances by introducing complex hand-crafted features [9], nonlinear kernel [7], multi-scale [12] and CNN features [10]. However, the run speeds of those trackers reduce due to extra costs. The CF-based trackers using CNN features have even lost the real-time capability.

The SVM-based approaches. The tracking performance of the SVM-based methods can be improved by two aspects: model [1] and features [2-4]. Moreover, Zuo et al. [13] propose a SCF method that exploits circulant matrix to formulate SVM model. Because the circulant matrix can be...
operated via Discrete Fourier Transform (DFT), the SVM model is efficiently learned by iterative updating in each frame, which also improve the tracking speed of SVM tracker. However, this model suffers from boundary effects [8] due to the introduction of periodic assumption. Besides, the SCF method learns the appearance model by exploiting the information of current frame, without the inter-frame information, which is hard to learn an effective appearance model for visual tracking.

In this paper, we present a novel tracking approach named STRSCF that introduces a spatial weight and a temporal regularization term, and rewrite its optimization algorithm. By utilizing the weighted dense samples, we make the trained classifier more focus on the object rather than boundary region. With the temporal regularization term, we learn an effective SCF model, and optimize it in closed form solution. Furthermore, we extend the proposed STRSCF model to multi-dimensional feature space. The procedure of our STRSCF tracker is shown in Figure 1.

![Figure 1](image)

**Figure 1.** Overall procedure of the proposed STRSCF tracker. Our tracker framework includes two main steps: updating and tracking.

2. Method

In this section, we first give the SCF model and then introduce our Spatial-Temporal Regularized Support Correlation Filter (STRSCF) model and optimization algorithm. Finally, we extend our STRSCF model to multi-dimensional feature space.

2.1. The SCF formulation

We keep the same notations as [13] to write the SCF formulation in the following. Given a sequence image $x$, we use a circulant matrix $X$ to represent the full set of translated image versions, and each row of $X$ stands for one candidate of a target. We define the eigenvectors of $X$ in Fourier domain:

$$X = V^H \text{Diag}(\hat{x}) V,$$

where $\hat{x}$ is the Fourier transform of sequence image $x$, and $\text{Diag}(\cdot)$ denotes the diagonal matrix from a vector. $V$ denotes the base vector and $V^H$ is its Hermitian transpose.

The SCF model is aimed at learning a set of variables: classifier $w$ and bias $b$ by,

$$y = \text{sgn} \left( \mathcal{F}^{-1} \left( \hat{x} \odot \hat{w} \right) + b \right),$$

where $\text{sgn}(\cdot)$ denotes a sign function, and $\mathcal{F}^{-1}(\cdot)$ denotes the Inverse Discrete Fourier Transform (IDFT). $\odot$ denotes the element-wise multiplication operator. $\hat{x}^*$ denotes the complex conjugate of $\hat{x}$, and $y$ is its label set.
For a \( n \times n \) image \( x \), the model variables including \( w \) and \( b \) in (2) can be learned from the training samples of \( X = [x_1; x_2; \ldots; x_n] \) and the corresponding labels \( y = [y_1; y_2; \ldots; y_n] \), by solving the following global optimization problem:

\[
\min_{w, b} \sum_{i=1}^{n} ||y_i - \mathcal{S}^{-1}(\mathbf{x_i} \circ \mathbf{w}) - b||^2 + \mu ||\mathbf{w} - \mathbf{w}_{t-1}||^2 + C ||y - \mathcal{S}^{-1}(\mathbf{x_i} \circ \mathbf{w}) - b||^2, \quad \text{s.t. } e \geq 0,
\]

where \( C \) is the regularization parameter, \( 1 \) is a notation and denotes a vector of 1s. Given the target position \( \mathbf{p}_t \) and each sample position \( \mathbf{p}_i \) in the image \( x \), we can compute the class label of \( \mathbf{x}_i \) using a confidence map [14] that is written as:

\[
m(\mathbf{p}_i, \mathbf{p}_t) = \gamma \exp\left(-\alpha ||\mathbf{p}_i - \mathbf{p}_t||^2\right),
\]

where \( \gamma \) is a normalization constant, and \( \exp(\cdot) \) is an exponential function. \( \beta \) and \( \alpha \) denote the shape and scale parameters, respectively.

According to the confidence map \( m(\mathbf{p}_i, \mathbf{p}_t) \), we define the class labels \( y_i \) as follows:

\[
y_i = \begin{cases} 
1, & \text{if } m(\mathbf{p}_i, \mathbf{p}_t) \geq \theta_a, \\
-1, & \text{if } m(\mathbf{p}_i, \mathbf{p}_t) \leq \theta_b.
\end{cases}
\]

where \( \theta_a \) is the lower threshold, and \( \theta_b \) is the upper threshold.

2.2. The proposed STRSCF model

In general, the region near by the target central position \( \mathbf{p}_t \) contains the major part of the target. On the other hand, the classifier should pay closer attention to the central region, rather than the boundary region of \( x \). Therefore, motivated by the spatial regularization model in [8], we introduce a spatial weight \( s \) to enhance central region and penalize boundary region, where the spatial weight \( s \) is a Gaussian function defined by,

\[
s(\mathbf{p}_i, \mathbf{p}_t) = \rho e^{-\frac{||\mathbf{p}_i - \mathbf{p}_t||^2}{\sigma^2}},
\]

where \( \rho \) is a normalization constant that normalize \( s \) to the range between 0 and 1, and \( \sigma \) is a scale parameter. In spatial domain, we take the spatial weight \( s \) to the image \( x \) based on element-wise multiplication, and (1) becomes:

\[
X = V^T \text{Diag}(s \circ x)V,
\]

In online tracking, the target and its background always show continuous change. Thus the learned classifiers maintain continuity from frame to frame. On the other hand, the current classifier should be similar to the previous one. In order to make the learned classifier similar to the previous, we introduce a temporal regularization term \( ||\mathbf{w} - \mathbf{w}_{t-1}||^2 \) into the SCF model, and (3) can be written as,

\[
\min_{w, b} \sum_{i=1}^{n} ||y_i - \mathcal{S}^{-1}(\mathbf{x_i} \circ \mathbf{w}) - b||^2 + \mu ||\mathbf{w} - \mathbf{w}_{t-1}||^2 + C ||y - \mathcal{S}^{-1}(\mathbf{x_i} \circ \mathbf{w}) - b||^2, \quad \text{s.t. } e \geq 0,
\]

where \( \mu \) is the temporal regularization parameter, and \( \mathbf{w}_{t-1} \) is the previous support correlation filter.

The optimization problem of (8) can be decomposed into three subproblems: subproblem \( e \), subproblem \( (w, b) \) and subproblem \( y \). Each subproblem has its own solution when the other variables are given. Therefore, the STRSCF model can be efficiently solved by the following algorithm:

Subproblem \( e \). Given label set \( y \) and model variables \( (w, b) \), (8) can be written as:

\[
\min_{e} ||e - y \circ \mathcal{S}^{-1}(\mathbf{x} \circ \mathbf{w}) + b||^2, \quad \text{s.t. } e \geq 0.
\]

When we define \( e_0 = y \circ \mathcal{S}^{-1}(\mathbf{x} \circ \mathbf{w}) - 1 \), the close-form solution of subproblem \( e \) is,

\[
e = \max\{e_0, 0\},
\]
Subproblem \( \{w, b\} \). Given \( e \), class label \( y \) and the previous classifier \( w_{t-1} \), we get \( y \circ y = 1 \) due to \( y_i \in \{1, -1\} \). With \( f = y \circ (1 + e) \), (8) can be written as:

\[
\min_{[w, b]} \left\| w \right\|^2 + \mu \left\| w - w_{t-1} \right\|^2 + C \left\| \mathcal{F}^{-1}(\hat{x}^* \circ \hat{w}) + b1 - f \right\|^2_2. \tag{11}
\]

Before solving the subproblem (11), we define \( \bar{x} \) as the mean vector of \( x \), and let \( x_c = x - \bar{x} \). Similarly, \( \bar{f} \) is the mean of \( f \), and \( \hat{f} = f - \bar{f} \). The close-form solution of (11) are,

\[
\begin{align*}
\hat{w} &= \frac{\hat{x}^* \circ \hat{f} + \mu w_{t-1}}{\hat{x}^* \circ \hat{x} + \frac{\mu}{\varepsilon}}, \\
b &= \bar{f}.
\end{align*}
\tag{12}
\]

Subproblem \( y \). Given \( e \) and model variables \( \{w, b\} \), the class labels \( y = [y_1; y_2; \ldots; y_n] \) can be computed by:

\[
\min_{y_i \in \{1, -1\}} \left\| y_i (w^T x_i + b) - 1 - e \right\|^2, \quad i = 1, 2, \ldots, n^2, \tag{13}
\]

The solution of Subproblem \( y \) is,

\[
y_i = \begin{cases} 1, \text{ if } w^T x_i + b \geq 0, \\ -1, \text{ if } w^T x_i + b \leq 0. \end{cases} \tag{14}
\]

2.3. STRSCF for multi-dimensional features
We further extend our STRSCF model to multi-dimensional feature space. The multi-dimensional features include Color Names (CN) and Histogram of Oriented Gradient (HOG). For one RGB image, we extract \( L \)-dimensional image features for our tracker, which includes 11-dimensional CN features and 33-dimensional HOG features. Therefore, our STRSCF model can be written as:

\[
\min_{y, e \in \{w, b\}} \left\| w \right\|^2 + C \left\| y \circ \left( \mathcal{F}^{-1} \left( \sum_i (\hat{x}^{(i)})^* \circ \hat{w}^{(i)} \right) + b1 \right) - 1 - e \right\|^2 + \mu \left\| w - w_{t-1} \right\|^2, \quad \text{s.t. } e \geq 0, \tag{15}
\]

where \( x^{(i)} \) denotes the \( i \)-th dimensional of image features. \( w^{(i)} \) denotes the support correlation filter in the corresponding channel. The final support correlation filter \( w \) consists of \( w^{(i)} (l = 1, 2, \ldots, L) \), and is denoted by \( w = [w^1; w^2; \ldots; w^L] \).

3. Experiments
In this section, we first introduce the experimental setup including the benchmark dataset, evaluation metrics and parameter settings that we used, and then analysis the characteristics of our STRSCF tracker. Finally, we give a general comparison with several state-of-the-art trackers.

3.1. Experimental setup
We evaluate the performances of STRSCF method on a benchmark dataset (OTB-13) [15] of 50 sequences. For the evaluation protocol and metrics, the one-pass evaluation (OPE) protocol is used to evaluate the tracking performances of all trackers. It directly reports the results by the success plots and precision plots with 20 pixels threshold. Moreover, we report the run speeds via mean FPS.

We implement the proposed method in MATLAB on a PC with Intel i5-3210 CPU (2.50 GHz) and 4 GB memory. For fair evaluations, we use fixed parameter values in general comparison. In the proposed STRSCF model, we set \( C = 10^4 \) in (12), scale and shape parameters \( (\alpha = 1.04, \beta = 1.5) \) in (4), and the threshold \( (\theta_0 = 0.6, \theta_1 = 0.3) \) in (5). The parameters \( \sigma = \text{mean}(w, d) \) in (6), where \( (w, d) \) denotes the size of the initial target. The cell size and orientations of HOG features are set to 4 and 9,
respectively. To further analyse the performance effect, we set temporal regularization parameter $\mu$ to the range between 0 and 0.1, and fix $\mu=0.01$ for general comparison.

3.2. Analyses of the proposed method

We implement different versions of our STRSCF model. Table 1 shows the characteristics and tracking performances of those versions, which includes OPE and mean FPS. Note that the mean FPS is estimated on the same sequence for each version, which is sequence doll with 3872 frames.

| Trackers  | Feature type | High dimension feature | Temporal regularization parameter $\mu$ | OPE precision (20 pixels) | OPE success (AUC) | Mean FPS |
|-----------|--------------|------------------------|----------------------------------------|---------------------------|-------------------|----------|
| STRSCF    | CN+HOG       | yes                    | 0.1                                    | 0.791                     | 0.575             | 26       |
| STRSCF    | CN+HOG       | yes                    | 0.05                                   | 0.802                     | 0.581             | 25       |
| STRSCF    | CN+HOG       | yes                    | 0.02                                   | 0.834                     | 0.611             | 25       |
| STRSCF    | CN+HOG       | yes                    | 0.01                                   | 0.876                     | 0.628             | 26       |
| STRSCF    | HOG          | yes                    | 0                                    | 0.829                     | 0.608             | 25       |
| STRSCF    | Raw          | no                     | 0.01                                  | 0.675                     | 0.480             | 19       |
| SCF       | Raw          | no                     | —                                     | 0.633                     | 0.468             | —        |
| KCF       | HOG          | yes                    | —                                     | 0.740                     | 0.514             | —        |

As shown in Table 1, the STRSCF model with HOG and CN features performs the better precision and success rate. Moreover, our STRSCF model is sensitive to the temporal regularization parameter, and it get a best OPE with $\mu=0.01$. Besides, the STRSCF tracker using CN and HOG features can runs at 26 FPS, which reaches the baseline of the real-time tracking.

3.3. General comparison on OTB-13

We compare the STRSCF tracker against 9 state-of-the-art trackers, including CSK [6], KCF [7], SAMF [12], Staple [9], DCFNet [10], CNN-SVM [3], Scale-DLSSVM [4], SCF [13] and SiamFC [16]. Among them, CSK, KCF, SAMF, Staple and DCFNet are CF-based trackers; CNN-SVM, Scale-DLSSVM and SCF are SVM-based trackers; SiamFC is Siamese networks-based tracker. In addition, the CNN-SVM and DCFNet trackers exploit the CNN representations of image, which can also be classified into CNN-based trackers.

As shown in Figure 2, the proposed method (STRSCF) gets a 0.628 success rate and a 0.876 precision on all 50 test sequences, which is the best OPE result among all the compared trackers.
4. Conclusion
In this paper, we propose an effective and efficient STRSCF method for real-time visual tracking. By the introduction of the spatial weight and temporal regularization term, we learn a robust appearance model using a developed optimization algorithm. In addition, we extend our STRSCF to the HOG and CN feature space for effective tracking. Experiment results show that our STRSCF not only reaches the baseline of the real-time tracking, but also outperforms several state-of-the-art trackers in accuracy.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (No. 61461028) and the National Natural Science Foundation of China (No. 61861027).

References
[1] Hare S, Saffari A, Torr P H S. Struck: Structured output tracking with kernels. Proc. International Conference on Computer Vision, 2011, pp. 263-270.
[2] Zhang J, Ma S, Sclaroff S. MEEM: Robust Tracking via Multiple Experts Using Entropy Minimization. Proc. European Conference on Computer Vision, 2014, pp. 188-203.
[3] Hong S, You T, Kwak S, et al. Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network. Proc. International Conference on Machine Learning, 2015, pp. 597-606.
[4] Ning J, Yang J, Jiang S, et al. Object Tracking via Dual Linear Structured SVM and Explicit Feature Map. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4266-4274.
[5] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2544-2550.
[6] Henriques J F, Caseiro R, Martins P, et al. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. Proc. European Conference on Computer Vision, 2012, pp. 702-715.
[7] Henriques J F, Caseiro R, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(3), pp. 583-596.
[8] Danelljan M, Häger G, Khan F S, et al. Learning Spatially Regularized Correlation Filters for Visual Tracking. Proc. IEEE International Conference on Computer Vision, 2015, pp. 4310-4318.
[9] Bertinetto L, Valmadre J, Golodetz S, et al. Staple: Complementary Learners for Real-Time Tracking. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1401-1409.
[10] Wang Q, Gao J, Xing J, et al. DCFNet: Discriminant Correlation Filters Network for Visual Tracking. arXiv:1704.04057. 2017.
[11] Babenko B, Yang M-H, Belongie S. Robust Object Tracking with Online Multiple Instance Learning. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2011, 33(8), pp. 1619-1632.
[12] Li Y, Zhu J. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration. Proc. European Conference on Computer Vision Workshops, 2015, pp. 254-265.
[13] Zuo W, Wu X, Lin L, et al. Learning Support Correlation Filters for Visual Tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, pp. 1-1.
[14] Zhang K, Zhang L, Liu Q, et al. Fast Visual Tracking via Dense Spatio-temporal Context Learning. Proc. European Conference on Computer Vision, 2014, pp. 127-141.
[15] Wu Y, Lim J, Yang M-H. Online Object Tracking: A Benchmark. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2411-2418.
[16] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-Convolutional Siamese Networks for Object Tracking. Proc. European Conference on Computer Vision Workshops, 2016, pp. 850-865.