Saliency Guided Visual Tracking via Correlation Filter With Log-Gabor Filter

MINGXIN YU, YUHUA ZHANG, YONGKE LI, ZHI-LONG LIN, JIANZENG LI, AND CHANGLONG WANG
Shijiazhuang Campus of Army Engineering University, Shijiazhuang 050003, China
Corresponding author: Changlong Wang (wang-oec@126.com)
This work was supported by the National Natural Science Foundation of China under Grant 61801507.

ABSTRACT Correlation filter (CF) based tracking algorithms have tremendously contributed to the field of visual tracking due to the high computational efficiency and competitive performance. Nonetheless, most CF-based trackers are vulnerable to the influence of occlusion and boundary effect, which results in suboptimal performance. In this article, we propose saliency guided visual tracking via correlation filter with log-Gabor filter to robustify its performance under occlusion and boundary effect challenges. Firstly, we propose the CF with log-Gabor filter to get a robust appearance model. The log-Gabor filter is adopted to preprocess the sequence to gain the log-Gabor feature, which provides important cues for tracking since it encodes the texture information. Secondly, considering the prior information, we embed the novel saliency guided adaptive spatial feature selection to filter learning to preserve the spatial structure in the lower manifold and mitigate boundary distortion. Thirdly, the occlusion estimating strategy, performing on-line evaluation of tracking, triggers the motion estimation module to optimize the optimal location. Experiments on benchmark databases demonstrate the enhanced discrimination and interpretability of the proposed tracker and its superiority over other trackers.

INDEX TERMS CF with log-Gabor filter, saliency prior information, adaptive feature selection, motion estimation module.

I. INTRODUCTION
Visual tracking is a classical and rapidly evolving research topic in computer vision with various real-world applications including video surveillance, human-computer interaction, unmanned aerial vehicles (UAVs) and autonomous driving. The objective is the target trajectory estimation given only its initial location. Although significant achievements have been accomplished within the last few years [1]–[5], stably tracking arbitrary objects is still a complicated yet systematic task due to several factors, such as objection deformation, occlusion, and motion blur. In order to ensure accurate tracking, the research of visual tracking with different challenges still has broad prospects and important research significance.

Recently, CF-based approaches have witnessed astonishing improvements in visual tracking robustness. A discriminative correlation filter is trained to distinguish the target from the surrounding. The intrinsic idea of exploiting the circulant structure shows competitive performance and high computational speed, but brings undesired boundary effect. Danelljan et al. [6] introduce spatial regularized constraints into DCF to restrain correlation filter coefficients depending on their spatial location, which solves the boundary effects efficiently. A STRCF model [7] is presented by incorporating both spatial and temporal regularization into the DCF framework to encourage temporal consistency. However, the constraints are usually fixed for different objects and not changed during the tracking process, which cannot fully exploit the diversity information of different objects along the timeline. SACF [8] introduces a spatial alignment module, which is incorporated into a differentiable CF based network to address boundary effect. LADCF [9] is proposed to perform adaptive spatial feature selection to avoid boundary distortion, which considers the diversity and the redundancy of the entire feature input. Although the result is promising, there is still a need to capture prior assumptions regarding the object appearance. In this article, it is also witnessed that the use of the prior target content information has turned out...
to be extremely effective in numerous tough sequences. The saliency map can introduce the target content information, i.e. the shape and variation, which can serve as the prior information. Adopting the saliency-embedded adaptive spatial feature selection to optimize the filter, the tracker can enable compress sensing by selecting temporal consistency spatial feature. The adaptive feature selection could suppress the background to track the irregular, non-rigid and temporally changing target.

For online single object tracking, the appearance model is one of the most key components, on which extensive studies have been performed. At present, numerous trackers employ multi-channel hand-crafted features (such as Histogram of Oriented Gradients (HOG) [10], Color-Naming (CN) feature [11], histogram of local intensities), which have their own properties across different scenarios. Yuan et al. [12] explore a multiple feature fused model which can adaptively combine these advantages of different features perfectly. Yu et al. [13] discuss the possible strategy of information entropy weighted feature for preferable feature representation. Preferably, an appearance model should capture color features as well as texture information. Zhang et al. [14] propose a tracker, which computes Gabor responses on local patches and estimates the position of the target by finding the position with the maximal sum of responses of all Gabor kernels. Nevertheless, the bandwidth of Gabor filters is limited to one octave and the Gabor filter concentrates on the lower frequencies in feature extraction, which results in the lack of the high frequencies of the image. To tackle the limitations of traditional Gabor filter, an alternative to the Gabor function is the log-Gabor function proposed by Field. Bastos et al. [15] designs an iris recognition algorithm by using 2D Log-Gabor filters, and the experiments show that 2D log-Gabor filters is an effective alternative to encode the texture features. Inspired by the discussions above, in this study, an alternative representation using CF with log-Gabor filter to encode HOG, CN features, and log-Gabor features is presented and evaluated.

It is critical for the trackers to be self-recovered or reinitialized from failures. The re-detection step is an alternative strategy in case of tracking failure or occlusion. Several interesting trackers make use of several common techniques, such as support vector machine (SVM), boosting, random fern and so on. Dong et al. [16] propose a classifier-pool to identify whether the current tracking state is occlusion or not, and an appropriate tracking strategy is proposed for each tracking state. Ma et al. [17] use a discriminative correlation filter to estimate the confidence of current tracking state and train an online learned random forest classifier to re-detect the target. Thus, how to identify the current tracking state accurately is a critical problem, as re-detection at appropriate time will significantly improve the tracking accuracy. Therefore, one of the major aspects of this article is the motion estimation module. According to the occlusion estimating strategy, the motion information is used to predict or correct the position of the target.

The main contributions of this work can be summarized as follows:

1. A novel saliency guided adaptive spatial feature selection is proposed for spatial-temporal filter training enabling an optimal discriminative feature selection. On the basis of the adaptive spatial regularization, the proposed tracker can mitigate spatial boundary effect and background clutter.

2. The CF with log-Gabor filter containing comprehensive description of the target is proposed, which equips the tracker with a more robust feature representation across varieties of challenging attributes.

3. The motion estimation module is triggered by the color histogram-based occlusion estimating strategy. And the motion-aware strategy utilizes the Kalman filter to predict or correct the position of the target to prevent the tracker from being corrupted.

4. Qualitative and quantitative experiments on OTB-50, OTB-100 and UAV-123 have demonstrated that our approach provides promising visual tracking performance.

The remainder of the paper is organized as follows. Section II describes the related works of visual tracking. The pipeline of the proposed method is introduced in Section III. Section IV details several evaluations. Section V concludes this article.

II. RELATED WORKS

A. DISCRIMINATIVE CORRELATION FILTERS

Bolme et al. propose the minimum output of squared error (MOSSE) [17] that achieves good performance with an astonishing tracking speed of 669 frames per second. By employing the kernel trick and circulant matrix on the correlation filter, the kernelized correlation filter (KCF) [19] gets desirable success. KCF exploits the multi-channel HOG features to enhance the feature representation ability.Similarly, the color naming features are introduced to robust tracking. The learned CF-based trackers using high-dimensional deep features extracted from CNN have superior robustness in photometric and geometric variations [20]. To adapt to the scale changes, the discriminative scale space tracking (DSST) trains the filter using the scale pyramid [21]. Contrary to DSST which needs to train an extra correlation filter, Li et al. [22] propose a scale-adaptive module which is integrated into one correlation filter. The features of multiple scale targets are cropped from the original image which can decrease computational burden. SAMF addresses the scale adaptation problem using multi-scale searching scheme [23]. Danelljan et al. introduce a spatial regularization component to restrain correlation filter coefficients, which solves the boundary effects efficiently caused by periodic assumption [6]. BACF [24] exploits the variations of foreground and background to build a discriminative classifier. This not only alleviates boundary effect but also maintains real-time tracking speed. TFCR [25] proposes a target-focusing loss function to pay attention to target region, and simultaneously reduces the effect of background. The DCF-based tracker
has been extended by exploiting sophisticated tracking techniques, such as contextual information [26], [27], sparse representation [28], deep neural network [29].

**B. SPATIALLY REGULARIZED DISCRIMINATIVE CORRELATION FILTER**

SRDCF introduces a spatial regularization term within DCF to address the boundary effect. Given a $d$-dimension feature map $X$ extracted from a given image region, $Y^d \in \mathbb{R}^{M \times N}$ is the Gaussian shaped label. The SRDCF is formulated by minimizing the following objective,

$$
E(F) = \sum_{k=1}^{d} \| X^k \ast F^k - Y^k \|^2 + \sum_{k=1}^{d} \| W \odot F^k \|^2,
$$

where $\odot$ denotes the Hadamard product, $\ast$ denotes the convolution operator, $W$ is the regularization matrix, $F$ is the learned filter and $F^k$ is the $k$-th channel of $F$ accordingly. The regularization matrix penalizes the filter $F$ by assigning higher weights to the outside the target region and lower weights to the inside target region. The regularization matrix is constructed through the following equation:

$$
W(i, j) = \xi + \eta \left( \frac{i - x_0}{\omega} \right)^2 + \eta \left( \frac{j - y_0}{h} \right)^2,
$$

where $(i, j)$ stands for the coordinate in the search region, $(x_0, y_0)$ denotes the center of the search region, $(\omega, h)$ is the target size, while $\xi$ and $\eta$ are the fixed parameters.

**III. PROPOSED METHOD**

In this section, we will introduce the proposed algorithm that is adaptive to cope with various challenges. Section III.A mainly introduces saliency-embedded adaptive feature selection. Section III.B introduces temporal consistency saliency weight map. Section III.C introduces CF with log-Gabor filter. Section III.D introduces motion estimation module. The main step of the algorithm is illustrated in Fig. 1.

**A. SALIENCY-EMBEDDED ADAPTIVE FEATURE SELECTION**

On the basis of the saliency guided regularization, the adaptive feature selection process aims at selecting the optimal spatial mask in a lower dimensional manifold to avoid boundary distortion. Exiting spatial regularization methods only regularize the filter with simple pre-defined constraints without considering the target shape and variation. A saliency detection algorithm, similar to the human visual system with the ability of rapid significance detection, extracts the target in the frame [30]. Introducing the saliency prior information, the adaptive spatial feature selection is also embedded to the learned tracker.

The feature selection process selects specific elements in the filter, which can be formulated as:

$$
\theta_\psi = \text{diag} (\psi) \theta,
$$

where $\text{diag} (\psi)$ is the diagonal matrix generated from indicator vector of selected features $\psi$. The elements in $\psi$ are either 0 or 1 which means maintaining or discarding the corresponding element. It maintains spatial structure. The number of the selected spatial elements is determined by:

$$
M = D^2 \ast \sigma,
$$

where $D^2$ is the number of the spatial feature. $\sigma$ is the selected rate. $\theta_j$ denotes the $j$-th spatial element in the filter. The spatial
element selection is determined by sorting the $l_2$-norm of each spatial vector $\|\theta_l\|_2$ and preserving the top $M$ vectors.

The learning algorithm should be passive to make the updated filter similar to the previous one. Temporal smoothness means that the learned filter in successive frames should be changed smoothly along the timeline. As discussed above, a new objective function to minimize reconstruction error is proposed:

$$\arg\min_\theta \|\theta \ast x - y\|_2^2 + \lambda_1 \|W \odot \theta\|_2^2 + \lambda_2 \|\theta - \theta_{last}\|_2^2$$

$$+ \lambda_3 \|\psi\|_0 \quad \text{s.t.} \quad \theta = \psi = \text{diag}(\psi) \theta$$

$$\|\psi\|_0 = \|\theta\|_0,$$  \hspace{1cm} (6)

where $x$ denotes the multi-channel features, and $W$ denotes the spatial prior regularization weight map. The corresponding filter is regarded as $\theta$, and the learned filter in the previous frame is denoted as $\theta_{last}$.

$l_0$-norm is non-convex. At the meantime, $l_1$-norm can be used for sparsity [31]. The objective function can be converted into the following form:

$$\arg\min_\theta \|\theta \ast x - y\|_2^2 + \lambda_1 \|W \odot \theta\|_2^2 + \lambda_2 \|\theta - \theta_{last}\|_2^2$$

$$+ \lambda_3 \|\theta\|_1,$$  \hspace{1cm} (6)

$l_1$-norm can be calculated by the trick that $l_2$-norm for each location is calculated and then $l_1$-norm is performed. The objective function is converted into the following form:

$$\arg\min_\theta \sum_{k=1}^d \|\theta_k \ast x_k - y\|_2^2 + \lambda_3 \left\| \sum_{k=1}^d \theta_k \odot \theta_k \right\|_1$$

$$+ \lambda_1 \sum_{k=1}^d \|W \odot \theta_k\|_2^2 + \lambda_2 \sum_{k=1}^d \|\theta_k - \theta_{last,k}\|_2^2,$$  \hspace{1cm} (7)

where $d$ denotes the dimension of the feature map.

To fully exploit the convexity of Eq. (7), the model can be minimized to obtain an optimal solution using the alternating direction method of multipliers (ADMM):

$$\arg\min_\theta \sum_{k=1}^d \|\theta_k \ast x_k - y\|_2^2 + \lambda_3 \left\| \sum_{k=1}^d \theta'_k \odot \theta'_k \right\|_1$$

$$+ \lambda_1 \sum_{k=1}^d \|W \odot \theta_k\|_2^2 + \lambda_2 \sum_{k=1}^d \|\theta_k - \theta_{last,k}\|_2^2$$

$$\text{s.t.} \quad \theta'_k = \theta_k,$$  \hspace{1cm} (8)

The augmented Lagrangian form of Eq. (8) can be formulated as the following equation:

$$\arg\min_\theta \sum_{k=1}^d \|\theta_k \ast x_k - y\|_2^2 + \lambda_2 \sum_{k=1}^d \|\theta_k - \theta_{last,k}\|_2^2$$

$$+ \lambda_1 \sum_{k=1}^d \|W \odot \theta_k\|_2^2$$

$$+ \lambda_3 \left\| \sum_{k=1}^d \theta'_k \odot \theta'_k \right\|_1$$

$$+ \frac{\mu}{2} \sum_{k=1}^d \|\theta_k - \theta'_k + \frac{\eta_k}{\mu}\|_2^2,$$  \hspace{1cm} (9)

where $\mu$ denotes the penalty parameter and $\eta_k$ denotes the Lagrange multiplier.

By employing the ADMM algorithm, the equation should be divided into the following subproblems:

$$\hat{\theta}_k = \arg\min_\theta \sum_{k=1}^d \|\theta_k \ast x_k - y\|_2^2 + \frac{\mu}{2} \sum_{k=1}^d \|\theta_k - \theta'_k + \frac{\eta_k}{\mu}\|_2^2$$

$$+ \lambda_2 \sum_{k=1}^d \|\theta_k - \theta_{last,k}\|_2^2,$$  \hspace{1cm} (10)

Then Eq. (10) can be iteratively solved by solving the three subproblems. The solution to each subproblem is detailed as follows:

Subproblem $\theta$: Using the Parseval’s theorem, the first row of Eq. (10) can be rewritten in the Fourier domain as:

$$\hat{\theta}_k = \arg\min_\theta \|\hat{\theta}_k \odot \hat{x}_k - \hat{y}\|_2^2 + \frac{\mu}{2} \|\hat{\theta}_k - \hat{\theta}'_k + \frac{\hat{\eta}_k}{\mu}\|_2^2$$

$$+ \lambda_2 \|\hat{\theta}_k - \hat{\theta}_{last,k}\|_2^2,$$  \hspace{1cm} (11)

The model has a closed-form solution, and can be expressed as:

$$\hat{\theta}_k = \frac{\hat{x}_k \odot \hat{\theta}'_k + \lambda_2 \hat{\theta}_{last,k} + \frac{\mu}{2} \hat{\theta}'_k - \frac{\eta_k}{\mu}}{\hat{x}_k \odot \hat{\theta}'_k + \lambda_2 + \frac{\mu}{2}},$$  \hspace{1cm} (12)

Subproblem $\theta'$:

$$\hat{\theta}'_k = \arg\min_\theta \lambda_3 \left\| \sum_{k=1}^d \theta_k \odot \theta'_k \right\|_1$$

$$+ \lambda_1 \sum_{k=1}^d \|\theta_k - \theta_{last,k}\|_2^2$$

$$+ \frac{\mu}{2} \sum_{k=1}^d \|\theta_k - \theta'_k + \frac{\eta_k}{\mu}\|_2^2,$$  \hspace{1cm} (13)

where the spatial size of the feature map is $D \times D$, $\theta'_k$ denotes the $j$-th element in the $k$-th channel filter. To be more specific, Eq. (13) can be written as follows:

$$\hat{\theta}'_k = \arg\min_\theta \lambda_3 \left\| \sum_{k=1}^d \theta_k \odot \theta'_k \right\|_1$$

$$+ \lambda_1 \sum_{k=1}^d \|\theta_k - \theta_{last,k}\|_2^2$$

$$+ \frac{\mu}{2} \sum_{k=1}^d \|\theta_k - \theta'_k + \frac{\eta_k}{\mu}\|_2^2,$$  \hspace{1cm} (14)
constant $\mu$ in the previous state, $\theta\eta$ where:

$$
\text{strategy is introduced}
$$

To ensure the stability of model updating, the updating strategy to capture the latest changes to enhance the adaptability. Then

$$
The appearance of the target varies from frame to frame, $\theta\eta$:

Subproblem $\eta$: Lagrangian multipliers can be updated as:

$$
\eta_{k+1} = \eta_k \left( \theta_{k+1} - \theta'_{k+1} \right),
$$

where $\eta_k$ denotes the Fourier transform of the Lagrangian in the previous state, $\theta_{k+1}$ and $\theta'_{k+1}$ are the solutions to the two subproblems above at iteration $i + 1$. The regularization constant $\mu$ is commonly set as $\mu = \min (\mu_{\text{max}}, \beta \mu)$.

The appearance of the target varies from frame to frame, and the tracker needs to develop the proper model update strategy to capture the latest changes to enhance the adaptability. To ensure the stability of model updating, the updating strategy is introduced:

$$
\theta_k = (1 - \alpha) \theta_{\text{last}} + \alpha \theta_k,
$$

B. TEMPORAL CONSISTENCY SALIENCY WEIGHT MAP

The core idea of this section is the temporal consistency saliency weight map. It is introduced to ensure adaptability to target variations between successive frames. The difficulties are posed by a wide spectrum of appearance variations in unconstrained scenarios. Instead of fixed weight map disregarding appearance variation, the temporal consistency saliency weight map is introduced, which can deal with complex and changeable scenes. The process of introducing the saliency map into the regularization weight map to highlight the target and suppress the background is illustrated in Fig. 2.

The saliency map is simply multiplied with the conventional regularization weight map to get a new weight map to better reflect the shape of the target. The saliency detection method in [32] is adopted to detect the saliency within the search region and get a saliency map $S$. To make the saliency map work as the weight map, the saliency map can be adjusted as:

$$
W_c (i, j) = \frac{W_{s_{\text{max}}} - W (i, j)}{W_{s_{\text{max}}} - W_{s_{\text{min}}}},
$$

where $W_{s_{\text{max}}}$ denotes the maximum value of $W$, and $W_{s_{\text{min}}}$ denotes the minimum value of $W$. To put emphasis on the effect of saliency, the gap between the target region and the background should be widened. The saliency map can be adjusted as:

$$
W_s (i, j) = \begin{cases} 
W_c (i, j) \odot 10 & W_c (i, j) > \rho_1, \\
W_c (i, j) \odot 1/10 & W_c (i, j) < \rho_2, \\
W_c (i, j) & \text{otherwise}.
\end{cases}
$$

The saliency map is incorporated to the fixed regularization weight map $W_{\text{reg}}$ to obtain an improved regularization weight map which can suppress the background adaptively. The improved regularization weight map can be expressed as follows:

$$
W = W_s \odot W_{\text{reg}},
$$

FIGURE 2. The process of introducing the saliency map into the regularization weight map.
where $W_{reg}$ is the same as the regularization weight map of SRDCF.

It is robust to constrain the saliency map to be smooth across time dimension. In this article, the color histogram is extracted to judge the similarity of the two adjacent saliency regions. The color histogram of the search region in the current frame is denoted as $hist_k$, and the color histogram in the previous frame is denoted as $hist_{k-1}$. Bhattacharyya distance is utilized to measure the similarity $similarity_k$ of the two histograms. The weight is expressed as follows:

$$\omega_k = [\sin(\pi \cdot similarity_k - \pi/2)]/6, \quad (23)$$

As discussed above, the saliency map can be obtained via

$$W_{sk} = \omega_k W_{sk-1} + (1 - \omega_k) W_{sk}, \quad (24)$$

C. CF WITH LOG-GABOR FILTER

In order to address the problem that the single feature cannot contain the complementary description of the target and cannot adapt to the complex situation variation, in this section, CF with log-Gabor filter is presented to integrate HOG feature, CN feature, and log-Gabor feature.

HOG feature describes the abundant gradient characteristics. It maintains a good invariance of the photometric and local geometric transformations. CN feature can capture rich color characteristics. It has less dependence on the deformation, rotation and scale change. Image texture features reflect spatial structure information of the target and the relationship between the target and the surrounding background. Introducing the texture feature to gain a robust feature presentation is beneficial to visual tracking. The log-Gabor filter is a band-pass filter for detecting texture and edges, which is sensitive to orientations, scale information and different frequencies. A typical alternative idea to encode the texture information statistically is multi-orientation and multi-scale log-Gabor filter. It can describe the frequency of images more accurately, and it is consistent with measurements on human visual systems.

Two-dimensional log-Gabor filter is widely used as a tool for feature extraction in the field of computer vision. The frequency response of Log-Gabor filter in polar coordinate is defined as

$$LG(f, \theta) = \exp \left( - \log \left( \frac{f/f_m}{2} \right)^2 \right) \exp \left( - \frac{(\theta - \theta_n)}{2\sigma_\theta^2} \right), \quad (25)$$

where $f$ and $\theta$ denote the frequency and angle, $f_m$ is the center frequency, $\sigma_f$ denotes the width parameter for frequency. $\theta_n$ is the central angle and $\sigma_\theta$ is the width parameter of the angle.

The details of the construction regarding the log-Gabor filter are described in [33]. The log-Gabor feature is the convolution of the image and the log-Gabor filter. Given an input image $I(x, y)$, where $(x, y)$ represents the pixel location, the convolution results is shown as below:

$$A(x, y) = I(x, y) * LG, \quad (26)$$

Then, the log-Gabor feature is regarded as the texture feature representation.

The HOG feature is $x_{hog} = [x_{hog_1}, x_{hog_2}, \cdots, x_{hog_d}]$ (where $d1$ is the number of channels of the HOG feature), the CN feature is $x_{cn} = [x_{cn_1}, x_{cn_2}, \cdots, x_{cn_d}]$ (where $d2$ is the number of channels of the CN feature), and the log-Gabor feature is $x_{log-Gabor} = [x_{log-Gabor_1}, x_{log-Gabor_2}, \cdots, x_{log-Gabor_d}]$ (where $d3$ is the number of channels of the log-Gabor feature). It can be expressed as $x = [x_{hog}, x_{cn}, x_{log-Gabor}]$ by cascading these three features as an enhancement to feature representation.

D. MOTION ESTIMATION MODULE

In this study, the motion estimation module activated by the occlusion estimating strategy is integrated to optimize the possible position. The tracker will continuously learn the wrong samples when the target is severely occluded or missing. In this situation, most existing trackers fail to recover from occlusion since the filter is contaminated due to inappropriate updating on noisy samples. However, motion information could contribute to the localization.

Histograms for different objects can be different markedly. The color histogram of the target region in the first frame is denoted as $H_1$. The color histogram of the target region located by the tracker mentioned in Section III. A is denoted as $H_t$. The similarity of the two histograms is proposed to reflect the occlusion level of the target. Bhattacharyya distance is introduced to measure the similarity.

$$d(H_1, H_t) = \sqrt{1 - \frac{1}{\sqrt{H_1 H_t N^2}} \sum_i \sqrt{H_1(i) H_t(i)}}, \quad (27)$$
where \( N \) denotes the number of bins in the color histogram, 
\[
\hat{H}_k = \frac{1}{N} \sum_l H_k (l).
\]
The smaller the value is, the smaller the occlusion level is.

The parameter of position is initialized to \((x_1, y_1)\). The velocity and acceleration \((v_{x1}, v_{y1}), (a_{x1}, a_{y1})\) are set to \((0, 0)\) in the first frame.

The possible location of the target in the second frame can be expressed as follows:
\[
\begin{aligned}
    x_{t+1} &= x_t + v_{x_t} \Delta t + 1/2 * a_{x_t} * \Delta t^2, \\
    y_{t+1} &= y_t + v_{y_t} \Delta t + 1/2 * a_{y_t} * \Delta t^2.
\end{aligned}
\]
(28)

The velocity in the second frame \((v_{x2}, v_{y2})\) is updated by Eq. (29).
\[
\begin{aligned}
    v_{x_t} &= x_t - x_{t-1}, \\
    v_{y_t} &= y_t - y_{t-1}, \\
    a_{x_t} &= v_{x_t} - v_{x_{t-1}}, \\
    a_{y_t} &= v_{y_t} - v_{y_{t-1}}.
\end{aligned}
\]
(29)

The acceleration in the third frame \((a_{x2}, a_{y2})\) is updated by Eq. (30).

\[
\begin{aligned}
    v_{x_t} &= x_t - x_{t-1}, \\
    v_{y_t} &= y_t - y_{t-1}, \\
    a_{x_t} &= v_{x_t} - v_{x_{t-1}}, \\
    a_{y_t} &= v_{y_t} - v_{y_{t-1}}.
\end{aligned}
\]
(30)

where \((x_t, y_t)\) is the position detected by the tracker proposed in Section III.A. The scheme to estimate instantaneous velocity and acceleration among three contiguous frames is shown in Fig. 4.

\[
\begin{array}{c}
(a) \quad (b) \\
\hline
(c) \quad (d) \\
\hline
\end{array}
\]

FIGURE 4. Instantaneous motion estimation.

The location in the following frame \((x_{t1}, y_{t1})\) can be predicted by Eq. (28), and the position can be served as the center of the search region of correlation filter.

Assuming that the motion model of the target is a constant acceleration model, the Kalman filter [34] can be described to predict the position.

\[
P (t|t-1) = AP (t-1|t-1) + BM (t),
\]
(31)

\[
C (t|t-1) = AC (t-1|t-1) A' + Q,
\]
(32)

\[
K = C (t|t-1) H' \\
\]
\[
= HC (t|t-1) H' + R,
\]
(33)

\[
P (t|t) = P (t|t-1) + K [N (t) - HP (t|t-1)],
\]
(34)

\[
C (t|t) = (I - KH) C (t|t-1),
\]
(35)

where \( P (t|t-1) \) is the current predicted position of the target, and \( P (t-1|t-1) \) denotes the result of the previous state optimization. \( C (t|t-1) \) is the covariance corresponding to \( P (t|t-1) \), similarly \( C (t-1|t-1) \) is covariance corresponding to \( P (t-1|t-1) \). \( A' \) denotes the transpose matrix of \( A \) and \( \overline{Q} \) represents the covariance of the noise. \( M (t) \) denotes the motion model. \( K \) is the Kalman gain in the current frame and \( H' \) regards the transpose matrix of \( H \), and \( R \) denotes the measuring error.

If the occlusion level \( d_t \) is larger than an adaptive threshold, \( \tau = 1.4 * \frac{1}{T} \sum_{k=1}^{T} d_k \), there is a possibility that the target is severely occluded or contaminated. The location predicted by the Kalman filter \( P (t|t-1) \) is considered as the tracking result.

Blindly updating correlation filters would cause the tracker to learn wrongly detected objects, thus using a smaller learning rate is necessary. The learning rate of the filter is updated by:

\[
\alpha_t = \alpha \exp \left( \frac{1}{T} \sum_{k=1}^{T} d_k \right) / d_t - 1.
\]
(36)

If the occlusion level \( d_t \) is lower than the threshold, it indicates that there is no model drift detected. In this case, the basic tracker could continuously perform the tracking task, and the Kalman filter will be utilized to correct the location.

The position of the target detected by the tracker mentioned in Section III.A is used as the measurement value \( N (t) \). The optimal estimation of the current position \( P (t|t) \) is achieved.

IV. EXPERIMENTS

The experiments on three benchmark datasets (OTB-50, OTB-100 and UAV-123) are organized as follows: (1) Experimental parameters setting is introduced in Section IV.A; (2) The quantitative evaluation and qualitative evaluation are introduced in Section IV.B and Section IV.C, respectively; (3) A ablation study is complemented in Section IV.D.

A. EXPERIMENTAL PARAMETERS SETTING

The experiment were conducted on a PC with an i5 2.50GHz CPU with 4GB memory using MATLAB R2017b. The size
FIGURE 6. The precision plots and success plots of the proposed tracker compared with other trackers on different attributes on OTB-100 database.
of hog grid cell is 4. The number of hog orientations is 9. The regularization parameters $\lambda_1$, $\lambda_2$, and $\lambda_3$ are set to 0.0001, 15, and 0.03, respectively. The initial penalty parameter $\mu$ is set to 20. The maximum iteration of ADMM is 2. The coefficient of enhanced weight map $\rho_1$ and $\rho_2$ is set to 0.8 and 0.3, respectively. The selected rate $\sigma$ is set to 5%, and the initial learning rate $\alpha$ is set to 0.95.

B. QUANTITATIVE EVALUATION

The proposed tracker is evaluated on OTB-100 database and compared with other state-of-the-art trackers, including GFS-DCF [35], SiamRPN [36], SiamFC [37], DaSiamRPN [38], LADCF [9], ECO_HC [4], C_COT [39], BACF [24], SRDCF [6], using distance precision plots and overlap success plots. The distance precision is regarded as the percentage of images where the distance between the predicted location and the ground truth is no more than 20 pixels. The overlap success is defined as the percentage of images whose overlap is beyond 0.5. The precision plots and the success plots obtained by using fully integrated OTB toolkit are shown in Fig. 6. Among these trackers, the proposed tracker obtains the best performance with the average success rate of 68.4%, which achieves a gain of {2.8%, 3.7%, 4.2%} over LADCF, C_COT, DaSiamRPN. Among the state-of-the-art trackers, the proposed tracker provides the best result with a precision rate, and a gain of {1.2%, 1.8%, 1.9%} over C_COT, LADCF, DaSiamRPN. An obvious performance level-up is achieved in the proposed tracker thanks to the filter learning based on adaptive spatial feature selection, which can put emphasis on the relevant feature and maintain the spatial structure.

To further evaluate the performances of trackers, the quantitative analysis of 11 attributes summarized in the OTB-100 database is conducted. The challenging factors, such as occlusion, and...
Algorithm 1 The Overall Tracking Algorithm

Input: the \( t^{th} \) frame video \( I_t \), the initial predicted position by Eq. (28) \( (x_t, y_t) \), the filter model in previous frame \( \theta_{t-1} \)

Output: the position in the current frame \( (x_t, y_t) \), the filter model in the current frame \( \theta_t \)

for \( t = 2 \) to end do

Crop out searching window in frame \( t \) centered at the predicted position \( (x_t, y_t) \) and extract features \( x \) via CF with log-Gabor filter;

Get the temporal consistency saliency map in the search region \( S \) and gain the improved regularization weight map \( W \) by Eq. (20), (21), (22);

Convolute learned filter \( \theta_{t-1} \) with \( x \) and detect the position;

Compare the color histogram of the predicted target with the color histogram of the target in the first frame and estimate the occlusion level \( d_t \).

if \( d_t > \tau \), then

the Kalman filter predicts the position \( (x_t, y_t) \) by Eq. (31) and the learning rate is updated by Eq. (36);

else

the Kalman filter corrects the position \( (x_t, y_t) \) by Eq. (34).

end

According to predicted position \( (x_t, y_t) \) in frame \( t \), crop out searching window in frame \( t \) and extract features via CF with log-Gabor filter;

Train the saliency guided filter based on the adaptive spatial feature selection by Eq. (12), (17), (18);

Update the filter model \( \theta_t \) by Eq. (19).

end

scale variation can be reflected by 11 attributes contained in the OTB-100 database. As can be seen from Fig. 6, the proposed tracker achieves the optimal performance on almost all attributes. The proposed tracker shows superior performance in dealing with occlusion. The proposed tracker performs favorably against LADCF (by 2.5%). The conventional trackers utilize a passive strategy to update the CF and track the target in the following frame, which contributes to overfitting to the recent polluted samples and tracking failure. The proposed tracker takes advantage of the motion estimation module to predict the position of the target and prevents from the model drift. These results demonstrate that our tracker is able to locate the target in these complicated scenarios more precisely than the other state-of-the-art trackers.

Fig. 7 shows the experimental results evaluated on the OTB-50 database. In the success plot, the proposed tracker precisely than the other state-of-the-art trackers.

The proposed tracker has achieved remarkable performance on the UAV-123 benchmark which is a uniform aerial video benchmark, as illustrated in Fig. 8. The precision rate gains the performance by 75.1%. Besides, the precision rate of proposed tracker increases by 1.7%, 2.6%, and 3.2% in comparison with those of SiamRPN, LADCF, and C_COT respectively. Our method achieves a substantial improvement over SiamRPN, LADCF, and C_COT, with a gain of 1.1%, 1.8%, and 1.9% in AUC. The comparison supports the effectiveness of the proposed method.

C. QUALITATIVE EVALUATION

Fig. 9 illustrates the qualitative comparisons of the proposed tracker compared with classical trackers carried on the OTB-100 dataset. A study on the sequence dragonbaby (the first row of Fig. 9), the appearance of the baby is changed severely (e.g. #27, #30) and only the proposed method can track the baby successfully (e.g. #34). A case study is carried out on the skating1 dataset (the second row of Fig. 9), the skater undergoes deformation as it moves (e.g. #78) and is occluded by others severely (e.g. #178, #180). Owing to the use of motion information, the tracker can track the skater. Benfitting from the motion estimation module, even when the baby undergoes fast motion (e.g. #48), motion information guarantees the robustness of the tracker. In the sequence freeman4 (the third row of Fig. 9), the box can be tracked well although it is occluded in a long term (e.g. #66, #69).

The proposed tracker, can track it continuously after leaving the obstruction (e.g. #70). The target in sequence panda (the fourth row of Fig. 9) has significant appearance variation as

FIGURE 7. The precision plots and success plots of the proposed tracker compared with other trackers on different attributes on OTB-50 database.

FIGURE 8. The precision plots and success plots of the proposed tracker compared with other trackers on different attributes on UAV-123 database.
it moves and turns (e.g. #147, #382), and most trackers cannot track it precisely except for the proposed tracker. The box undergoes severe deformation due to rotation, while the proposed tracker can still identify the box through the saliency prior information.

D. ABLEATION STUDY
The proposed tracker contains three important components: (1) the saliency-embedded adaptive feature selection; (2) CF with log-Gabor filter; (3) motion estimation module. To evaluate their separate contributions to our tracker, we implement several variants of our tracker.

We denote the tracker without the saliency-embedded adaptive feature selection as OURS_SAL, without the CF with log-Gabor filter as OURS_GAR, and without the motion estimation module as OURS_MOD. The comparison results are shown in Fig. 10. The precision plot of our tracker decreases by 1%, most likely because OURS_SAL does not consider the saliency map as prior target information. Benefitting from the credibility of the foreground object as prior information, adaptive spatial feature selection contributes to avoiding the disturbance of target losing. The precision rate of OURS_GAR is lower than OURS by 1.4%. It is due to the fact that OURS_GAR does not introduce log-Gabor feature, which can not capture the abundant texture feature of the target. The lack of texture feature leads to performance degradation. The performance of our tracker is certainly be enhanced compared with OURS_MOD, which benefits from motion information. Once the state of the target is detected as occlusion, the Kalman filter is utilized to predict the location.

V. CONCLUSION
In this article, we propose a robust visual tracking method based on the saliency-embedded adaptive spatial feature selection to handle occlusion and boundary effect challenges. CF with log-Gabor filter containing a comprehensive description of the target has been systematically proposed to deal with varieties of challenging attributes. On the one hand, the saliency map serves as the prior information of adaptive
spatial feature selection to suppress the boundary effect. The saliency map takes the shape and variation of the target into consideration. By design, spatial-temporal feature selection highlights the relevant feature and reduces the information redundancy. On the other hand, the motion estimation module utilizes the motion information to predict or correct the optimal position of the target to avoid the passive update of the filter, which will certainly be a significant boost in the tracking paradigm.

However, it is interesting to explore how to balance the tradeoff between the accuracy and efficiency. When there is more than one target in the scene, the relationship among the targets should be taken account of. The improvement and research can be performed in the future work.

REFERENCES

[1] T. Zhang, S. Liu, N. Ahuja, M.-H. Yang, and B. Ghanem, “Robust visual tracking via consistent low-rank sparse learning,” Int. J. Comput. Vis., vol. 111, no. 2, pp. 171–190, Jan. 2015.

[2] Z. Kalal, K. Mikolajczyk, and J. Matas, “Learning-tracking-detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 7, pp. 1409–1422, Jan. 2018.

[3] 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.

[4] 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.

[5] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, “Learning dynamic Siamese network for visual object tracking,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1763–1771.

[6] M. Danelljan, G. Hager, 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.

[7] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, “Learning spatial-temporal regularized correlation filters for visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4904–4913.

[8] M.-D. Zhang, Q. Wang, J.-L. Xing, J. Gao, P.-X. Feng, W.-M. Hu, and S. Maybank, “Visual tracking via spatially aligned correlation filters network,” in Proc. Europ. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 484–500.

[9] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, “Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking,” IEEE Trans. Image Process., vol. 28, no. 11, pp. 5596–5609, Nov. 2019.

[10] D. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2005, pp. 886–893.

[11] M. Danelljan, F. S. Khan, M. Felsberg, and J. V. D. Weijs, “Adaptive color attributes for real-time visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1090–1097.

[12] D. Yuan, X. Zhang, J. Liu, and D. Li, “A multiple feature fused model for visual object tracking via correlation filters,” Multimedia Tools Appl., vol. 78, no. 19, pp. 27271–27290, Oct. 2019.

[13] M.-X. Yu, Y.-H. Zhang, Y.-K. Li, J.-Z. Li, and C.-L. Wang, “Distractor-aware long-term correlation tracking based on information entropy weighted feature,” IEEE Access, vol. 8, pp. 29417–29429, 2020.

[14] S. Zhang, X. Lan, Y. Qi, and P. C. Yuen, “Robust visual tracking via basis matching,” IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 3, pp. 421–430, Mar. 2017.

[15] C. A. C. M. Bastos, T. I. Ren, G. D. C. Cavalcanti, “Iris segmentation and recognition using 2D log-Gabor filters,” in Proc. Int. Conf. Intell. Data Eng. Automated Learn., Aug. 2012, pp. 443–450.

[16] X. Dong, J. Shen, D. Yu, W. Wang, J. Liu, and H. Huang, “Occlusion-aware real-time object tracking,” IEEE Trans. Multimedia, vol. 19, no. 4, pp. 763–771, Apr. 2017.
MINGXIN YU received the B.S. degree from the Shijiazhuang Campus of Army Engineering University, Shijiazhuang, China, where she is currently pursuing the Ph.D. degree in unmanned aerial vehicle engineering. Her main research interests include computer vision, visual tracking, and machine learning.

YUHUA ZHANG received the B.S. degree from the Hubei University of Technology, Wuhan, China, and the M.S. and Ph.D. degrees from the National University of Defense Technology, Changsha, China. She is currently working with the Department of Unmanned Aerial Vehicle Engineering, Shijiazhuang Campus of Army Engineering University, Shijiazhuang. Her main research interest includes UAV-assisted image processing.

YONGKE LI received the B.S. and M.S. degrees from the Ordnance Engineering College, Shijiazhuang, China. He is currently an Associate Professor with the Department of Unmanned Aerial Vehicle Engineering, Shijiazhuang Campus of Army Engineering University, Shijiazhuang. His main research interests include pattern recognition and scene classification.

ZHI-LONG LIN received the B.S. and M.S. degrees from the Shijiazhuang Campus of Army Engineering University, Shijiazhuang, China. He is currently an Assistant Professor with the Department of Unmanned Aerial Vehicle Engineering, Shijiazhuang Campus of Army Engineering University. His current research interests include object detection, visual tracking, and deep learning.

JIANZENG LI received the B.S. and M.S. degrees from the Ordnance Engineering College, Shijiazhuang, China. He is currently an Associate Professor with the Department of Unmanned Aerial Vehicle Engineering, Shijiazhuang Campus of Army Engineering University, Shijiazhuang. His current research interests include intelligent system and pattern recognition.

CHANGLONG WANG received the B.S. degree in radar engineering from the Electronic Engineering College, Hefei, China, in 1987, the M.S. degree in signal and information processing from the Harbin Institute of Technology, Harbin, China, in 1992, and the Ph.D. degree in electronics and information technology from Xi’an Jiaotong University, Xi’an, China, in 2006. He is currently a Professor with the Department of Unmanned Aerial Vehicle Engineering, Shijiazhuang Campus of Army Engineering University, Shijiazhuang, China. His main research interests include pattern recognition and signal processing.