Structural Correlation Filter for Robust Visual Tracking

Si Liu[1] Tianzhu Zhang[2] Xiaochun Cao[1] * Changsheng Xu[2]

[1] State Key Laboratory Of Information Security, Institute of Information Engineering, Chinese Academy of Sciences

[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Abstract

In this paper, we propose a novel structural correlation filter (SCF) model for robust visual tracking. The proposed SCF model takes part-based tracking strategies into account in a correlation filter tracker, and exploits circular shifts of all parts for their motion modeling to preserve target object structure. Compared with existing correlation filter trackers, our proposed tracker has several advantages: ***(1)*** *Due to the part strategy, the learned structural correlation filters are less sensitive to partial occlusion, and have computational efficiency and robustness.* ***(2)*** *The learned filters are able to not only distinguish the parts from the background as the traditional correlation filters, but also exploit the intrinsic relationship among local parts via spatial constraints to preserve object structure.* ***(3)*** *The learned correlation filters not only make most parts share similar motion, but also tolerate outlier parts that have different motion. Both qualitative and quantitative evaluations on challenging benchmark image sequences demonstrate that the proposed SCF tracking algorithm performs favorably against several state-of-the-art methods.*

1. Introduction

Visual tracking is one of the most fundamental problems in computer vision with various applications in video surveillance, human computer interaction and vehicle navigation. Although great progress has been made in recent years, it remains a challenging problem due to factors such as illumination changes, geometric deformations, partial occlusions, fast motions and background clutters.

Tracking algorithms can be generally categorized as either generative or discriminative methods. Generative trackers typically formulate tracking problem as searching for the best image regions which are similar to the tracked targets [25, 38, 17, 34]. Different from generative trackers, discriminative approaches cast tracking as a classification problem that distinguishes tracked targets from backgrounds [2, 3, 16, 11, 32]. Recently, correlation filter based discriminative tracking methods have been proven to be able to achieve fairly high speed and robust tracking performance [5, 7, 13, 33, 12, 8, 14, 22, 21, 19]. Conventionally, correlation filters are designed to produce correlation peaks for each interested target in the scene while yielding low responses to background, which are usually used to detect expected patterns. As proved by Convolution Theorem, the correlation in time domain corresponds to an element-wise multiplication in Fourier domain. Thus, the intrinsic idea of correlation filter is that the correlation can be calculated in Fourier domain in order to avoid the time-consuming convolution operation. Due to its computational efficiency, correlation filters have attracted considerable attention to visual tracking. Although achieved the appealing results both in accuracy and robustness, these correlation filter based trackers cannot deal with partial occlusion well. Figure 1 shows one example about the tracking results on the *lemming* sequence of two correlation filter based trackers, namely DSST [7] and KCF [13], which have achieved state-of-art results and have beaten all other attended track-

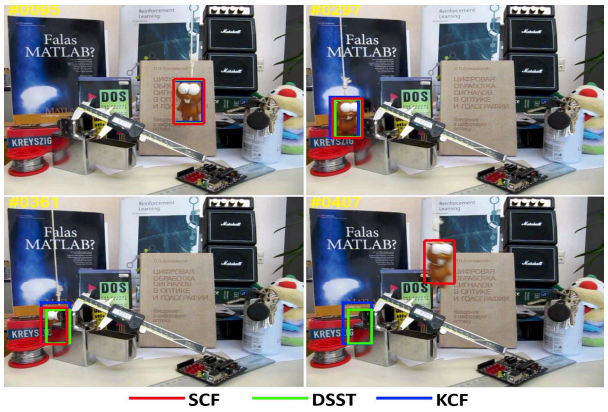

Figure 1. Comparisons of our approach with state-of-the-art correlation filter trackers in challenging situations of partial occlusion on the Lemming sequence [30]. Our SCF tracker takes part-based tracking strategy into account for translation estimation, and performs robustly to partial occlusion after the 361th frame than the DSST [7] and KCF [13] methods.

*Indicates corresponding author

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.467

(a) The KCF tracker (b) The SCF tracker

Figure 2. Comparisons of our approach with the KCF on the Jogging sequence [30] for part position estimation. The (m, n) is circular shift with the maximal value of the response map of each part, which exploits the motion information of each part.

ers in terms of accuracy in the VOT challenge [20]. However, these two trackers fail to track the target object when partial occlusion happens.

To deal with the above issues, [19, 21] have made successful attempts to apply part-based tracking strategy to correlation filter tracking. In general, part-based tracking strategy models object appearance based on multiple parts of target. Obviously, when target is partially occluded, remaining visible parts can still provide reliable cues for tracking. Therefore, this strategy can be helpful to gain robustness against partial occlusions. In [19, 21], object parts are independently tracked by the KCF tracker [13], and these trackers fail to exploit spatial constraints among object parts. As a result, as shown in Figure 2, object parts move independently and have different directions, which eventually leads the tracker to drift away. In fact, there is little change between two consecutive frames as the time interval is small [22], and most parts should have similar directions to preserve object structure. Moreover, in [31, 19, 21], experimental results have shown that the relationships among parts are effective.Therefore introducing structural constraints among parts in correlation filter is supposed to be advantageous.

Motivated by the above observations, we propose a novel Structural Correlation Filter (SCF) for object appearance modeling, which has the following advantages: **(1)** The proposed SCF appearance model has the advantages of both part based trackers and correlation filter trackers, such as, less sensitive to partial occlusion, computational efficiency and robustness. **(2)** The proposed SCF appearance model exploits spatial layout structure among object parts, which is ignored by all the previous correlation filter trackers [5, 7, 13, 33, 12, 8, 14, 22, 21, 19] to the best of our knowledge. Due to this advantage, our proposed model not only exploits the intrinsic relationship among object parts to learn their correlation filters jointly, but also preserves the spatial layout structure among object parts. **(3)** The proposed SCF appearance model is robust for outlier parts, which have different motion from most of other parts. As shown in Figure 2, the jointly learned correlation filters of all parts not only make most parts have similar motion, but also tolerate outlier parts that have different motion.

Based on the above structural appearance model, we propose a robust and efficient SCF tracking approach. In the proposed tracker, an object is made up of a set of parts, each with an associated correlation filter. We learn the parameters of correlation filters for all parts jointly. The learned correlation filters not only distinguish object part from background, but also exploit spatial constraints among parts to preserve object structure. During tracking, the correlation filter of each part has a response map, which can help predict the part state (position) by searching for the location of the maximal value of the map. Then, the target object location is estimated as a weighted average of translations of all parts. Here, the weight of each part is the maximum value of its response map. In the experimental results, we show that it is practical and robust to exploit the intrinsic relationship among parts to learn their correlation filters jointly by preserving object structure, and it helps locate target object more accurately and is less sensitive to partial occlusion. As a result, the incorporation of structural constraints leads to substantial performance improvements.

2. Related Work

Visual tracking has been extensively studied [26, 30, 24, 28]. In this section, we introduce the methods closely related to this work: correlation filter trackers, part based trackers, and the KCF tracker [13] in detail.

Correlation Filter Trackers: Correlation filters have attracted considerable attention recently to visual tracking due to its computational efficiency and robustness. Bolme et al. encode target appearance by learning an adaptive correlation filter [5]. Heriques et al. exploit the circulant structure of adjacent image patches [12], and is further improved using HOG features [13]. Danelljan et al. exploit adaptive color attributes in [8], and use adaptive multi-scale correlation filters to handle scale variations in [7]. Zhang et al. [33] incorporate context information into filter learning. Hong et al. [14] propose a biology-inspired framework with short-term processing and long-term processing. In [22], Ma et al. introduce online random fern classifier for long-term tracking. In [21, 19], part based strategy is used in correlation filter. Different from the existing correlation filter trackers, we propose a novel structural correlation filter to preserve object structure for object appearance modeling.

Part based Trackers: Instead of learning a holistic appearance model, various part-based tracking algorithms have been proposed to gain robustness against partial occlusions [18, 29, 37, 31, 21, 19, 10, 39]. The Frag tracker [1] models object appearance with histograms of local parts. Kwon et al. [18] represent a non-rigid target object by a number of local patches with color histograms. Ce-

hovin et al. [29] uses the global and local appearance based on object parts. Godec et al. [10] extend the Hough forest for online object tracking. Different from the existing part based trackers, we introduce part-based tracking strategy in correlation filter to model the relationships among parts and preserve object structure. Due to correlation filter, motion information of parts can be effectively exploited.

The KCF Tracker: The KCF tracker [13] achieves very impressive results on Tracking Benchmark [30]. The key idea is that many negative samples are employed to enhance the discriminative ability of the track-by-detector scheme while exploring the structure of circulant matrix for high efficiency. In the following, we briefly introduce the main idea. Readers may refer to [13] for more details.

The KCF tracker models object appearance using a correlation filter $\mathbf{w}$ trained on an image patch $\mathbf{x}$ of $M \times N$ pixels, where all the circular shifts of $\mathbf{x}_{m,n}$, $(m,n) \in \{0,1,\ldots,M-1\} \times \{0,1,\ldots,N-1\}$, are generated as training samples with Gaussian function label $\mathbf{y}_{m,n}$. The goal is to find the optimal weights $\mathbf{w}$ in (1).

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{m,n} |\langle \phi(\mathbf{x}_{m,n}), \mathbf{w} \rangle - \mathbf{y}_{m,n}|^2 + \lambda \|\mathbf{w}\|^2. \quad (1)$$

Here, ϕ denotes the mapping to a kernel space and λ is a regularization parameter. Using the fast Fourier transformation (FFT) to compute the correlation, the objective function (1) is minimized as $\mathbf{w} = \sum_{m,n} \alpha(m,n)\phi(\mathbf{x}_{m,n})$, and the coefficient α is calculated as in (2).

$$\alpha = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\mathbf{y})}{\mathcal{F}(\langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle) + \lambda}\right) \quad (2)$$

Where $\mathbf{y} = \{\mathbf{y}(m,n)\}$, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and its inverse.Given the learned α and target appearance model $\bar{\mathbf{x}}$, the tracking task is carried out on an image patch $\mathbf{z}$ in the new frame with the search window size $M \times N$ by computing the response map as in (3). Here, $\odot$ is the Hadamard product. Then, the target position is detected by searching for the location of the maximal value of $\bar{\mathbf{y}}$.

$$\bar{\mathbf{y}} = \mathcal{F}^{-1}(\mathcal{F}(\alpha) \odot \mathcal{F}(\langle \phi(\mathbf{z}), \phi(\bar{\mathbf{x}}) \rangle)). \quad (3)$$

3. Structural Correlation Filter Tracking

In this section, we give a detailed description of our structural correlation filter based tracking method that makes use of the structural correlation filter model to learn correlation filters of all parts jointly to preserve target object structure. Next, we will sequentially introduce the structural correlation filter and SCF tracker.

3.1. Structural Correlation Filter Model

The objective function of the KCF tracker (1) can equivalently be expressed in its dual form (4).

$$\min_{\alpha} \frac{1}{4\lambda}\alpha^\top \mathbf{X}\mathbf{X}^\top \alpha + \frac{1}{4}\alpha^\top \alpha - \alpha^\top \mathbf{y} \quad (4)$$

Here, the vector α contains $M \times N$ dual optimization variables $\alpha_{m,n}$, and $\mathbf{X} = [\mathbf{x}_{0,0}, \ldots, \mathbf{x}_{m,n}, \ldots \mathbf{x}_{M-1,N-1}]^\top$. The two solutions are related by $\mathbf{w} = \frac{\mathbf{X}^\top \alpha}{2\lambda}$.

The KCF tracker (4) is to learn a holistic appearance model, which is not robust for partial occlusion. To deal with this issue, we apply part-based tracking strategy to the correlation filter. Given a target object, its K parts with $M \times N$ pixels can be sampled. Then, our goal is to learn K optimal weights $\mathbf{w}_k$ or α_k via (5).

$$\min_{\{\mathbf{u}_k\}_{k=1}^K} \sum_{k=1}^{K} \frac{1}{4\lambda}\mathbf{u}_k^\top \mathbf{G}_k \mathbf{u}_k + \frac{1}{4}\mathbf{u}_k^\top \mathbf{u}_k - \mathbf{u}_k^\top \mathbf{y} \quad (5)$$

Here, for clarity, we adopt $\mathbf{u}_k$ to denote the dual optimization variables, and $\mathbf{G}_k = \mathbf{X}_k \mathbf{X_k}^\top$. The $\mathbf{X_k}$ is all training samples of the k-th part, where $k = 1, \ldots, K$.

Note that, the basic idea of (4) is to select discriminative training samples $\mathbf{x}_{m,n}$ via $\alpha_{m,n}$ to distinguish the target object from the background. Here, the training samples $\mathbf{x}_{m,n}$, $(m,n) \in \{0,1,\ldots,M-1\} \times \{0,1,\ldots,N-1\}$ are the all possible circular shifts, which represent the possible motion of the target object. Therefore, selecting training samples $\mathbf{x}_{m,n}$ via $\alpha_{m,n}$ can predict the motion or state of target object. Ideally, the individual parts should stay close to each other to cover the entire target. As shown in Figure 2, most parts of target object move in the same way between two consecutive frames. Therefore, they should select the similar circular shifts to make them have similar motion. Considering that some parts may have different motions, and they may select different discriminative training samples. Based on the above observation, it is clear that most parts have similar $\mathbf{u}_k$ to make them move in the same direction to preserve target object structure, and some parts may have separate motion. Therefore, in (5), we assume all $\mathbf{u}_k$ can be written as $\mathbf{u}_k = \mathbf{u}_0 + \mathbf{v}_k$, where the vectors $\mathbf{v}_k$ are small when the selected circular shifts of all parts are similar to each other. That is to say, $\mathbf{u}_0$ carries the information of the commonality, and $\mathbf{v}_p$ carries the information of the specialty (outlier) and should be sparse.

$$\begin{aligned} \min_{\{\mathbf{u}_k\}_{k=1}^K} & \sum_{k=1}^{K} \tfrac{1}{4\lambda}\mathbf{u}_k^\top \mathbf{G}_k \mathbf{u}_k + \tfrac{1}{4}\mathbf{u}_k \top \mathbf{u}_k - \mathbf{u}_k^\top \mathbf{y} + \gamma \|\mathbf{v}_k\|_1 \\ & \text{s.t.} \quad \mathbf{u}_k = \mathbf{v}_k + \mathbf{u}_0, \quad k = 1, \ldots, K \end{aligned} \quad (6)$$

Motivated by the above points, the part based correlation filters (5) can be reformulated as structural correlation filter model (6), which can learn the correlation filters of all parts jointly, and distinguish the parts from the background. Moreover, the SCF model is less insensitive to partial occlusion, but has computational efficiency and robustness.

3.2. The Proposed SCF Tracker

Based on the structural correlation filter model, we propose a novel SCF tracker with several important modules,

including model updating, target state estimation, kernel selection, and feature representation.

Model Updating: In tracking, object appearance will change because of a number of factors such as illumination and pose changes. Hence it is necessary to update part classifiers over time. In the proposed tracker, the model consists of the learned target appearance $\bar{\mathbf{x}}_k$ and the transformed classifier coefficients $\mathbf{u}_k$. Moreover, different parts of targets may suffer from different appearance changes, illumination variation or partial occlusion. If we simply combine all parts with the same weight, their correlations filters may be unfairly emphasized. Therefore, for each part, we have its weight π_k to emphasize its importance. For each part, its model parameters at time t are updated as in (7).

$$\begin{aligned} \mathcal{F}(\mathbf{u}_k)^t &= (1-\eta)\mathcal{F}(\mathbf{u}_k)^{t-1} + \eta\mathcal{F}(\mathbf{u}_k) \\ \mathcal{F}(\bar{\mathbf{x}}_k)^t &= (1-\eta)\mathcal{F}(\bar{\mathbf{x}}_k)^{t-1} + \eta\mathcal{F}(\mathbf{x}_k) \\ \pi_k^t &= (1-\eta)\pi_k^t + \eta\pi_k \end{aligned} \tag{7}$$

Where η is a learning rate parameter. The $\mathbf{u}_k$ is computed by simple linear interpolation, and the $\bar{\mathbf{x}}_k$ is updated by taking the current appearance into account. The π_k is the maximal value of the response map of the k-th part.

Target State Estimation: The target state estimation includes position prediction and scale decision. *(1) Position Estimation.* Given the learned model $\bar{\mathbf{x}}_k$, $\mathbf{u}_k$ of part k, its new position is detected by searching for the location of the maximal value of $\bar{\mathbf{y}}_k$ as in (3). Then, we can obtain its translation $\mathbf{s}_k$. For simplicity, the translation of the target object is calculated as $s = \sum_k \pi_k s_k$, which shows more robust tracking parts with larger detection scores have higher effect on the target position estimation. *(2) Scale Handling.* To handle scale variation, windows with different sizes are sampled around the target, and are correlated with the learned filter. Subsequently, the window with the highest correlation score can be predicted as the new state. This searching strategy is also used in the DSST [7].

Kernel Selection: Inspired by the effectiveness of the Gaussian kernel in the existing correlation filter trackers, the $\mathbf{G}_k$ is computed with the same kernel $\kappa(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\frac{|\mathbf{x}_1 - \mathbf{x}_2|^2}{\delta^2})$, which is defined as $\kappa(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle$ with a mapping ϕ. We compute the full kernel correlation for each part in (5) and (6) efficiently in the Fourier domain. The details are discussed in Section 4.

Feature Representation: Similar to [13], we use HOG features with 31 bins. However, our tracker is quite generic and any dense feature representation with arbitrary dimensions can be incorporated.

4. Optimization

In this section, we present how to solve the optimization problem (6) using the fast first order Alternating Direction Method of Multipliers (ADMM) [6] approach. By introducing augmented Lagrange multipliers to incorporate the equality constraints into the objective function, we obtain the Lagrangian function in (8) that can be optimized through a sequence of simple closed form update operations in (9) where θ_k and $\beta_k > 0$ are Lagrange multipliers and penalty parameters, respectively.

$$\begin{aligned} &L(\{\mathbf{u}_k, \mathbf{v}_k, \theta_k, \beta_k\}_{k=1}^K, \mathbf{u}_0) \\ &= \sum_{k=1}^K \frac{1}{4\lambda}\mathbf{u}_k^\top \mathbf{G}_k \mathbf{u}_k + \frac{1}{4}\mathbf{u}_k^\top \mathbf{u}_k - \mathbf{u}_k^\top y + \gamma\|\mathbf{v}_k\|_1 \\ &+ \theta_k^\top(\mathbf{u}_k - \mathbf{v}_k - \mathbf{u}_0) + \frac{\beta_k}{2}\|\mathbf{u}_k - \mathbf{v}_k - \mathbf{u}_0\|^2 \end{aligned} \tag{8}$$

$$\Rightarrow \min_{\{\mathbf{u}_k, \mathbf{v}_k, \theta_k, \beta_k\}_{k=1}^K, \mathbf{u}_0} L(\{\mathbf{u}_k, \mathbf{v}_k, \theta_k, \beta_k\}_{k=1}^K, \mathbf{u}_0) \tag{9}$$

The ADMM method iteratively updates one of the variables $\mathbf{u}_0$, $\{\mathbf{v}_k\}_{k=1}^K$, $\{\mathbf{u}_k\}_{k=1}^K$, and the Lagrange multiplier $\{\theta_k\}_{k=1}^K$ by minimizing (9), while keeping the others fixed to their most recent values. By updating these variables iteratively, the convergence can be guaranteed [6]. Consequently, we have four update steps corresponding to all the variables with closed form solutions as follows.

Step 1: Update $\mathbf{u}_0$ (with others fixed): The $\mathbf{u}_0$ is updated by solving the optimization problem (10) with the closed form solution (11).

$$\mathbf{u}_0 = \arg\min_{\mathbf{u}_0} \sum_{k=1}^K -\theta_k^T \mathbf{u}_0 + \frac{\beta_k}{2}\|\mathbf{u}_k - \mathbf{v}_k - \mathbf{u}_0\|^2 \tag{10}$$

$$\Rightarrow \mathbf{u}_0 = \frac{1}{K}\sum_{k=1}^K \mathbf{u}_k - \mathbf{v}_k + \frac{1}{\beta_k}\theta_k \tag{11}$$

Step 2: Update $\mathbf{v}_k$ (with others fixed): The minimization problem (8) with respect to $\{\mathbf{v}_k\}_{k=1}^K$ is decomposed into K independent subproblems. The k-th subproblem to update $\mathbf{v}_k$ can be equivalently rewritten as (12).

$$\begin{aligned} \mathbf{v}_k = \arg\min_{\mathbf{v}_k} &\gamma\|\mathbf{v}_k\|_1 + \theta_k^\top(\mathbf{u}_k - \mathbf{v}_k - \mathbf{u}_0) \\ &+ \frac{\beta_k}{2}\|\mathbf{u}_k - \mathbf{v}_k - \mathbf{u}_0\|^2 \end{aligned} \tag{12}$$

The solution of (12) can be obtained by rearranging it into the optimization problem (13) with the closed form solution (14).

$$\mathbf{v}_k = \arg\min_{\mathbf{v}_k} \frac{\gamma}{\beta_k}\|\mathbf{v}_k\|_1 + \frac{1}{2}\left\|\mathbf{v}_k - (\mathbf{u}_k + \frac{1}{\beta_k}\theta_k - \mathbf{u}_0)\right\|^2 \tag{13}$$

$$\Rightarrow \mathbf{v}_k = \mathcal{S}_{\frac{\gamma}{\beta_k}}(\mathbf{u}_k + \frac{1}{\beta_k}\theta_k - \mathbf{u}_0) \tag{14}$$

Here, $\mathcal{S}_\lambda(\mathbf{x}_i) = sign(\mathbf{x}_i)\max(0, |\mathbf{x}_i| - \lambda)$ is the soft-thresholding operator for a vector $\mathbf{x}$.

Algorithm 1: The optimization for (9) via ADMM.

Input : Training Data: $\mathbf{G}_k$ and $\mathbf{y}$. Initialization of λ, γ, $\mathbf{u}_k = 0$, $\mathbf{u}_0 = 0$, $\mathbf{v}_k = 0$, $\theta = 0$, and $\beta > 0$.

Output: Correlation filters $\mathbf{u}_k$, $k = 1, \ldots, K$.

```
while not converged do
    Update u_0 via (11);
    for k = 1 to K do
        Update v_k via (14);
        Update u_k as in (16);
        Update θ_k as in (17);
    end
end
```

Step 3: Update $\mathbf{u}_k$ (with others fixed): The minimization problem (8) with respect to $\{\mathbf{u}_k\}_{k=1}^K$ is decomposed into K independent subproblems. The k-th subproblem to update $\mathbf{u}_k$ can be equivalently rewritten as (15).

$$\mathbf{u}_k = \arg\min_{\mathbf{u}_k} \frac{1}{4\lambda}\mathbf{u}_k^\top \mathbf{G}_k \mathbf{u}_k + \frac{1}{4}\mathbf{u}_k^\top \mathbf{u}_k - \mathbf{u}_k^\top \mathbf{y} + \theta_k^\top(\mathbf{u}_k - \mathbf{v}_k - \mathbf{u}_0) + \frac{\beta_k}{2}\|\mathbf{u}_k - \mathbf{v}_k - \mathbf{u}_0\|^2 \quad (15)$$

Then, for each $\mathbf{u}_k$, it is updated by solving the optimization problem (15) with the closed form solution (16).

$$\mathbf{u}_k = (\frac{1}{2\lambda}\mathbf{G}_k + \frac{1}{2}\mathbf{I} + \beta_k\mathbf{I})^{-1}(\mathbf{y} - \theta_k + \beta_k\mathbf{v}_k + \beta_k\mathbf{u}_0) \quad (16)$$

Here, $\mathbf{I}$ is a $MN \times MN$ identity matrix.

Step 4: Update Multiplier θ_k: The Lagrange multipliers are updated as in (17), where $\rho > 1$,

$$\theta_k = \theta_k + \beta_k(\mathbf{u}_k - \mathbf{v}_k - \mathbf{u}_0); \quad \beta_k = \rho\beta_k \quad (17)$$

The ADMM algorithm that solves (9) is shown in Algorithm 1, where convergence is reached when the change in the objective function or solution $\mathbf{u}_k$ is below a pre-defined threshold (e.g., $\tau = 10^{-3}$ in this work). In addition, we set $\beta_1 = \cdots \beta_k = \cdots = \beta_K = \beta$. Here, we note that other penalty update rules and stopping criteria can be used for this optimization problem as discussed in [6]. As shown in Algorithm 1, the major computation cost is the fifth step to update $\mathbf{u}_k$ with matrix inverse and multiplication in spatial domain. However, it can be calculated very efficiently in the Fourer domain by considering the circulant structure property of $\mathbf{X}_k$. Assume $\mathbf{x}$ is the base sample of $\mathbf{X}_k$, the $\mathbf{u}_k$ can be updated with only the base sample as (18).

$$\hat{\mathbf{u}}_k = \frac{\hat{\mathbf{y}} - \hat{\theta}_k + \beta_k\hat{\mathbf{v}}_k + \beta_k\hat{\mathbf{u}}_0}{\frac{1}{2\lambda}\hat{\mathbf{x}}^* \odot \hat{\mathbf{x}} + \frac{1}{2} + \beta_k} \quad (18)$$

Here, the fraction denotes element-wise division, $\mathbf{x}^*$ is the complex-conjugate of $\mathbf{x}$, $\hat{\mathbf{x}}$ denotes the Discrete Fourier Transform (DFT) of the generating vector $\hat{\mathbf{x}} = \mathcal{F}(\mathbf{x})$, and $\odot$ denotes the element-wise product. Finally, the $\mathbf{u}_k$ can be obtained via $\mathbf{u}_k = \mathcal{F}^{-1}(\hat{\mathbf{u}}_k)$. Moreover, to make the SCF tracker faster, the Algorithm 1 can be implemented in matrix form without the for loop.

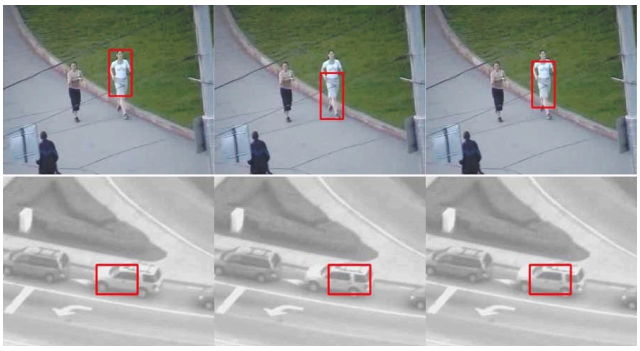

Figure 3. The generated 3 parts based on the target's ratio.

5. Experimental Results

We first introduce experimental setup including parameters, datasets, and evaluation metrics. Then, we provide both quantitative and qualitative comparisons with state-of-the-art trackers.

5.1. Experimental Setup

Parameters: The γ in (6) is set to 0.01. All the other parameters are set to the same values as the KCF tracker. To generate the parts, we use the spatial layout as shown in Figure 3 to sample 3 parts based on the target's height-width ratio. Note that, any other part sampling methods can also be adopted. We use the same parameter values and initialization for all the sequences. All the parameter settings are available in the source code to be released for accessible reproducible research.

Datasets and Evaluation Metrics: We evaluate the proposed method on a large benchmark dataset [30] that contains 50 videos with comparisons to state-of-the-art methods. The performance of our approach is quantitatively validated by three metrics used in [30] including distance precision (DP), centre location error (CLE) and overlap precision (OP). The DP is computed as the relative number of frames in the sequence where the centre location error is smaller than a certain threshold. As in [30], the DP values at a threshold of 20 pixels are reported. The CLE is computed as the average Euclidean distance between the ground-truth and the estimated centre location of the target. The OP is defined as the percentage of frames where the bounding box overlap surpasses a threshold. We report the results at a threshold of 0.5, which correspond to the PASCAL evaluation criteria. We provide results using the average DP, CLE and OP over all 50 sequences. In addition, we plot the

Table 1. Comparison with state-of-the-art trackers on the 50 benchmark sequences. Our approach performs favorably against existing methods in overlap precision (OP) (%) at an overlap threshold 0.5, distance precision (DP) (%) at a threshold of 20 pixels and centre location error (CLE) (in pixels). The top rank 3 values are highlighted by bold and different colors: red, blue, and green, respectively.

Metrics	**SCF**	TLD [16]	Struck [11]	CSK [12]	VTD [17]	KCF [13]	L1APG [4]	LOT [23]	DFT [27]	MEEM [32]	TGPR [9]	RPT [19]	MUSTer [14]	DSST [7]	SCM [41]	MIL [3]	ASLA [15]
OP	**79.7**	52.1	55.9	44.3	49.3	62.3	44.0	41.3	44.4	69.8	65.1	**70.7**	**78.4**	66.7	61.6	37.3	51.1
DP	**86.6**	60.8	65.6	54.5	57.6	74.0	48.5	52.2	49.6	**83.0**	71.4	81.9	**86.5**	73.7	64.9	47.5	53.2
CLE	**22.5**	48.1	50.6	88.8	47.4	35.5	77.4	58.2	69.3	**21.4**	47.2	35.9	**17.3**	41.3	54.1	62.3	71.1

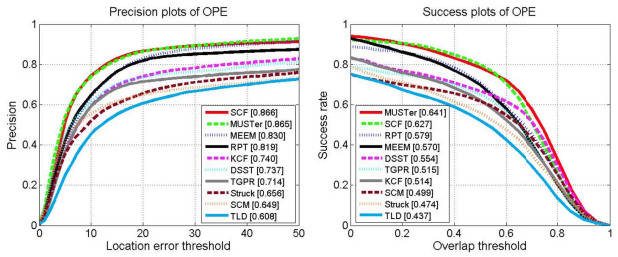

Figure 4. Precision and success plots over all the 50 sequences using one-pass evaluation (OPE). The legend contains the area-under-the-curve score for each tracker. The proposed SCF method performs favorably against the state-of-the-art trackers.

precision and success plots as in [30].

5.2. Comparison with State-of-the-Art

We evaluate the proposed tracker on the benchmark with comparisons to 34 trackers including 29 trackers in [30] including SCM [41], MTT [35, 36], and TLD [16], and other 5 recently published state-of-the-art trackers with their shared source code: MEEM [32], TGPR [9], RPT [19], MUSTer [14], DSST [7]. The details of the 29 trackers in the benchmark can be found in [30]. We present the results using average OP, DP and CLE over all sequences in Table 1, and report the results in one-pass evaluation (OPE) using the distance precision and overlap success rate in Figure 4 and attribute-based evaluation in Figure 5.

Table 1 shows that our algorithm performs favorably against state-of-the-art methods. Among the trackers in the literature, the MUSTer method achieves the best results with an average OP of 78.4%, DP of 86.5%, and CLE of 17.3 pixels. Our algorithm performs well with OP of 79.7%, DP of 86.6%, and CLE of 22.5 pixels. These results show the proposed SCF tracker achieves slightly better tracking performance than the MUSTer. Note that, the proposed SCF tracker can be improved more by considering other tracking strategy, such as, long-term strategy, and keypoint matching strategy in the MUSTer tracker [14]. Overall, our SCF tracker achieves significantly improvement than other existing trackers. The details are as follows. *(1)* MEEM and RPT are top 2 existing methods with average OP of 69.8% and 70.7% respectively. Our approach achieves better tracking performance by 9.8% and 9%. *(2)* The proposed SCF method performs well against the MUSTer (by 0.1%), MEEM (by 3.6%), and RPT (by 4.7%) methods in terms of average DP. *(3)* Among the other existing trackers, MEEM provides the best results with an average CLE of 21.4 pixels. Our approach achieves comparable results with an average CLE of 22.5 pixels. *(4)* Compared with the correlation filter trackers, the proposed SCF method performs well against the KCF (by 17.4%) and DSST (by 13%) methods in terms of average OP, and achieves performance gain of 12.6% and 12.9% in term of average DP. In term of average CLE, the proposed SCF method has about 13.0 pixels and 18.8 pixels improvement.

Figure 4 contains the precision and success plots illustrating the mean distance and overlap precision over all the 50 sequences. In both precision and success plots, our approach shows comparable results as the MUSTer and significantly outperforms the best existing correlation filter methods (DSST and KCF). Note that, when overlap threshold is from 0.2 to 0.6, the proposed SCF method achieves slightly better than the MUSTer in success plots of OPE. In summary, the precision plot demonstrates that our approach performs well against the existing methods (KCF, MEEM, TGPR, SCM, Struck).In Figure 5, We analyze the tracking performance based on attributes of image sequences [30], which annotates 11 attributes to describe the different challenges in the tracking problem, e.g., occlusions or out-of-view. These attributes are useful for analyzing the performance of trackers in different aspects. Due to space constraints, we present the success and precision plots of OPE for 4 attributes in Figure 5 and more results can be found in the supplementary material. For presentation clarity, we present the top 10 performing methods in each plot. We note that the proposed tracking method performs well in dealing with challenging factors including deformation, occlusion, out-of-plane rotation, and out of view.

5.3. Qualitative Comparison

We compare our algorithm with the top 9 existing trackers in our evaluation (MUSTer [14], RPT [19], MEEM [32], DSST [7], KCF [13], TGPR [9], SCM [41], Struck [11], and TLD [16]) on 10 challenging sequences in Figure 6. Overall, these trackers perform well, but the existing trackers have the following issues: The MUSTer drifts when fast motion happens (*couple*). The RPT does not perform well in scale variation (*singer2*, *walking2*, *lemming*, and *tiger1*), fast motion (*couple*), and partial occlusion (*jogging-2*). The MEEM cannot handle partial occlusion well (*suv*, *walking2*,

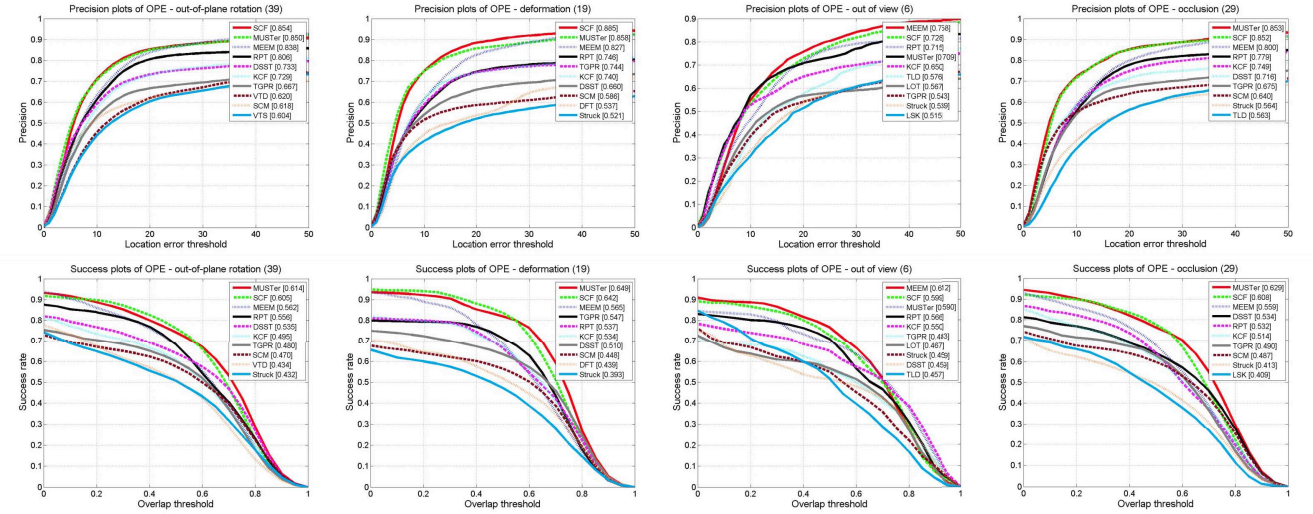

Figure 5. The precision and success plots of OPE over four tracking challenges of out-of-plane rotation, deformation, out-of-view, and occlusion. The legend contains the AUC score for each tracker. Our SCF method performs favorably against the state-of-the-art trackers.

and *jogging-2*). The KCF, DSST, and Struck methods drift when target objects undergo heavy occlusion (*jogging-2*) and fast motion (*couple*). The SCM and TLD methods do not follow targets undergoing significant deformation and fast motion (*tiger1* and *lemming*) well. The TGPR does not perform well in fast motion (*couple*) and partial occlusion (*suv*). Overall, the proposed SCF tracker performs well in tracking objects on these challenging sequences. In addition, we compare the center location error frame-by-frame on the 10 sequences in Figure 7, which shows that our method performs well against existing trackers. More results can be found in the supplementary material.

6. Conclusion

In this paper, we propose a novel structural correlation filter namely SCF to model target appearance for robust visual tracking. The proposed SCF model fuses part-based tracking strategy into correlation filter tracker, and exploits circular shifts of all parts for their motion modeling to preserve target object structure. As a result, it not only has the advantages of existing correlation filter trackers, such as, computational efficiency and robustness, but also can be less sensitive to partial occlusion, preserve object structure, and enable the capture of outlier parts to have different motion. Both qualitative and quantitative evaluations on challenging benchmark image sequences demonstrate that the proposed SCF tracking algorithm performs favorably against several state-of-the-art methods. In the future, we will evaluate the proposed SCF model on more datasets and make use of co-learning algorithm [40] to obtain more improvement.

Acknowledgment

This work is supported by National Natural Science Foundation of China (No.61572493, 61225009, 61432019, 61303173, 61572498, 61532009, 61572296, Grant U1536203), 100 Talents Programme of The Chinese Academy of Sciences, National Basic Research Program of China (973 Program No. 2012CB316304), and "Strategic Priority Research Program" of the Chinese Academy of Sciences (XDA06010701).

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *CVPR*, pages 798–805, 2006. 2

[2] S. Avidan. Ensemble tracking. In *CVPR*, pages 494–501, 2005. 1

[3] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *CVPR*, 2009. 1, 6

[4] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l_1 tracker using accelerated proximal gradient approach. In *CVPR*, 2012. 6

[5] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, pages 2544–2550, 2010. 1, 2

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 4, 5

[7] M. Danelljan, G. Hger, F. S. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *BMVC*, 2014. 1, 2, 4, 6

[8] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer. Adaptive color attributes for real-time visual tracking. In *CVPR*, pages 1090–1097, 2014. 1, 2

Figure 6. Tracking results of the top 10 trackers (denoted in different colors and lines) in our evaluation on 10 challenging sequences (from left to right and top to down are shaking, singer2, jumping, suv, lemming, skating1, tiger1, couple, walking2, and jogging-2, respectively).

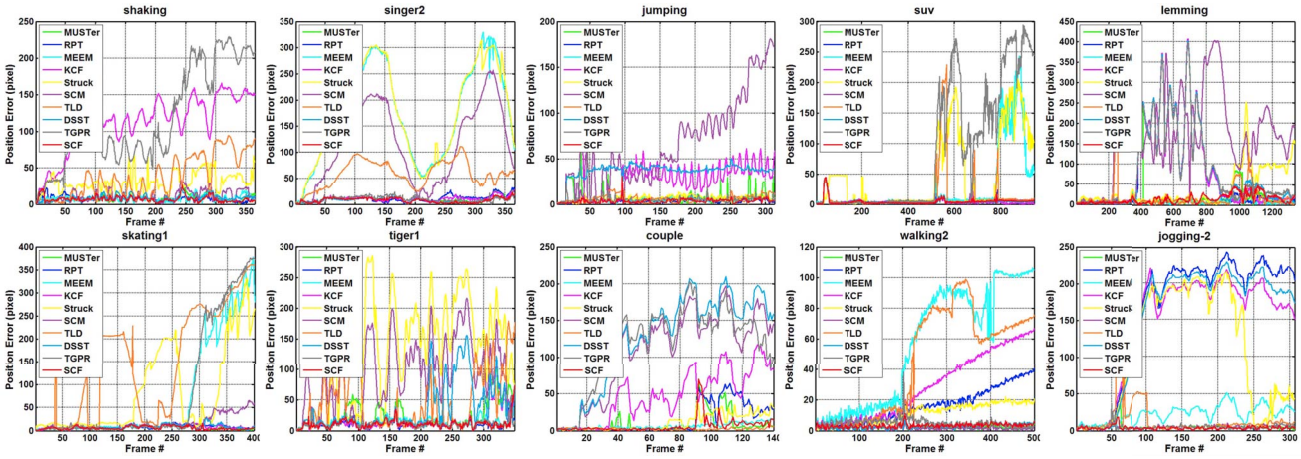

Figure 7. Comparison of center location errors (in pixels) on 10 challenging sequences. Generally, our method achieves better.

[9] J. Gao, H. Ling, W. Hu, and J. Xing. Transfer learning based visual tracking with gaussian process regression. In *ECCV*, 2014. 6

[10] M. Godec, P. M. Roth, and H. Bischof. Hough-based tracking of non-rigid objects. In *ICCV*, 2011. 2, 3

[11] S. Hare, A. Saffari, and P. Torr. Struck: Structured output tracking with kernels. In *ICCV*, 2011. 1, 6

[12] J. Henriques, R. Caseiro, P. Martins, and J. Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*, 2012. 1, 2, 6

[13] J. F. Henriques, R. Caseiro, P. M. 0004, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(3):583–596, 2015. 1, 2, 3, 4, 6

[14] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In *CVPR*, pages 749–758, 2015. 1, 2, 6

[15] X. Jia, H. Lu, and M.-H. Yang. Visual tracking via adaptive structural local sparse appearance model. In *CVPR*, 2012. 6

[16] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1409–1422, 2012. 1, 6

[17] J. Kwon and K. M. Lee. Visual tracking decomposition. In *CVPR*, 2010. 1, 6

[18] J. Kwon and K. M. Lee. Highly non-rigid object tracking via patch-based dynamic appearance modeling. *PAMI*, 35(10):2427–2441, 2013. 2

[19] Y. Li, J. Zhu, and S. C. H. Hoi. Reliable patch trackers: Robust visual tracking by exploiting reliable patches. In *CVPR*, pages 353–361, 2015. 1, 2, 6

[20] F. LIRIS. The visual object tracking vot2014 challenge results. 2014. 2

[21] T. Liu, G. Wang, and Q. Yang. Real-time part-based visual tracking via adaptive correlation filters. In *CVPR*, pages 4902–4912, 2015. 1, 2

[22] C. Ma, X. Yang, C. Zhang, and M.-H. Yang. Long-term correlation tracking. In *CVPR*, pages 5388–5396, 2015. 1, 2

[23] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan. Locally orderless tracking. In *CVPR*, pages 1940–1947, 2012. 6

[24] Y. Pang and H. Ling. Finding the best from the second bests - inhibiting subjective bias in evaluation of visual tracking algorithms. In *ICCV*, 2013. 2

[25] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental Learning for Robust Visual Tracking. *IJCV*, 77(1):125–141, 2008. 1

[26] S. Salti, A. Cavallaro, and L. D. Stefano. Adaptive appearance modeling for video tracking: Survey and evaluation. *TIP*, 21(10):4334–4348, 2012. 2

[27] L. Sevilla-Lara and E. Learned-Miller. Distribution fields for tracking. In *CVPR*, pages 1910–1917, 2012. 6

[28] A. Smeulder, D. Chu, R. Cucchiara, S. Calderara, A. Deghan, and M. Shah. Visual tracking: an experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1442–1468, 2013. 2

[29] L. Čehovin, M. Kristan, and A. Leonardis. Robust visual tracking using an adaptive coupled-layer visual model. *PAMI*, 35(4):941–953, 2013. 2, 3

[30] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 1, 2, 3, 5, 6

[31] R. Yao, Q. Shi, C. Shen, Y. Zhang, and A. van den Hengel. Part-based visual tracking with online latent structural learning. In *CVPR*, 2013. 2

[32] J. Zhang, S. Ma, and S. Sclaroff. MEEM: Robust tracking via multiple experts using entropy minimization. In *ECCV*, 2014. 1, 6

[33] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang. Fast visual tracking via dense spatio-temporal context learning. In *ECCV*, volume 8693, pages 127–141, 2014. 1, 2

[34] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Low-rank sparse learning for robust visual tracking. In *ECCV*, 2012. 1

[35] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via multi-task sparse learning. In *CVPR*, 2012. 6

[36] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via structured multi-task sparse learning. *International Journal of Computer Vision*, 101(2):367–383, 2013. 6

[37] T. Zhang, C. Jia, C. Xu, Y. Ma, and N. Ahuja. Partial occlusion handling for visual tracking via robust part matching. In *CVPR*, 2014. 2

[38] T. Zhang, S. Liu, N. Ahuja, M.-H. Yang, and B. Ghanem. Robust Visual Tracking via Consistent Low-Rank Sparse Learning. *International Journal of Computer Vision*, 111(2):171–190, 2015. 1

[39] T. Zhang, S. Liu, C. Xu, S. Yan, B. Ghanem, N. Ahuja, and M.-H. Yang. Structural sparse tracking. In *CVPR*, 2015. 2

[40] T. Zhang and C. Xu. Cross-domain multi-event tracking via co-pmht. *ACM Trans. Multimedia Comput. Commun. Appl.*, 10(4):31:1–31:19, July 2014. 7

[41] W. Zhong, H. Lu, and M.-H. Yang. Robust object tracking via sparsity-based collaborative model. In *CVPR*, pages 1838–1845, 2012. 6