IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 25, 2019, accepted April 18, 2019, date of publication April 22, 2019, date of current version May 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2912527

Learning an Orientation and Scale Adaptive Tracker With Regularized Correlation Filters

KE TAN AND ZHENZHONG WEI
Key Laboratory of Precision Opto-mechatronics Technology, Ministry of Education, Beihang University, Beijing 100083, China

Corresponding author: Zhenzhong Wei (zhenzhongwei@buaa.edu.cn)

This work was supported in part by the National Science Fund for Distinguished Young Scholars of China under Grant 51625501, and in part by the Co-Construction Project of Scientific Research and Postgraduate Cultivation -Achievements Transformation and Industrialization Program-On-line dynamic detection system of train pantograph-catenary operating conditions provided by the Beijing Municipal Education Commission.

ABSTRACT Spatial and temporal constraints are very important for correlation filter (CF)-based trackers. However, the existing methods usually fail to regularize the CF learning by the spatio-temporal information because of the ineffective target representation. To address the issue, we propose a novel Orientation and Scale adaptive tracker with Regularized Correlation Filters (OSRCF) for visual tracking. First, we directly employ the target region to build a spatio-temporally regularized CF, which is solved efficiently by the alternating direction method of multipliers (ADMM). Especially, the spatial regularization module can alleviate boundary effects by suppressing the outside background, while the temporal one can handle rapid appearance changes by smoothing the CF updating. Second, we obtain a non-axis-aligned bounding box for the target region representation by orientation and scale adaptive strategy, which cooperates a straightforward orientation estimation with a discriminative scale space correlation filter. We perform comprehensive experiments on two recent visual tracking benchmark datasets: VOT2017 and OTB2015. The results show that our OSRCF tracker outperforms top-ranked methods with handcrafted features in VOT2017 and achieves outstanding performance compared to some state-of-the-art methods in OTB2015.

INDEX TERMS Visual tracking, orientation and scale adaption, correlation filters.

I. INTRODUCTION

Visual tracking is a fundamental research topic in computer vision due to its numerous applications in areas such as robotics, surveillance, vehicle navigation and human computer interactions. Although much work has been done over the past decades see [1]–[4] to cite a few, it is still a challenging problem to design an all-situation-handled tracker that can handle various critical situations, such as illumination changes, geometric deformations, partial occlusions, fast motions and background clutters.

Existing tracking methods can be classified roughly into two categories: generative methods [5]–[15] and discriminative methods [16]–[26]. Generative methods learn the representation of an object, often a set of basis vectors from a subspace or a series of templates, to search for the region which is the most similar to the tracked object. Discriminative methods instead learn a classifier to distinguish the tracked object from environment by machine learning techniques, such as Support Vector Machines (SVM), boosting techniques and neural networks.

Recently, Discriminative Correlation Filter (DCF) based methods have achieved great success in modern object tracking benchmarks [27], [28]. The standard DCF based trackers can utilize all spatial shifts of training and testing samples by exploiting the Fast Fourier Transform (FFT) at both learning and detection stages, which brings ultrahigh speed. However, this leads to unwanted boundary effects because of the periodic assumption of the samples. To solve the problem, Galoogahi et al. [29] propose a zero-padding correlation filter which reduce the number of unrealistic examples. Then Danneljan et al. [30] propose a spatially regularized correlation filter by penalizing filter values outside the object boundaries. However, these trackers suffer from the inaccurate target region because the target shape is approximated by an axis-aligned rectangle. To solve the problem, Lukezic et al. [31] exploit a color model to segment the target from the background, which is then used to produce a spatial

The associate editor coordinating the review of this manuscript and approving it for publication was Shuhan Shen.

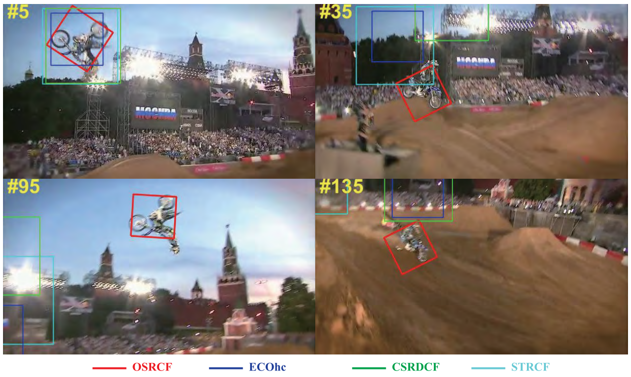


FIGURE 1. Comparisons of our approach with the top-ranked trackers in challenging situations of rotations on the Motocross1 sequence.

constraint regularization for correlation filter. But the color model is sensitive to illumination changes and easily drifts away when too much background is contained in the bounding box, which leads to false segmentation and tracking.

Another problem of the correlation filter is that the DCF is not rotation-invariant since the filter template strongly relies on spatial layout of the tracked object. However, most DCF based trackers only focus on the problem of position and scale estimation [32], [33], and ignore the rotation of the target. This results from the common axis-aligned representation of the target region and the slight changes of orientation in most testing sequences. However, if there is a significant orientation change, even the top-ranked trackers fail (see Fig.1). In [34], a rotation adaptive correlation filter is proposed, but it does not consider the boundary effects, which obtains unsatisfactory results.

Motivated by the above observations, we propose a novel Orientation and Scale adaptive tracker with Regularized Correlation Filters. Our key idea is to obtain a non-axis-aligned bounding box for the target region representation by an effective orientation and scale adaptive estimation scheme, which is then used to constrain the spatial regularization of correlation filters. To alleviate a rapid model degradation by large appearance changes and occlusion, we incorporate a target-region-based temporal regularization [35] to the spatial constraint correlation filters by penalizing the variations of appearance model. Meanwhile, we solve the model with an ADMM algorithm, which can empirically converge within very few iterations. Evaluations on recent visual tracking benchmark datasets show that our OSRCF tracker outperforms most state-of-the-art methods.

The contributions of this paper are as follows.

- A novel OSRCF model is developed by combining target-region-based spatial and temporal regularizations, thereby alleviating boundary effects and being robust for rapid appearance changes.
- An ADMM algorithm is developed for solving OSRCF efficiently.
- An orientation and scale adaptive tracker is proposed, which not only produces non-axis-aligned target region for spatio-temporal regularizations, but also improve the performance of tracking rotated targets.

The rest of this paper is organized as follows. Section II gives a brief overview of the related work. The proposed method is described in section III. In section IV we provide description and results of the performed experiments. Conclusions are finally presented in section V.

II. RELATED WORK

In this section, we provide a brief review on DCF based trackers and then discuss the methods closely related to this work.

A. DCF BASED TRACKERS

In the DCF based trackers, a target is tracked by correlating a filter over a larger search window and the location with the maximum value in the correlation response indicates the location of the target. Bolme et al. [20] first adopted correlation filters in tracking applications by minimizing the total squared error between the actual and the desired correlation output on a set of grayscale sample. By computing the correlation in Fourier domain, their MOSSE filter operates at hundreds of frames per second. Then Henriques et al. explored the circulant structure, called CSK [21], to extend [20] to kernel-based learning with dense sampling. Subsequently, KCF [36] boosted the performance of CSK by extending the input features from a single channel to multiple channels (e.g., HOG). The above work constructs the standard formulations of the DCF framework for visual tracking.

Later, a great deal of work has significantly improved the DCF framework by, e.g., incorporation of powerful features e.g. color names [37] and CNN features [38]–[40], scale estimation [32], [33], rotation adaption [34], [41], long-term memory [42], alleviating boundary effects [29]–[31], [43], fusing multi-resolution feature maps [44], [45] and ensembles [46]–[48].

B. REGULARIZATIONS

The DCF based trackers learn optimized filters by solving a ridge regression on the set of all cyclic shifts versions of the target. To avoid overfitting, a regularization term is commonly added into the regression target. Standard DCF based approaches use unified weights for every training pixel, which is unreasonable because the pixel away from the target center is unreliable. In [30], a Tikhonov regularization is adopted to penalize the DCF coefficients depending on their spatial locations. Then [35] boosts the performance of [30] by adding a temporal regularization on the whole training sample. In [31], the target region is extracted by a color model and the outside pixels are ignored when learning. Recently, [49] explores the nonlocal information to accurately represent and segment the target, which is then used to regularize a correlation filter. Compared with these trackers, our approach has the following merits: (1) A non-axis-aligned target region is used to provide a spatial constraint regularization; (2) A temporal regularization based on the target region is added instead of the whole sample region.

C. ORIENTATION AND SCALE ESTIMATION

In the standard form, the DCF based trackers can only estimate the horizontal and vertical location of the target in the image. But in some applications such as robotics, it is also crucial to estimate the additional target-centric information, such as scale and orientation. To incorporate scale estimation in a tracking framework, [32] proposed a straightforward scale adaptive scheme to handle scale changes, where a scaling pool is used to search the desired scale factor with maximum response. In [33], a discriminative scale space filter was proposed by learning a separate scale filter with the samples of the target at a set of different scales. All these trackers ignore the orientation estimation. To solve the problem, a rotation adaptive correlation filter tracker is proposed in [34] and [41]. However, they did not consider the boundary effects. And in this paper, we extend regularized DCF based trackers with the capability of handling scale and orientation changes.

III. THE OSRCF TRACKER

In this section, we firstly review the standard DCF formulation, and then introduce the target-region-based spatial and temporal regularization for the DCF proposed in this paper. Meanwhile, an ADMM is developed to solve the model. Finally, we present the orientation and scale adaptive scheme used in our approach.

A. THE STANDARD DCF TRACKER

In the standard DCF formulation, the target appearance is modeled by a multichannel filter. The aim is to learn the filter from a single training image patch with the target located at the center. A D-dimensional feature map $x \in \mathbb{R}^d$ is extracted from the image patch, which has the spatial size $M \times N$. At each location $(m, n) \in \{0, 1, \cdots, M-1\} \times \{0, 1, \cdots, N-1\}$, a training sample $x_{m,n}$ is generated by the circular shift of x with Gaussian function label $y_{m,n}$. We denote feature layer $d \in \{1, 2, \cdots, D\}$ of x by x^d. The desired filter f consists of one $M \times N$ correlation filter f^d per feature layer x^d, which can be solved by minimizing the L_2 error of the correlation response compared to the label y,

$$\min_f \frac{1}{2}\sum_{d=1}^{D}\left\| f^d \star x^d - y \right\|^2 + \frac{\lambda}{2}\sum_{d=1}^{D}\left\| f^d \right\|^2 \tag{1}$$

where $\star$ denotes circular correlation and λ is the weight of the regularization term. Using Parseval's formula, Eq.1 can be transformed to a regression objective in the Fourier domain,

$$E\left(F\right) = \frac{1}{2}\sum_{d=1}^{D}\left\| F^d \odot X^d - Y \right\|^2 + \frac{\lambda}{2}\sum_{d=1}^{D}\left\| F^d \right\|^2 \tag{2}$$

Here F,X,Y are the Fourier transforms of f,x,y and the operator $\odot$ denotes Hadamard product. Eq.2 has a closed-form solution by

$$F^d = \frac{\overline{Y} \odot X^d}{\overline{X^d} \odot X^d + \lambda}, \quad d = \{1, 2, \cdots, D\} \tag{3}$$

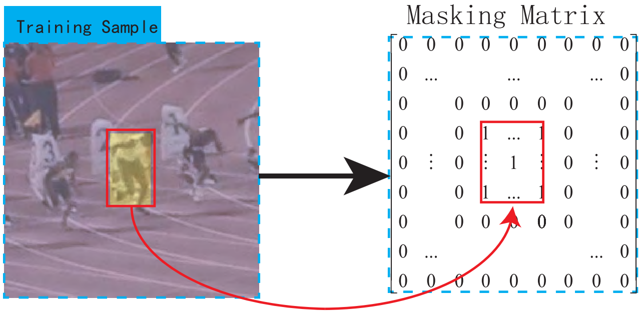


FIGURE 2. The masking matrix $\mathcal{M}$ is determined by the target region, where the element is binary. The training sample is obtained by padding the target bounding box. After feature extraction, $\mathcal{M}$ is resized to the feature space scale and remains the same during the tracking period.

Here the bar $\overline{X}$ denotes the complex conjugation and all the operations in Eq.3 are point-wise.

To learn the filter in a robust way, the numerator A^d and denominator B of the filter is updated linearly according to [20] as

$$\begin{cases} A^d = (1-\eta)\,A^d + \eta\overline{Y} \odot X^d \\ B = (1-\eta)\,B + \eta\displaystyle\sum_{d=1}^{D}\overline{X^d} \odot X^d \\ F^d = \dfrac{A^d}{B+\lambda}, \quad d = 1, 2, \cdots, D \end{cases} \tag{4}$$

Here the scalar η is the learning rate parameter.

When a new frame t comes, a feature map x_t is extracted from the image patch centered around the predicted target location as the same way as the training image patch. The correlation response z between the filter f and feature map x_t is computed in the Fourier domain by

$$z = \mathcal{F}^{-1}\left(\sum_{d=1}^{D} F^d \odot X_t^d\right) \tag{5}$$

Here $\mathcal{F}^{-1}$ denotes the inverse DFT and the maximum value of z indicates the location of the target.

B. SPATIALLY AND TEMPORALLY REGULARIZED DCF

The standard DCF suffers from boundary effect and is sensitive to rapid appearance changes. To deal with these problems, we incorporate target-region-based spatial and temporal regularizations into the standard DCF. Given a target region $\mathcal{O}$, we define a masking matrix $\mathcal{M} \in \{0, 1\}^{M \times N}$ by

$$\mathcal{M}\,(m, n) = \begin{cases} 1, & (m, n) \in \mathcal{O} \\ 0, & (m, n) \notin \mathcal{O} \end{cases} \tag{6}$$

The construction process is shown in Fig.2, And each element of $\mathcal{M}$ indicates whether the pixel should be active or inactive in learning, so it is very important to obtain the accurate target region. Different from other methods using a segmentation model, we propose an orientation and scale adaptive strategy to get the non-axis-aligned bounding box for target region representation, which is detailed described in Section III-C.

To ensure the coefficients of filter is zero outside of the target region, we introduce a constraint $f \equiv \mathcal{M} \odot f$, resulting a spatial regularization $\|\mathcal{M} \odot f\|^2$. Note that $\mathcal{M}$ remains the same once initialized because the training patch is regularized to the same size. Based on the target region, a temporal regularization term $\|\mathcal{M} \odot f - \mathcal{M} \odot f_{t-1}\|^2$ is exploited to smooth the CF learning, which leads to the following constrained optimization problem

$$\begin{aligned} \min_f \frac{1}{2} \sum_{d=1}^{D} \left\| f^d \star x^d - y \right\|^2 + \frac{\lambda}{2} \sum_{d=1}^{D} \left\| \mathcal{M} \odot f^d \right\|^2 \\ + \frac{\mu}{2} \sum_{d=1}^{D} \left\| \mathcal{M} \odot f^d - \mathcal{M} \odot f_{t-1}^d \right\|^2 \\ s.t. f \equiv \mathcal{M} \odot f \end{aligned} \tag{7}$$

where λ and μ are the constraint penalty values. We define an auxiliary variable $g^d = \mathcal{M} \odot f^d$, and we have $f_{t-1}^d \equiv \mathcal{M} \odot f_{t-1}^d$, then Eq.7 can be formulated as

$$\begin{aligned} \min_f \frac{1}{2} \sum_{d=1}^{D} \left\| f^d \star x^d - y \right\|^2 + \frac{\lambda}{2} \sum_{d=1}^{D} \left\| g^d \right\|^2 \\ + \frac{\mu}{2} \sum_{d=1}^{D} \left\| g^d - f_{t-1}^d \right\|^2 \\ s.t. f^d - g^d \equiv 0 \end{aligned} \tag{8}$$

The above problem can be solved by the following augmented Lagrangian

$$\begin{aligned} \mathcal{L}(f, g, h) = \frac{1}{2} \sum_{d=1}^{D} \left\| f^d \star x^d - y \right\|^2 + \frac{\lambda}{2} \sum_{d=1}^{D} \left\| g^d \right\|^2 \\ + \frac{\mu}{2} \sum_{d=1}^{D} \left\| g^d - f_{t-1}^d \right\|^2 + \sum_{d=1}^{D} (f^d - g^d) h^d \\ + \frac{\gamma}{2} \sum_{d=1}^{D} \left\| f^d - g^d \right\|^2 \end{aligned} \tag{9}$$

where h is a Lagrangian multiplier and γ is the constraint penalty value. Specially, coefficients in f residing outside the target region are suppressed to zero by assigning higher weights to γ. We introduce $s^d = \frac{h^d}{\gamma}$, then Eq.9 can be formulated as

$$\begin{aligned} \mathcal{L}(f, g, h) = \frac{1}{2} \sum_{d=1}^{D} \left\| f^d \star x^d - y \right\|^2 + \frac{\lambda}{2} \sum_{d=1}^{D} \left\| g^d \right\|^2 \\ + \frac{\mu}{2} \sum_{d=1}^{D} \left\| g^d - f_{t-1}^d \right\|^2 + \frac{\gamma}{2} \sum_{d=1}^{D} \left\| f^d - g^d + s^d \right\|^2 \end{aligned} \tag{10}$$

Using Parseval's formula, Eq.10 can be transformed as

$$\begin{aligned} &\mathcal{L}(F, G, S) \\ &= \frac{1}{2} \sum_{d=1}^{D} \left\| F^d \odot X^d - Y \right\|^2 + \frac{\lambda}{2} \sum_{d=1}^{D} \left\| G^d \right\|^2 \\ &\quad + \frac{\mu}{2} \sum_{d=1}^{D} \left\| G^d - F_{t-1}^d \right\|^2 + \frac{\gamma}{2} \sum_{d=1}^{D} \left\| F^d - G^d + S^d \right\|^2 \end{aligned} \tag{11}$$

Then we adopt ADMM algorithm to solve the above problem by solving the following subproblems at each iteration,

$$\begin{cases} F^{(i+1)} = \min_F \frac{1}{2} \sum_{d=1}^{D} \left\| F^d \odot X^d - Y \right\|^2 \\ \qquad + \frac{\gamma}{2} \sum_{d=1}^{D} \left\| F^d - G^d + S^d \right\|^2 \\ G^{(i+1)} = \min_G \frac{\lambda}{2} \sum_{d=1}^{D} \left\| G^d \right\|^2 + \frac{\mu}{2} \sum_{d=1}^{D} \left\| G^d - F_{t-1}^d \right\|^2 \\ \qquad + \frac{\gamma}{2} \sum_{d=1}^{D} \left\| F^d - G^d + S^d \right\|^2 \\ S^{(i+1)} = S^{(i)} + F^{(i+1)} - G^{(i+1)} \end{cases} \tag{12}$$

The minimizations in Eq.12 have a closed-form solution:

$$F = \frac{\overline{Y} \odot X + \gamma G - H}{\overline{X} \odot X + \gamma} \tag{13}$$

$$G = \mathcal{F}(\mathcal{M} \odot \mathcal{F}^{-1}(\frac{\mu F_{t-1} + \gamma F + H}{\lambda + \mu + \gamma})) \tag{14}$$

and the Lagrange multiplier H and the constraint penalty γ are updated as follows:

$$H^{(i+1)} = H^{(i)} + \gamma^{(i)} (F - G) \tag{15}$$

$$\gamma^{(i+1)} = \beta \gamma^{(i)} \tag{16}$$

Since the operations in Eq.13 are fully element-wise, the computation cost for F is $\mathcal{O}(DMN)$. The solution for Eq.14 requires a single inverse FFT and another FFT, so the complexity of solving G is $\mathcal{O}(DMN \log MN)$. Hence, the overall computational complexity of the algorithm is $\mathcal{O}(DMN \log MNN_I)$, where N_I is the number of iterations. Note that the algorithm can empirically converge within very few iterations. The procedure of filter learning is summarized in the Algorithm 1.

C. ORIENTATION AND SCALE ADAPTIVE SCHEME

The spatial and temporal constraint as described in Section III-B depends on the masking matrix $\mathcal{M}$, which is obtained from the target region. To estimate it accurately, we propose an orientation and scale adaptive scheme. The orientation search is in a straightforward way. We define a rotating pool $\mathcal{R} = \{r_1, r_2, \cdots, r_k\}$, and use bilinear interpolation to get the rotated image patches from the original image space. For each $r_i \in \mathcal{R}$, we extract an image patch of

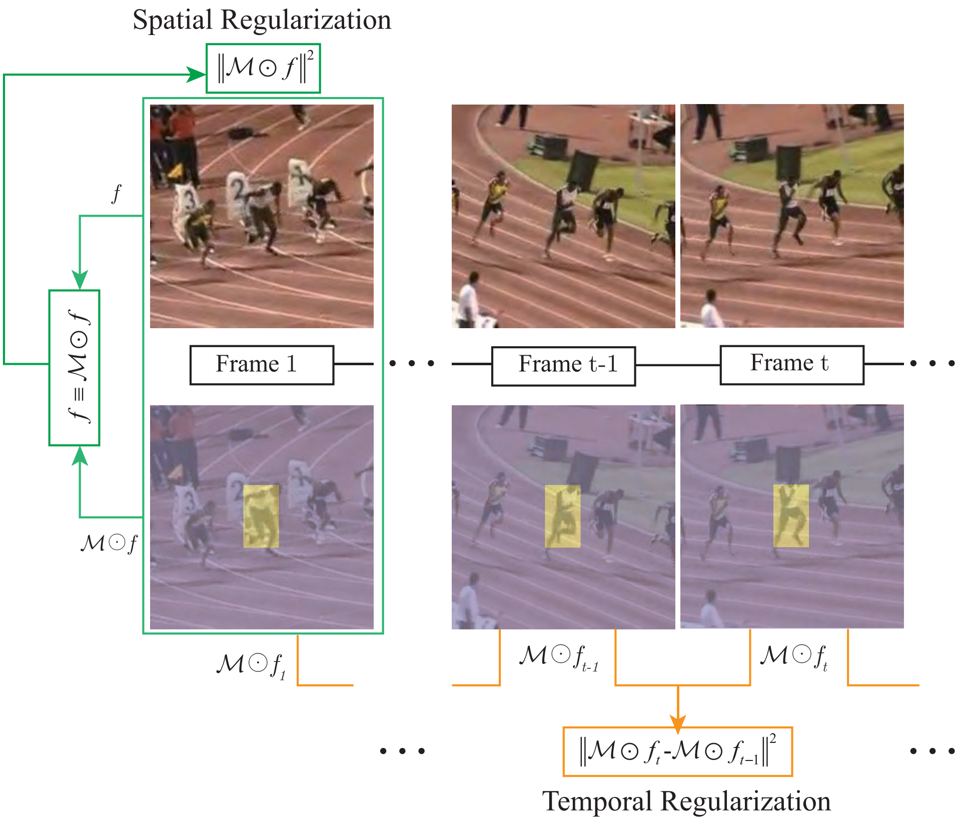


FIGURE 3. Model learning with the combination of the target-region-based spatial and temporal regularization on the standard DCF.

Algorithm 1 Learning Spatially and Temporally Regularized DCF

```
Input: Feature map $X$, desired correlation response $Y$,
    masking matrix $\mathcal{M}$, previous filter $F_{t-1}$
Output: Optimized filter $F$
Computing the standard DCF $F$ based on Eq.3
Initializing the filter $G = \mathcal{F}(\mathcal{M} \odot \mathcal{F}^{-1}(F))$
while not converged do
    Update $F$ via Eq.13
    Update $G$ via Eq.14
    Update $H$ via Eq.15
end while
return $F \Leftarrow G$
```

orientation r_i+r_{t-1} centered around the target with fixed size. Here, r_{t-1} denotes the target orientation in the frame $(t-1)$. Then a feature map x_i is extracted from each patch and the final response is calculated by

$$z_i = \mathcal{F}^{-1}(\sum_{d=1}^{D} F^d \odot X_i^d), \quad i = 1, 2, \cdots, k \tag{17}$$

where F is the learnt filter described in Section III-B and the maximum of $\{z\}$ indicates the desired orientation. As the target translation is implied in the response map of proper orientation, the final movement needs to be tuned to obtain the real location of the target.

To estimate the scale of the target, we use a discriminative scale space filter F_{scale} described in [33]. Specifically, to construct the training samples, we extract feature maps x_{scale} using variable patch sizes centered around the target with a proper orientation. By maximizing the scale correlation scores

$$z_{scale} = \mathcal{F}^{-1}(F_{scale} \odot X_{scale}) \tag{18}$$

we obtain the relative change in scale compared to the previous frame.

D. TRACKING WITH OSRCF TRACKER

Based on the target-region-based spatially and temporally regularized filter and the orientation and scale adaptive scheme, we propose a OSRCF tracker. The target state estimation and model update steps of the tracking framework proceed as follows.

1) TARGET STATE ESTIMATION

The target is first localized by the response of the learned regularized DCF F_{t-1}. Orientation is estimated by maximizing the response of F_{t-1} on variable patch rotations and a new position p_t is obtained by tuning the translation. Scale is estimated by maximizing the response of F_{scale}. The procedure is visualized in Fig.4.

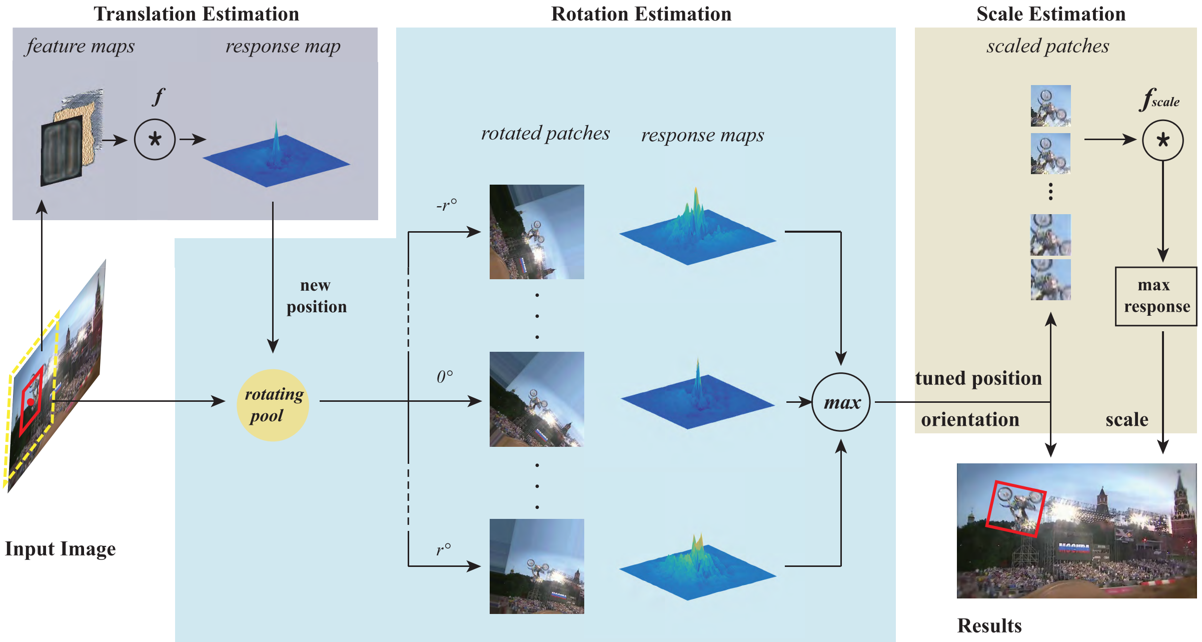


FIGURE 4. Visualization of the procedure used to estimate the target state. When a new frame comes, we estimate the translation, rotation and scale of the tracking object to obtain a non-axis-aligned target region, which is then used to update the model.

2) MODEL UPDATE

The masking matrix $\mathcal{M}$ is created by Eq.6 when initializing and the regularized filter F_t is computed with Algorithm 1. We adopt an incremental strategy to update the filters with learning rate η. A brief outline of our OSRCF approach is given in Algorithm 2.

IV. EXPERIMENTS

In this section, we present a comprehensive experimental evaluation of the proposed method. Section IV-A gives the implementation details. In section IV-B, comparative experiments on the recent benchmark VOT2017 [27] is conducted and demonstrate state-of-the-art performance. Furthermore, we compare our tracker with some deep-learning-based trackers in Section IV-C. Next, we conduct spatial and temporal penalty factors experiments to analyze the effects each term in Eq.9. We then carry out ablative studies in Section IV-E. Section IV-F gives a qualitative comparison with related competing method in the VOT2017 benchmark dataset. Finally, we evaluate our tracker on OTB2015 [28] dataset.

A. IMPLEMENTATION DETAILS

Our tracker is implemented by native Matlab without optimization. The experiments are conducted on an Intel E5-1650v4 CPU (3.60 GHz) PC with 16 GB memory. Our proposed OSRCF tracker runs at about 10 fps. We set the initial position, orientation and size of the target based on a bounding box centered on the object in the first frame. Note that, the selected bounding box can be either a rectangle or a polygon. The HOG and Color Names (CN) features are used in our tracker. And for fair comparison, we firstly compare the proposed method with DCF based trackers only using hand-crafted features. Then, we further compare our tracker with some deep-learning-based methods.

Algorithm 2 The OSRCF Tracking Algorithm

```
Input: Image $I_t$. Previous target position $p_{t-1}$, orientation $r_{t-1}$ and scale $s_{t-1}$. Masking matrix $\mathcal{M}$, Regularized DCF $F_{t-1}$. Scale filter $F_{scale,t-1}$.
Output: Estimated target position $p_t$, orientation $r_t$ and scale $s_t$. Updated regularized DCF $F_t$. Updated scale filter $F_{scale,t}$.
Target State Estimation
Extract feature maps $x_t$ from $I_t$ at $p_{t-1}$, $r_{t-1}$ and $s_{t-1}$.
Compute the correlation response $z_t$ using Eq.5.
Set $p_t$ to the target position that maximizes $z_t$.
for all $r_i \in \mathcal{R}$ do
    Extract feature maps $x_{t,i}$ from $I_t$ at $p_t$, $r_i + r_{t-1}$ and $s_{t-1}$.
    Compute the correlation response $z_{t,i}$ using Eq.17.
end for
Set $r_t$ and new $p_t$ to the target orientation and position that maximizes $\{z_{t,i}\}$.
Extract feature maps $x_{scale,t}$ from $I_t$ at $p_t$ and $r_t$.
Compute the correlation response $z_{scale,t}$ using Eq.18.
Set $s_t$ to the target scale that maximizes $z_{scale,t}$.
Model Update
Compute a new regularized DCF $F$.
Compute a new scale filter $F_{scale}$.
Update filter $F_t = (1-\eta)F_{t-1} + \eta F$.
Update filter $F_{scale,t} = (1-\eta_s)F_{scale,t-1} + \eta_s F_{scale}$.
```

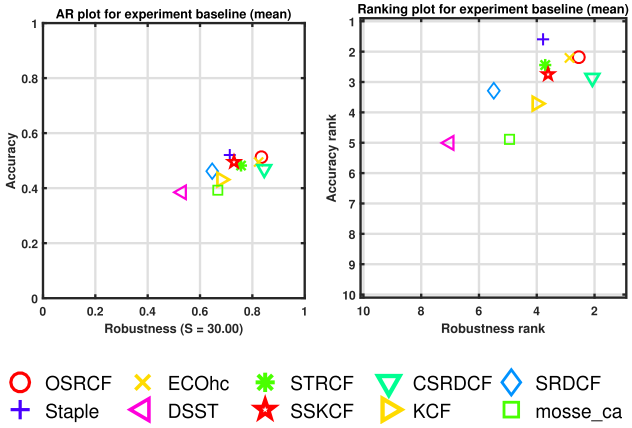


FIGURE 5. The performance of 10 trackers in accuracy and robustness. The results show our OSRCF tracker outperforms other DCF based trackers that do not apply deep convolutional features.

Here, we use VOT2017 dataset as our validation set for tuning all hyper parameters. The spatial regularization parameter is set to $\lambda = 0.01$ and temporal regularization parameter is set to $\mu = 4$. The augmented Lagrangian optimization parameters are set to $\gamma^{(0)} = 5$ and $\beta = 3$, the number of iterations is $N_I = 2$. The correlation filter adaptation rate is set to $\eta = 0.05$. Rotation from a fixed pool of 4 angles ranging from $-30°$ to $30°$ is used to detect the orientation change and the parameters for the scale filter are set to the same as in [33]. Further, we use the same parameter values and initialization for all the sequences.

B. THE VOT2017 BENCHMARK

The VOT2017 dataset [27] consists of 60 challenging sequences and is unarguably the most difficult sequence set in contrast to related benchmark. The VOT methodology resets a tracker upon failure to fully use the dataset and the expected average overlap (EAO) is used to measure the performance which combines the raw values of per-frame accuracies and failures in a principled manner.

Table 1 shows the comparison of our tracker with other 9 DCF based trackers with handcrafted features including ECOhc [45], STRCF [35], CSRDCF [31], Staple [50], SRDCF [30], DSST [33], SSKCF [51], KCF [36], and mosseca_ca [52]. Our OSRCF tracker achieves the EAO score of 0.2718. And compared with the CSRDCF and ECOhc methods which are the top-two DCF based trackers using handcrafted features [27], the performance gain is 6.1% and 14.0%, respectively. Fig.5 presents the detailed performance of these trackers in terms of accuracy (overlap with the ground truth) and robustness (failure rate) [53], [54]. Compared to the second best approach (CSRDCF), our method obtains a significant gain in accuracy.

C. COMPARISON WITH DEEP-LEARNING-BASED TRACKERS

We further compare our OSRCF with 8 deep-learning-based trackers MCCT [55], CCOT [44], SiamDCF [56], MCPF [57], CRT [58], DLST [59], RCPF [60], SiamFC [61] on VOT2017. Fig.6 presents the EAO score of each tracker and the best three are shown in red, blue and green fonts,

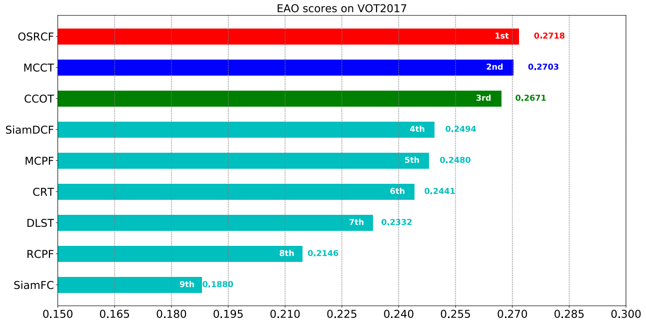


FIGURE 6. A comparison with 8 deep-learning-based trackers on VOT2017 dataset.

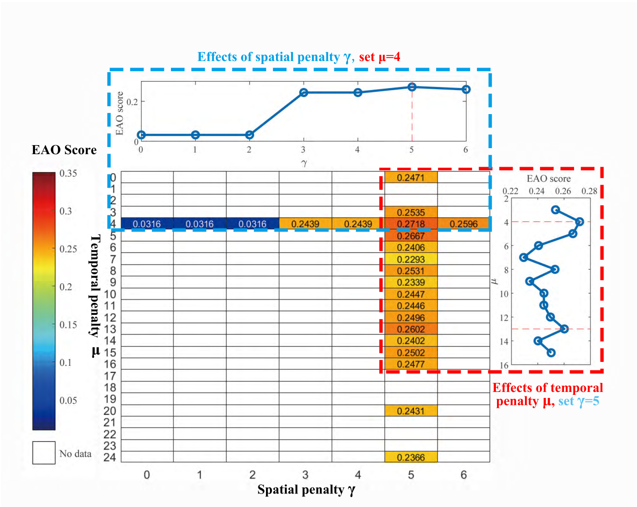


FIGURE 7. Spatial and temporal penalty factors experiments on VOT2017 dataset.

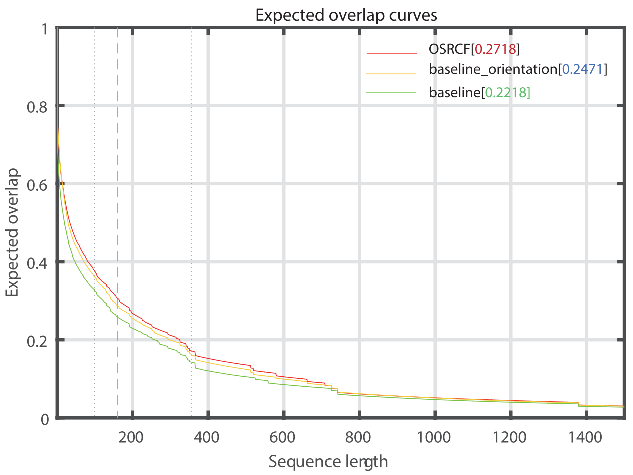


FIGURE 8. Ablation analysis on VOT2017 datasets and the EAO score is denoted in the figure.

TABLE 1. Comparison with state-of-the-art methods on the VOT2017 dataset. The results are presented by EAO and the best three results are shown in red, blue and green fonts, respectively.

	OSRCF	ECOhc	STRCF	CSRDCF	Staple	SRDCF	DSST	SSKCF	KCF	mosse_ca
EAO	0.2718	0.2384	0.1765	0.2561	0.1694	0.1189	0.0788	0.1660	0.1349	0.1406

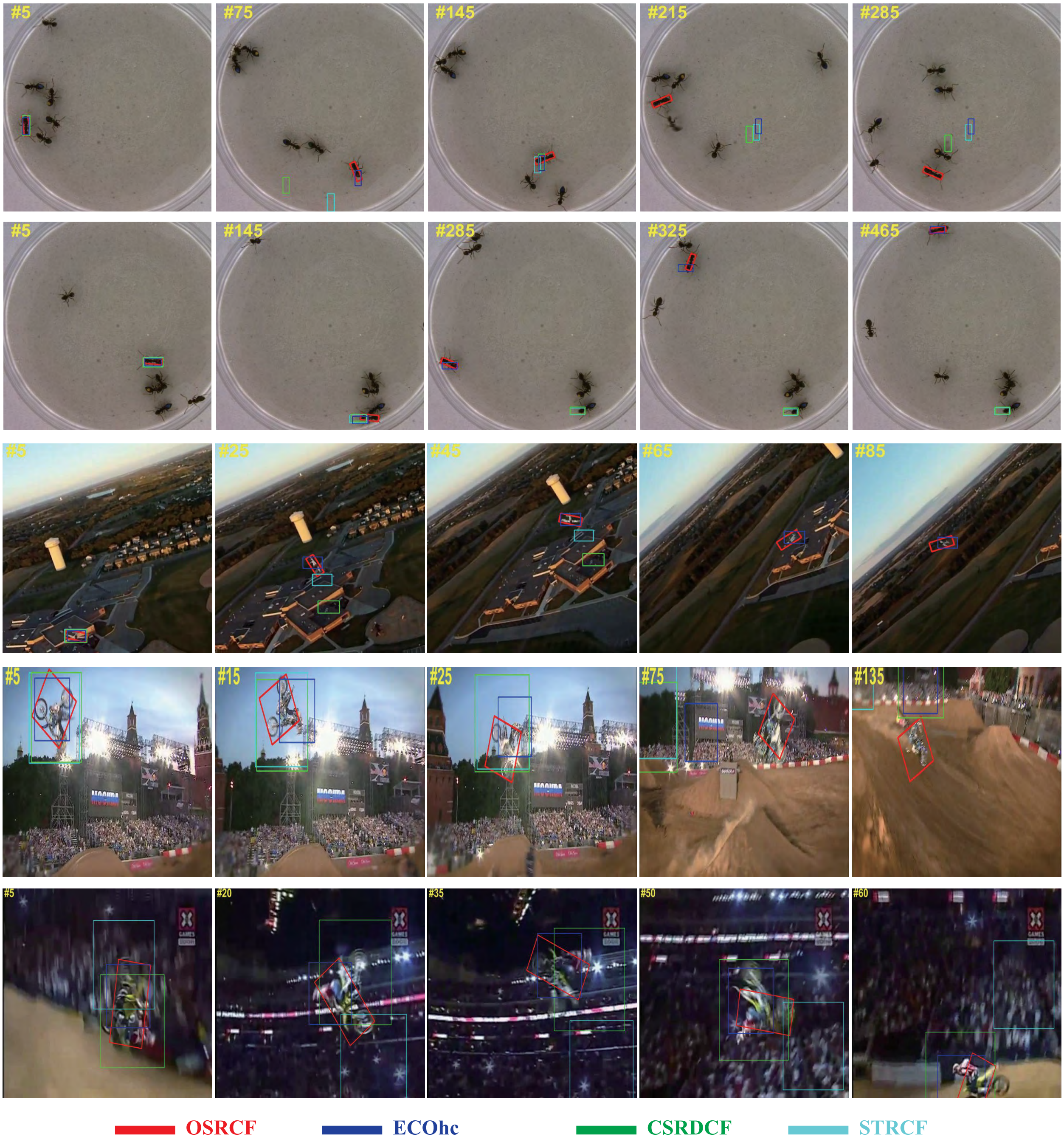
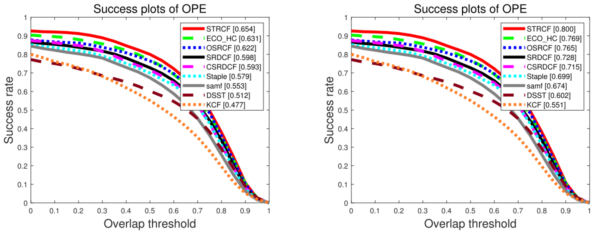


FIGURE 9. Comparisons of the proposed OSRCF tracker with the state-of-the-art trackers (ECOhc, CSRDCF, STRCF) in the unsupervised experiment on 5 challenging sequences (from top to down are ants1, ants3, drone_flip, motocross1, motocross2, respectively).

respectively. One can see that OSRCF performs better than six DCF based trackers with deep features (MCCT, CCOT, MCPF, CRT, RCPF) and two CNN matching based trackers (SiamDCF, SiamFC).

D. SPATIAL AND TEMPORAL PENALTY FACTORS EXPERIMENTS

In Eq.7, the constraint condition $f \equiv \mathcal{M} \odot f$ makes the coefficients in f residing outside the target region close to



(a) (b)

FIGURE 10. Success plots over all 100 sequences using one-pass evaluation on the OTB-2015 dataset. The score for each tracker is shown in the legend. Our OSRCF method performs favorably against the state-of-the-art trackers.

zero, which is controlled by the spatial penalty factor γ in Eq.9. And the λ controls the regularization on the coefficients within the target region, which is commonly set to a small value (0.01 in this paper). The temporal penalty factor μ makes the coefficients in f within the target region change more smoothly by assign higher value and vice versa. Fig.7 illustrates how the spatial and temporal penalty factors, γ and μ, affect the tracking performance on VOT2017. We first set $\mu = 4$ analyze the effects of spatial penalty factor γ. One can see that the best EAO score is achieved at $\gamma = 5$. Then we set $\gamma = 5$ and further analyze the effects of temporal penalty μ. From the graph, we can see that the values around 4 achieve better performances. Finally, we set $\gamma = 5$ and $\mu = 4$, which achieves EAO score 0.2718 on VOT2017 dataset.

E. ABLATION ANALYSIS

We use experiments to justify the effectiveness of our orientation adaptive scheme and spatial-constraint temporal regularization in OSRCF. We use the VOT2017 datasets for the ablation analysis.

We first evaluate the performance of the baseline method which only contains spatial regularization. Then we add the orientation adaption scheme and spatial-constraint temporal regularization to the baseline model, respectively. The results of EAO scores are reported in Fig.8. Compared with baseline method, the orientation adaption scheme (baseline_oriention) advances the performance by 11.4% in EAO score. On the other side, a gain of 10.0% in EAO score is achieved by considering the spatial-constraint temporal regularization. The overall OSRCF dramatically improves the performance by an EAO score of 22.5% against the baseline, which demonstrates the effectiveness of our orientation adaptive scheme and spatial-constraint temporal regularization in practical tracking.

F. QUALITATIVE COMPARISON

To demonstrate the effect of orientation adaptive scheme in our tracking algorithm, we make a qualitative comparison

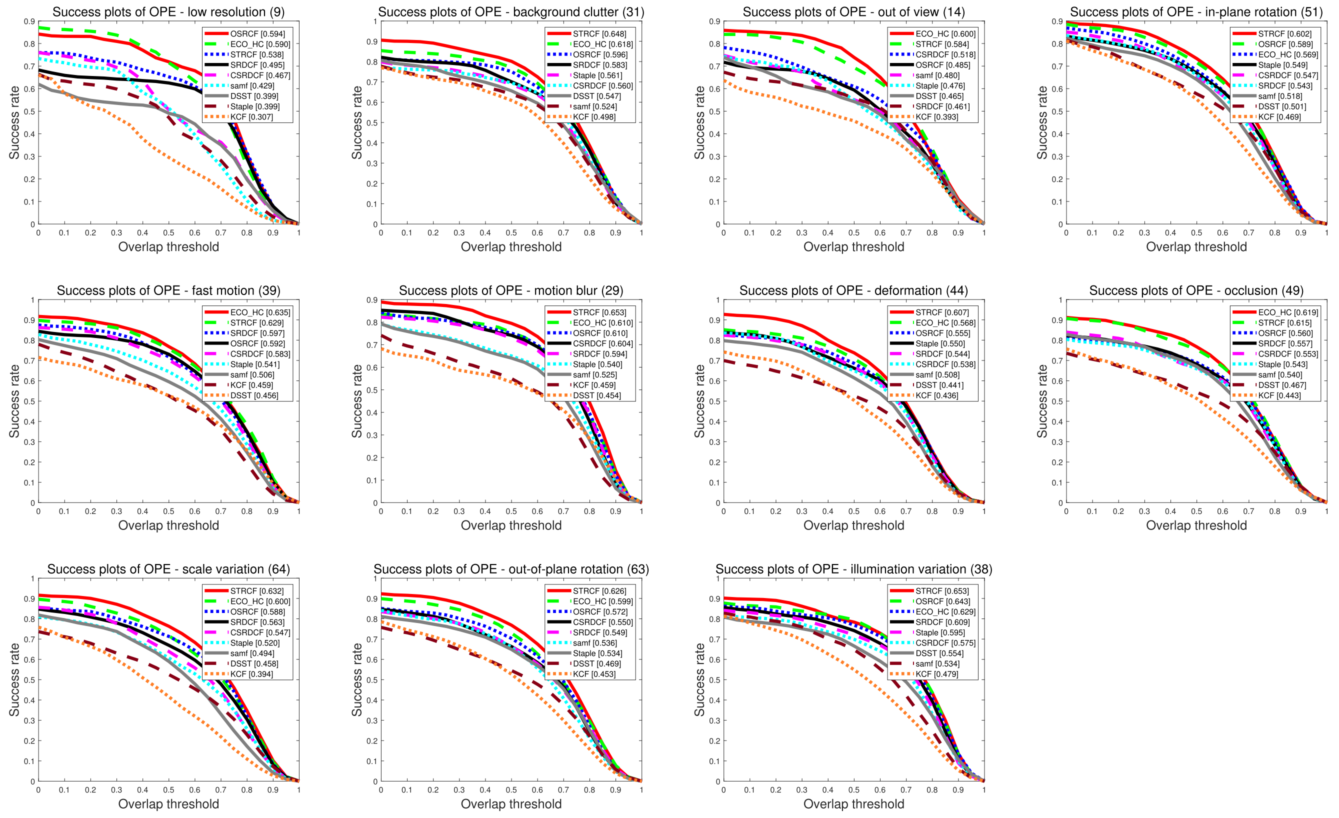


FIGURE 11. Overlap success plots over 11 tracking challenges of low resolution, background clutter, out-of-view, in-plane rotation, fast motion, motion blur, deformation, occlusion, scale variation, out-of-plane rotation and illumination variation.

with other top-four trackers (ECOhc, CSRDCF, STRCF) on 5 different sequences where there are significant orientation variations. As shown in Fig. 9, our OSRCF tracker achieve the best performance among all four trackers. Although the ECOhc is capable of tracking rotated objects, such as in the ants3, drone_flip and motocross2 sequence, it cannot estimate the orientation variations in all sequences and suffers from a significant rotation drift. In the motocross1 sequence, all the compared trackers struggle due to difficult lighting conditions and rotation motions, while our tracker robustly handles these factors. In addition to robustly tracking the target, our approach accurately estimates the orientation variations and is able to keep track of the target throughout the sequence.

G. THE OTB2015 DATASET

We evaluate our approach on the OTB2015 dataset which consists of 100 challenging videos. The OTB methodology uses mean overlap precision (OP) and area-under-the-curve (AUC) scores to evaluate the performance of trackers. The OP score is calculated as the percentage of frames in a video where the intersection-over-union (IOU) overlap with the ground-truth exceeds a certain threshold (0.5 in this experiment). The mean OP over all videos is plotted over the range of IOU thresholds [0 1] to get the success plot (see Fig.10(a)). The area under this plot gives the AUC score (see Fig.10(b)). We refer to [28] for details. Our method is compared with 8 recent DCF based trackers with handcrafted features (STRCF [35], ECO-HC [45], SRDCF [30], CSRDCF [31], Staple [50], samf [32], DSST [33], KCF [36]). The proposed OSRCF algorithm achieves the AUC score of 62.2% and OP of 76.5%. Overall, our algorithm achieves comparable results. The AUC results of 11 challenging attributes are detailedly illustrated in Fig.11.

V. CONCLUSIONS

We propose a target-region-based spatially and temporally regularized correlation filter to reduce the boundary effect and improve temporal robustness, which is solved by an ADMM method efficiently. To obtain the non-axis-aligned region of the tracked target, We introduce an orientation and scale adaptive scheme. Evaluation on VOT2017 shows that our OSRCF tracker achieves the top result among DCF based trackers with handcrafted features.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, 2006, Art. no. 13.

[2] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomputing*, vol. 74, no. 18, pp. 3823–3831, Nov. 2011.

[3] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, Jul. 2014.

[4] N. Wang, J. Shi, D.-Y. Yeung, and J. Jia, "Understanding and diagnosing visual tracking systems," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3101–3109.

[5] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 798–805.

[6] N. Alt, S. Hinterstoisser, and N. Navab, "Rapid selection of reliable templates for visual tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis.Pattern Recognit.*, Jun. 2010, pp. 1355–1362.

[7] S. He, Q. Yang, R. W. H. Lau, J. Wang, and M.-H. Yang, "Visual tracking via locality sensitive histograms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2427–2434.

[8] M. J. Black and A. D. Jepson, "EigenTracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Comput. Vis.*, vol. 26, no. 1, pp. 63–84, 1998.

[9] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 125–141, 2008.

[10] X. Mei and H. Ling, "Robust visual tracking using $l1$ minimization," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1436–1443.

[11] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2042–2049.

[12] T. Zhang *et al.*, "Structural sparse tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 150–158.

[13] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, Jun. 2000, pp. 142–149.

[14] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1296–1311, Oct. 2003.

[15] D. Wang, H. Lu, and M.-H. Yang, "Least soft-threshold squares tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2371–2378.

[16] G. Helmut, "Real-time tracking via on-line boosting," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2006, pp. 47–56.

[17] B. Babenko, M. H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 983–990.

[18] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[19] S. Hare *et al.*, "Struck: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, Oct. 2016.

[20] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.

[21] J. F. Henriques, C. Rui, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 702–715.

[22] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Dec. 2013, pp. 809–817.

[23] H. Fan and H. Ling, "SANet: Structure-aware network for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 42–49.

[24] I. Jung, J. Son, M. Baek, and B. Han, "Real-time MDNet," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, Sep. 2018, pp. 83–98.

[25] B. Chen, D. Wang, P. Li, S. Wang, and H. Lu, "Real-time 'actor-critic' tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, Sep. 2018, pp. 328–345.

[26] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1420–1429.

[27] M. Kristan *et al.*, "The visual object tracking vot2017 challenge results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Oct. 2017, pp. 1949–1972.

[28] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[29] H. K. Galoogahi, T. Sim, and S. Lucey, "Correlation filters with limited boundaries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4630–4638.

[30] M. Danelljan, G. Hääger, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4310–4318.

[31] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," *Int. J. Comput. Vis.*, vol. 126, no. 7, pp. 671–688, Jul. 2018.

[32] Y. Li and J. Zhu, "A scale adaptive Kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 254–265.

[33] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2017.

[34] Q. Du, Z.-Q. Cai, H. Liu, and Z. L. Yu, "A rotation adaptive correlation filter for robust tracking," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, Jul. 2015, pp. 1035–1038.

[35] F. Li, C. Tian, W. Zuo, L. Zhang, and M. H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4904–4913.

[36] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[37] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1090–1097.

[38] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Dec. 2015, pp. 621–629.

[39] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg, "Unveiling the power of deep tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 483–498.

[40] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.

[41] M. Zhang, J. Xing, J. Gao, X. Shi, Q. Wang, and W. Hu, "Joint scale-spatial correlation tracking with adaptive rotation estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 32–40.

[42] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5388–5396.

[43] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1144–1152.

[44] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 472–488.

[45] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR*, Jul. 2017, pp. 6931–6939.

[46] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 749–758.

[47] Y. Qi *et al.*, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4303–4311.

[48] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),*, Jun. 2016, pp. 1401–1409.

[49] K. Zhang, X. Li, H. Song, Q. Liu, and W. Lian, "Visual tracking using spatio-temporally nonlocally regularized correlation filter," *Pattern Recognit.*, vol. 83, pp. 185–195, Nov. 2018.

[50] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis.Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1401–1409.

[51] J.-Y. Lee and W. Yu, "Visual tracking by partition-based histogram backprojection and maximum support criteria," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2011, pp. 2860–2865.

[52] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1387–1395.

[53] M. Kristan *et al.*, "The visual object tracking vot2013 challenge results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 98–111.

[54] M. Kristan *et al.*, "The visual object tracking vot2014 challenge results," in *Proc. Eur. Conf. Comput. Vis. Workshops*, Mar. 2015, pp. 191–217.

[55] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4844–4853.

[56] W. Qiang, G. Jin, J. Xing, M. Zhang, and W. Hu. (2017). "Dcfnet: Discriminant correlation filters network for visual tracking." [Online]. Available: https://arxiv.org/abs/1704.04057

[57] T. Zhang, C. Xu, and M.-H. Yang, "Learning multi-task correlation particle filters for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 365–378, Feb. 2018.

[58] K. Chen and W. Tao, "Convolutional regression for visual tracking," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3611–3620, Jul. 2018.

[59] L. Yang, R. Liu, D. Zhang, and L. Zhang, "Deep location-specific tracking," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1309–1317.

[60] T. Zhang, S. Liu, C. Xu, B. Liu, and M.-H. Yang, "Correlation particle filter for visual tracking," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2676–2687, Jun. 2018.

[61] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.

KE TAN received the B.Eng. degree from the Department of Instrument Science and Opto-electronics Engineering, Beihang University, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree. His current research interests are object tracking, machine learning, and image processing.

ZHENZHONG WEI received the B.S. degree from the Automation Department, Beijing Institute of Petro-Chemical Technology, China, in 1997, and the M.S. and Ph.D. degrees from the School of Automation Science and Electrical Engineering, Beihang University, China, in 1999 and 2003, respectively, where he is currently a Professor with the School of Instrumentation Science and Opto-electronics Engineering. His research interests include machine vision and artificial intelligence. He was awarded as the Special Professor of Yangtze River Scholar, in 2016.