Mutation Sensitive Correlation Filter for Real-Time UAV Tracking with Adaptive Hybrid Label

Guangze Zheng, Changhong Fu*, Junjie Ye, Fuling Lin, and Fangqiang Ding

Abstract— Unmanned aerial vehicle (UAV) based visual tracking has been confronted with numerous challenges, *e.g.*, object motion and occlusion. These challenges generally introduce unexpected mutations of target appearance and result in tracking failure. However, prevalent discriminative correlation filter (DCF) based trackers are insensitive to target mutations due to a predefined label, which concentrates on merely the centre of the training region. Meanwhile, appearance mutations caused by occlusion or similar objects usually lead to the inevitable learning of wrong information. To cope with appearance mutations, this paper proposes a novel DCF-based method to enhance the sensitivity and resistance to mutations with an adaptive hybrid label, *i.e.*, MSCF. The ideal label is optimized jointly with the correlation filter and remains temporal consistency. Besides, a novel measurement of mutations called mutation threat factor (MTF) is applied to correct the label dynamically. Considerable experiments are conducted on widely used UAV benchmarks. The results indicate that the performance of MSCF tracker surpasses other 26 state-of-the-art DCF-based and deep-based trackers. With a real-time speed of ~38 frames/s, the proposed approach is sufficient for UAV tracking commissions.

I. INTRODUCTION

In recent years, visual object tracking has facilitates the applications of UAV, especially in path planning [1], traffic monitoring [2], and global pose estimation [3]. In the process of UAV object tracking, the target is specified primarily and tracked in the subsequent frames. However, the volatile environment as well as the object motion usually causes severe uncertainties. Besides, the limitation of computing resource and battery life on the aerial platform also impede the development of UAV-based tracking.

To solve the above problems, DCF-based tracking method has been introduced due to high computational efficiency and promising accuracy [5]. Since the proposition of KCF in [6], meaningful improvements have been made by follow-up researchers [7]–[9]. The general pipeline of DCF-based UAV tracking is summarized as follows. The current feature model is updated by the linear interpolation of previous one and current extracted feature with a learning rate. Subsequently, through the ridge regression, a filter is trained by minimizing the squared error of the correlation response with the current feature model. As for position estimation, a new frame is transmitted to UAV and processed with the filter to detect the object for continuous tracking. Since the computation is transformed into the frequency domain through fast discrete Fourier transform (DFT), the computation speed is considerably improved.

*Corresponding Author
The authors are with the School of Mechanical Engineering, Tongji University, 201804 Shanghai, China. changhongfu@tongji.edu.cn
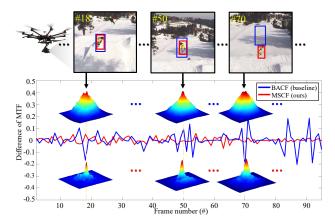
Fig. 1. The comparison of tracking performance between the presented MSCF tracker and baseline, BACF tracker [4]. The red and blue boxes are the tracking results from MSCF tracker and BACF tracker respectively. The difference of the proposed MTF metric between frames evaluates the change of mutations, where higher value means drastic transformation. The red curve with small fluctuations demonstrates that MSCF tracker possesses higher and stronger mutation sensitivity and anti-mutation power than BACF tracker.

Although DCF-based UAV tracking has gained excellent performance, there are still some problems caused by tracking mutations. First, the frequently-used Gaussian label pays most of attention to the centre of training region, but the object generally moves to the neighborhood in case of fast motion. Meanwhile, since target mutations, such as similar objects, are manifested as sub-peaks in the response map, they commonly represent severe threats to the detection of the target. Therefore, how to construct an adaptive label and improve the robustness of the tracker against the mutations has a become a crucial issue.

Based on the analysis above, this work proposes a UAV tracking approach with a mutation sensitive correlation filter (MSCF). Main work-flow of MSCF adopts the idea of genetic algorithm, which includes selection, crossover, and evolution. Response maps and peaks in them are regarded as chromosomes and genes respectively. In this way, the mutations can be detected through the proposed MTF measurement and the dangerous threats are consequently repressed. During the crossover phase, three labels, *i.e.*, ideal hybrid label, previous label, and current correlation value label, are hybridized to form an adaptive label. The filter evolves afterwards for object detection in the next frame. A qualitative comparison between MSCF and its baseline BACF [4] is shown in Fig. 1, where the tracking object is undergoing drastic deformation in appearance, and mutations are reflected as sub-peaks in response maps. The tracking procedure of MSCF is illustrated in Fig. 2. Main
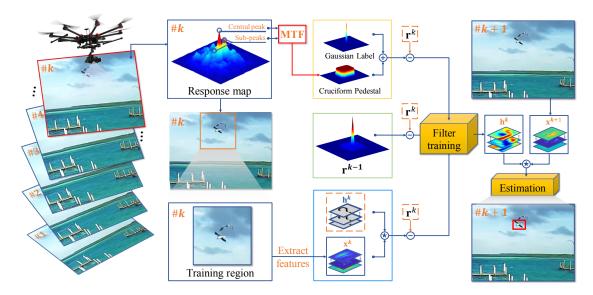
Fig. 2. Tracking procedure of the proposed MSCF tracker. Dashed boxes denote the variables to be solved in the main regression. As MTF in the **red** box is generated from search region in frame k, it is applied to adjust the altitude value of the cruciform pedestal. Besides, signal $-$ denotes the minus operation in the objective function, and $\star$ denotes circular correlation.

contributions of this work can be summarized as follows:

- A novel assessment measure of response maps is proposed, namely mutation threat factor (MTF), and the idea of genetic algorithm is adopted.
- A novel adaptive hybrid label with MTF adjustment is applied for the filter training.
- A novel MSCF tracker is developed to cope with tracking mutations by employing the adaptivity of the adaptive hybrid label. Numerous experiments are conducted on extensively-used UAV benchmarks. The results demonstrate the superior performance of the MSCF tracker against other 26 state-of-the-art trackers.

II. RELATED WORK

A. DCF for UAV Tracking

Since a prior work [10] introduced correlation filter into visual tracking, methods based on correlation filter have raised a paradigm shift. Subsequently, J. F. Henriques *et al*. [6] contributed a basic framework to DCF-based methods. Employments of scale estimation [11] and real background information [4] have also brought about further progress of DCF's performance. Meanwhile, efficient feature extraction [12], [13] contributes to more precise and comprehensive object representation. Nevertheless, standard DCFs commonly use a predefined Gaussian label for regression. These trackers are still weak to handle tracking mutations and prone to collapse as the wrong information is learned.

B. Adaptive Label for DCF

The mutation-adaptive potential is significant for UAV tracking, especially in extreme situations from aerial views. Nevertheless, the attention paid to adaptive label learning was insufficient. A. Bibi *et al*. [14] proposed a generic framework which adaptively changed the labels from frame to frame. R. Han *et al*. [15] improved the the label adaptivity by assigning different weights to the training samples. However,

the potential of the adaptive label has not been achieved thoroughly since the optimization of labels are still lack of direction. The deficiency of the label which merely focuses on the object centre is also unsolved due to the inflexibility of the label shape.

C. Evaluation of Response Map

Previous studies have proposed several methods to evaluate the credibility of response maps. D. S. Bolme *et al*. [10] took the advantage of the peak to sidelobe ratio (PSR) for failure detection. M. Wang *et al*. [16] proposed the average peak-to-correlation energy (APCE) to measure the fluctuated degree of response maps. Y. Han *et al*. [17] discovered the application of kurtosis to measure the peakedness and tail weight of the probability distributions of the sampled data. Whereas, certain challenges, *e.g*, similar objects or partial occlusion, generally bring specific mutations which are represented as sub-peaks in the response maps. These methods show insufficient concern for particular dangerous sub-peaks and thus are insensitive to mutations.

III. PROPOSED METHOD

To achieve higher adaptability to tracking mutations, the genetic idea is adopted in the process of tracking. Labels are regarded as chromosomes while the peaks in the response map are considered as genes. Since dangerous mutations are picked by employing MTF in selection phase, the labels are then hybridized into the adaptive hybrid label in the crossover phase. Finally, the filter and the label are trained in the evolution phase.

A. Selection Phase: Mutation Sensitivity

When the target appearance suffers from mutations, the response maps will fluctuate severely and consequently result in sub-peaks. In this section, the mutation threat factor (MTF) is proposed to represent the threats of sub-peaks
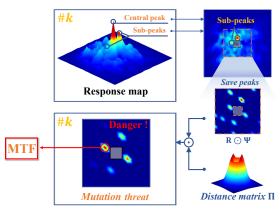
Fig. 3.   Main work-flow of MTF generation.

in the response map and the process is demonstrated in Fig. 3, where mutations are demonstrated as sub-peaks in the response map. Through the application of a distance matrix $\boldsymbol{\Pi}$, MTF is evaluated and utilized in the adaptive hybrid label in selection phase. The principle of selection phase is to find the most dangerous sub-peak to demonstrate the severity of current mutation. During the process of tracking, the threat of sub-peaks in the k-th frame is measured as:

$$\mathbf{M}^k = \frac{\mathbf{R}^k \odot \boldsymbol{\Psi}^k}{R_{max}^k} \odot \boldsymbol{\Pi} \quad , \tag{1}$$

where k denotes the k-th frame, $\mathbf{R}$ indicates the response map among the search region, and R_{max}^k is the maximum value in $\mathbf{R}$. Besides, $\boldsymbol{\Psi}$ represents a binary matrix to show the sub-peaks in the response map. MTF in k-frame is the maximum value of the calculated $\mathbf{M}^k$. $\boldsymbol{\Pi}$ is defined as:

$$\boldsymbol{\Pi} = \begin{cases} \dfrac{\nu}{1 + \delta \cdot \exp(d_{(i,j)})}, & d_{(i,j)} > d_{min} \\ 0 & , \quad d_{(i,j)} \le d_{min} \end{cases} , \tag{2}$$

where ν denotes the weights in $\boldsymbol{\Pi}$ while δ describes the influence degree of the distance. Both parameters are defined in advance. $d_{(i,j)}$ indicates the distance from the point with i-th row and j-th column to the centre in the matrix. Further distance from the centre leads to lower value in $\boldsymbol{\Pi}$. On account that the centre peak of the response map can affect the search of sub-peaks, the weights in the central area are set to 0 with a threshold d_{min}.

According to the definition above, the threat from mutations can be expressed elaborately by MTF and the feedback of MTF will be immediately utilized in the adaptive hybrid label.

B. Crossover Phase: Adaptive Hybrid Label

The proposed adaptive hybrid label $\mathbf{r}^k$ in the k-th frame is composed of three parts, *i.e.*, ideal hybrid label, previous label, as well as the circular correlation value in the training region. Through the crossover of the three components, the adaptability of label can realize regularization in both temporal and spatial domains.

Spatial regularization: ideal hybrid label. The optimized label consists of two parts: (i) a sharp peak from the Gaussian label; (ii) a cruciform pedestal in the bottom. As for the pedestal altitude value, the generated MTF is introduced to

adaptively adjust the value in case of mutations. Meanwhile, the size of the pedestal is initialized according to the target scale and a predefined size ratio parameter. Considering the appearance mutations, the optimal adaptive hybrid label $\boldsymbol{\Omega}$ is defined as:

$$\boldsymbol{\Omega} = \mathbf{y}_1 + (1 - \theta \cdot \max(\mathbf{M}^k))\mathbf{y}_2 \quad , \tag{3}$$

where $\mathbf{y}_1$ and $\mathbf{y}_2$ respectively represent Gaussian label and cruciform pedestal label. θ is a predefined coefficient, while $\max(\mathbf{M}^k)$ indicates the MTF.

Temporal regularization: previous label. To guarantee the label not to be influenced excessively by particular mutations, the consistency of the hybrid label is updated by referring to the label in the previous frame. The consistency of the adaptive hybrid label is also taken into account in objective optimization, which is demonstrated as the fourth term in Eq. (4).

Hybridization procedure: Different from conventional trackers, correlation filter $\mathbf{h}$ and the hybrid label $\mathbf{r}$ are trained alternately in the proposed method. During the training process in the k-th frame, the interaction of the filter and hybrid label contributes to a more rapid convergence. Therefore, the filter and the label have formed a mutual constraint relationship, which promotes the robustness of the tracker.

C. Evolution Phase: Objective Function

Based on the aforementioned two phases, the adaptability to mutations of the filter and the label can be realized through the evolution phase. Both the filter and the label evolve for higher adaptivity to mutations in solving the problem. The objective function of MSCF can be written as:

$$\begin{aligned} \mathcal{E}(\mathbf{h}^k, \mathbf{r}^k) = & \frac{1}{2} \left\| \mathbf{r}^k - \sum_{d=1}^{D} (\mathbf{P}\mathbf{h}_d^k) \star \mathbf{x}_d^k \right\|_2^2 \\ & + \frac{\lambda_1}{2} \sum_{d=1}^{D} \left\| \mathbf{h}_d^k \right\|_2^2 + \frac{\lambda_2}{2} \left\| \boldsymbol{\Omega}^k - \mathbf{r}^k \right\|_2^2 \quad , \\ & + \frac{\phi}{2} \left\| \mathbf{r}^k - \mathbf{r}^{k-1} \right\|_2^2 \end{aligned} \tag{4}$$

where d denotes the d-th feature channel, k indicates the k-th frame, $\mathbf{x} \in \mathbb{R}^T$ denotes the vectorized image, and $\mathbf{h} \in \mathbb{R}^T$ indicates the filter. Besides, $\mathbf{r}^k$ denotes the adaptive hybrid label. $\mathbf{P} \in \mathbb{R}^{N \times T}$ is a binary matrix which crops the mid N elements of $\mathbf{x}$, where $T \gg N$. T denotes the length of $\mathbf{x}$. In addition, λ_1, λ_2, and ϕ are regularization parameters. After transformed into frequency domain for computational efficiency, Eq. (4) can be expressed as:

$$\begin{aligned} \hat{\mathcal{E}}(\mathbf{h}^k, \mathbf{r}^k, \hat{\mathbf{g}}^k) = & \frac{1}{2T} \left\| \hat{\mathbf{r}}^k - \hat{\mathbf{X}}^k \hat{\mathbf{g}}^k \right\|_2^2 + \frac{\lambda_1}{2} \left\| \mathbf{h}^k \right\|_2^2 \\ & + \frac{\lambda_2}{2T} \left\| \hat{\boldsymbol{\Omega}}^k - \hat{\mathbf{r}}^k \right\|_2^2 + \frac{\phi}{2T} \left\| \hat{\mathbf{r}}^k - \hat{\mathbf{r}}^{k-1} \right\|_2^2 \quad , \\ s.t. \quad & \hat{\mathbf{g}}^k = \sqrt{T}(\mathbf{I}_D \otimes \mathbf{F}\mathbf{P}^{\mathsf{T}})\mathbf{h}^k \end{aligned} \tag{5}$$

where $\hat{\mathbf{g}}^k$ is an auxiliary variable vector and $\mathbf{F} \in \mathbb{C}^{T \times T}$ denotes the orthonormal matrix. $\hat{\mathbf{X}}^k$ is defined as $\hat{\mathbf{X}}^k = [diag(\hat{\mathbf{x}}_1^k)^{\mathsf{T}}, diag(\hat{\mathbf{x}}_2^k)^{\mathsf{T}}, ..., diag(\hat{\mathbf{x}}_D^k)^{\mathsf{T}}] \in \mathbb{C}^{T \times DT}$. $\mathbf{h}^k = [\mathbf{h}_1^{k\mathsf{T}}, \mathbf{h}_2^{k\mathsf{T}}, ..., \mathbf{h}_D^{k\mathsf{T}}]^{\mathsf{T}}$ and $\hat{\mathbf{g}}^k = [\hat{\mathbf{g}}_1^{k\mathsf{T}}, \hat{\mathbf{g}}_2^{k\mathsf{T}}, ..., \hat{\mathbf{g}}_D^{k\mathsf{T}}]^{\mathsf{T}}$ are representations of concatenated channels separately for $\mathbf{h}$ and $\hat{\mathbf{g}}$. Besides, the symbol $\hat{}$ denotes the DFT of a signal. $I_D \in$

505

$\mathbb{R}^{D \times D}$ is an identity matrix. The operator $\otimes$ and superscript T are Kronecker production and conjugate transpose operation respectively. By applying alternating direction method of multipliers (ADMM) [18], the augmented Lagrangian form of Eq. (5) is formulated as:

$$\hat{\mathcal{L}}(\mathbf{h}^k, \hat{\mathbf{g}}^k, \mathbf{r}^k, \hat{\zeta}^k) = \hat{\mathcal{E}}(\mathbf{h}^k, \mathbf{r}^k, \hat{\mathbf{g}}^k) + \mathscr{L}(\mathbf{h}^k, \hat{\mathbf{g}}^k, \hat{\zeta}^k) \quad . \quad (6)$$

where $\mathscr{L}(\mathbf{h}, \hat{\mathbf{g}}, \hat{\zeta}) = \frac{\mu}{2} \left\| \hat{\mathbf{g}} - \sqrt{T}(\mathbf{I}_D \otimes \mathbf{FP}^\mathsf{T})\mathbf{h} + \frac{\hat{\zeta}}{\mu} \right\|_2^2$ is the augmented Lagrangian item. μ denotes the penalty factor of the additional item and $\hat{\zeta} = [\hat{\zeta}_1, \hat{\zeta}_2, ..., \hat{\zeta}_D] \in \mathbb{C}^{DT \times 1}$ is the Lagrange multiplier. By decomposing Eq. (6) into subproblems, ADMM algorithm can be adopted as follows.

Subproblem h: Given $\hat{\mathbf{g}}^k$, $\mathbf{r}^k$ and $\hat{\zeta}$, the optimal $\mathbf{h}^{k*}$ can be obtained in closed-form by switching back to time domain:

$$\begin{aligned} \mathbf{h}_{i+1}^{k*} &= \arg\min_{\mathbf{h}^k} \left\{ \frac{\lambda_1}{2} \|\mathbf{h}^k\|_2^2 + \mathscr{L}(\mathbf{h}^k, \hat{\mathbf{g}}^k, \hat{\zeta}^k) \right\} \\ &= (\frac{\lambda_1}{T} + \mu)^{-1}(\mu\mathbf{g}^k + \zeta^k) \end{aligned} \quad , \quad (7)$$

where i denotes the i-th iteration.

Subproblem $\hat{\mathbf{g}}$: Once $\hat{\mathbf{h}}^k$, $\mathbf{r}^k$ and $\hat{\zeta}$ are fixed, the optimization of $\hat{\mathbf{g}}^{k*}$ can be gained as:

$$\hat{\mathbf{g}}_{i+1}^{k*} = \arg\min_{\hat{\mathbf{g}}^k} \left\{ \frac{1}{2T} \left\| \hat{\mathbf{r}}^k - \hat{\mathbf{X}}^k\hat{\mathbf{g}}^k \right\|_2^2 + \mathscr{L}(\mathbf{h}^k, \hat{\mathbf{g}}^k, \hat{\zeta}^k) \right\} \quad . \quad (8)$$

For the sake of lower computation complexity, we sample $\hat{\mathbf{x}}^k$ across all D channels in each pixel as:

$$\hat{\mathbf{g}}^{k*}(n) = \arg\min_{\hat{\mathbf{g}}^k(n)} \left\{ \begin{array}{l} \frac{1}{2T} \left\| \hat{\mathbf{r}}^k(n) - \hat{\mathbf{x}}^k(n)^\mathsf{T}\hat{\mathbf{g}}^k(n) \right\|_2^2 \\ + \frac{\mu}{2} \left\| \hat{\mathbf{g}}^k(n) - \hat{\mathbf{h}}^k(n) + \frac{\zeta(n)}{\mu} \right\|_2^2 \end{array} \right\} \quad , \quad (9)$$

where $\hat{\mathbf{h}}^k(n) = [\hat{\mathbf{h}}_1^k(n), \hat{\mathbf{h}}_2^k(n), ..., \hat{\mathbf{h}}_D^k(n)] \in \mathbb{C}^{D \times 1}$ and $\hat{\mathbf{h}}^k = \sqrt{D}\mathbf{FP}^\mathsf{T}\mathbf{h}^k$. Since $\hat{\mathbf{h}}^k$ can be solved efficiently in frequency domain, the solution of $\mathbf{g}^{k*}(n)$ is obtained by:

$$\hat{\mathbf{g}}^{k*}(n) = \frac{1}{\mu T} \left(\mathbf{I}_D - \frac{\hat{\mathbf{x}}^k(n)\hat{\mathbf{x}}^{k\mathsf{T}}(n)}{\mu T\mathbf{I}_D + \hat{\mathbf{x}}^{k\mathsf{T}}(n)\hat{\mathbf{x}}^k(n)} \right) \rho \quad , \quad (10)$$

where the vector ρ is formed as $\rho = \hat{\mathbf{x}}^k(n)\hat{\mathbf{r}}^k + \mu T\hat{\mathbf{h}}^k(n) - T\zeta(n)$. Sherman-Morrison formula [19] is used in the optimization of Eq. (10), *i.e.*, $(\mathbf{A} + \mathbf{uv}^\mathsf{T})^{-1} = \mathbf{A}^{-1} - (\mathbf{I}_D + \mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u})^{-1}\mathbf{A}^{-1}\mathbf{uv}^\mathsf{T}\mathbf{A}^{-1}$, where $\mathbf{A} = \mu T\mathbf{I}_D$, and $\mathbf{u} = \mathbf{v} = \hat{\mathbf{x}}^k(n)$. In consequence, Eq. (10) can be further simplified as:

$$\begin{aligned} \hat{\mathbf{g}}^{k*}(n) = &\frac{1}{\mu T}(\hat{\mathbf{x}}^k(n)\hat{\mathbf{r}}^k(n) + \mu T\hat{\mathbf{h}}^k(n) - T\zeta(n)) \\ &- \frac{\hat{\mathbf{x}}^k(n)}{\mu Tb}(\hat{\mathbf{r}}^k(n)\hat{\mathbf{s}}_{\mathbf{x}}^k(n) - T\hat{\mathbf{s}}_\zeta^k(n) + \mu T\hat{\mathbf{s}}_{\mathbf{h}}^k(n)) \end{aligned} \quad , \quad (11)$$

where $\hat{\mathbf{s}}_{\mathbf{x}}^k(n) = \hat{\mathbf{x}}^{k\mathsf{T}}(n)\hat{\mathbf{x}}^k(n)$, $\hat{\mathbf{s}}_\zeta^k(n) = \hat{\mathbf{x}}^{k\mathsf{T}}(n)\hat{\zeta}$, $\hat{\mathbf{s}}_{\mathbf{h}}^k(n) = \hat{\mathbf{x}}^{k\mathsf{T}}(n)\hat{\mathbf{h}}$ and $b = \hat{\mathbf{s}}_{\mathbf{x}}^k(n) + \mu T$.

Subproblem r: The optimal $\hat{\mathbf{r}}^{k*}$ can be obtained in closed-form, on condition that $\hat{\mathbf{h}}^k$, $\mathbf{g}^k$ and $\hat{\zeta}$ are fixed:

$$\begin{aligned} \mathbf{r}^{k*} = \arg\min_{\mathbf{r}^k} &\left\{ \begin{array}{l} \frac{1}{2} \left\| \hat{\mathbf{r}}^k - \hat{\mathbf{X}}^k\hat{\mathbf{g}}^k \right\|_2^2 \\ + \frac{(1+\Psi^2)\lambda_2}{2} \left\| \hat{\boldsymbol{\Omega}}^k - \hat{\mathbf{r}}^k \right\|_2^2 \\ + \frac{(1-\Psi^2)\phi}{2} \left\| \hat{\mathbf{r}}^k - \hat{\mathbf{r}}^{k-1} \right\|_2^2 \end{array} \right\} \\ = &\frac{\hat{\mathbf{X}}^k\hat{\mathbf{g}}^k + \lambda_2(1+\Psi^2)\hat{\boldsymbol{\Omega}}^k + \phi(1-\Psi^2)\hat{\mathbf{r}}^{k-1}}{1 + \lambda_2(1+\Psi^2) + \phi(1-\Psi^2)} \end{aligned} \quad . \quad (12)$$

Lagrangian multiplier update: On this occasion, Lagrangian multipliers can be updated as:

$$\hat{\zeta}_{i+1}^{k*} = \hat{\zeta}_i^{k*} + \mu_{i+1}(\hat{\mathbf{g}}_{i+1}^{k*} - \hat{\mathbf{h}}_{i+1}^{k*}) \quad . \quad (13)$$

The step size regularization constant μ_{i+1} is updated as $\mu_{i+1} = min(\mu_{max}, \beta\mu_i)$, where β is defined in advance.

D. Object Localization

As the feature in the searching region of a new frame is extracted, the object can be localized according to the response map $\mathbf{R}^k$, which is generated as:

$$\mathbf{R}^k = \mathcal{F}^{-1}\sum_{d=1}^D \hat{\mathbf{z}}_d^k \odot \hat{\mathbf{g}}_d^{k-1} \quad , \quad (14)$$

where $\mathcal{F}^{-1}$ indicates the inverse discrete Fourier transform (IDFT) operator and $\hat{\mathbf{z}}_d^k$ is the extracted feature map of frequency domain in the k-th frame.

IV. EXPERIMENT

In this section, the evaluation experiments of MSCF tracker are conducted on 243 challenging image sequences and altogether over 90,000 frames from three widely applied UAV benchmarks, namely UAV123@10fps [20], DTB70 [21], and UAVDT [22]. Performance of MSCF tracker is compared with other 15 CPU-based trackers, including KCF [6], SAMF [23], SAMF_CA [23], DSST [24], fDSST [11], Staple_CA [25], KCC [26], BACF [4], CSR-DCF [27], SRDCF [28], SRDCFdecon [28], STRCF [29], ECO-HC [30], MCCT-H [31], ARCF-H [32], and 11 GPU-based trackers, *i.e.*, DeepSTRCF [29], CoKCF [33], ECO [30], CCOT [13], MCPF [34], UDT [35], UDT+ [35], IBCCF [36], ASRCF [37], MCCT [31], and TADT [38].

A. Implementation Details

For fairness in evaluation, the parameters of the testes trackers are from their official versions, and the tests are conducted on the same platform.

1) Parameter setting: In terms of ADMM algorithm, we use three-step iterations for optimization. Other parameters are set as: $\mu = 0.1$, $\mu_{max} = 10000$, $\beta = 10$, $\gamma = 27$ and $\gamma_{max} = 10000$. The learning rate is set to 0.0158. The training process is performed every 2 frames in practice for boosting efficiency. In addition, the regularization parameters are set as: $\lambda_1 = 20$, $\lambda_2 = 840$, and $\phi = 1$. The hyper parameters for adaptive hybrid label represent as: $\delta = 0.01$, $\nu = 1$, $\eta = 0.044$ and the ratio of pedestal size is set to 2.5. All parameters remain fixed for all image sequences on three benchmarks. Source code is available here: https://github.com/vision4robotics/MSCF-tracker.

2) Features: MSCF tracker adopts a combination of histogram of oriented gradient (HOG) [39], color names (CN) [12] and gray-scale features for object representation.

3) Platform: All the experiments, including the test of the MSCF tracker and others, are conducted in MATLAB R2019a on a PC equipped with an Intel Core i7-8700K CPU (3.70GHz) and an Nvidia GeForce RTX 2080 GPU.

Precision plots on UAV123@10fps

MSCF [0.648]
CSR-DCF [0.643]
ECO-HC [0.640]
STRCF [0.627]
ARCF-H [0.612]
MCCT-H [0.596]
Staple_CA [0.587]
SRDCFdecon [0.584]
SRDCF [0.575]
Staple [0.573]
BACF [0.572]
KCC [0.531]
fDSST [0.516]
SAMF [0.467]
DSST [0.448]
KCF [0.406]

Precision plots on UAVDT

MSCF [0.713]
ARCF-H [0.705]
Staple_CA [0.695]
ECO-HC [0.694]
BACF [0.686]
DSST [0.681]
Staple [0.678]
CSR-DCF [0.674]
MCCT-H [0.668]
fDSST [0.666]
SRDCF [0.659]
KCC [0.649]
SRDCFdecon [0.644]
STRCF [0.629]
SAMF [0.586]
KCF [0.571]

Precision plots on DTB70

MSCF [0.664]
STRCF [0.649]
CSR-DCF [0.646]
ECO-HC [0.635]
ARCF-H [0.607]
MCCT-H [0.604]
BACF [0.581]
fDSST [0.534]
SAMF [0.519]
SRDCF [0.512]
SRDCFdecon [0.504]
Staple_CA [0.504]
Staple [0.489]
KCF [0.468]
DSST [0.463]
KCC [0.440]

Success plots on UAV123@10fps

MSCF [0.471]
ECO-HC [0.468]
STRCF [0.457]
CSR-DCF [0.450]
ARCF-H [0.434]
MCCT-H [0.433]
SRDCFdecon [0.429]
SRDCF [0.423]
Staple_CA [0.420]
Staple [0.415]
BACF [0.413]
fDSST [0.379]
KCC [0.374]
DSST [0.332]
SAMF [0.326]
KCF [0.265]

Success plots on UAVDT

MSCF [0.455]
BACF [0.432]
SRDCF [0.417]
ECO-HC [0.416]
ARCF-H [0.413]
STRCF [0.411]
SRDCFdecon [0.411]
MCCT-H [0.402]
Staple_CA [0.394]
Staple [0.393]
DSST [0.392]
CSR-DCF [0.389]
KCC [0.389]
fDSST [0.383]
SAMF [0.313]
KCF [0.290]

Success plots on DTB70

MSCF [0.459]
ECO-HC [0.448]
CSR-DCF [0.438]
STRCF [0.437]
ARCF-H [0.416]
MCCT-H [0.405]
BACF [0.398]
SRDCF [0.363]
fDSST [0.357]
Staple_CA [0.351]
SRDCFdecon [0.351]
SAMF [0.339]
Staple [0.336]
DSST [0.321]
KCC [0.291]
KCF [0.280]

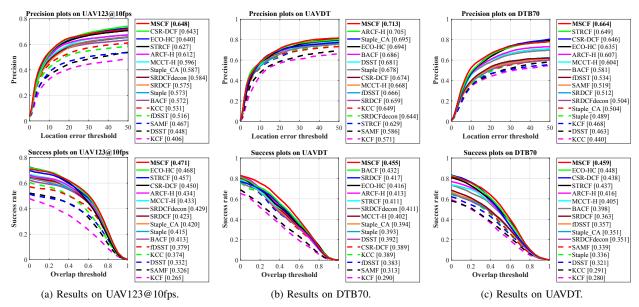(a) Results on UAV123@10fps.   (b) Results on DTB70.   (c) Results on UAVDT.

Fig. 4. Overall performance evaluation. PPs and SPs of the proposed MSCF tracker and other 15 state-of-the-art trackers. MSCF yields competitive performance on all three benchmarks in both terms of precision and AUC.

TABLE I

OVERALL AVERAGE PERFORMANCE EVALUATION. AVERAGE PRECISION, SUCCESS RATE AND FPS OF TOP 12 TRACKERS ON THREE BENCHMARKS.

THE BEST THREE PERFORMANCES ARE RESPECTIVELY HIGHLIGHTED WITH RED, GREEN, AND BLUE COLORS.

Trackers	BACF	Staple_CA	SRDCF	SRDCFdecon	MCCT-H	CSR-DCF	STRCF	ECO-HC	fDSST	Staple	ARCF-H	MSCF
Prec.	0.613	0.595	0.582	0.577	0.622	0.655	0.635	0.657	0.572	0.580	0.641	0.675
Succ.	0.414	0.388	0.401	0.397	0.414	0.426	0.435	0.444	0.373	0.382	0.421	0.462
FPS	43.8	52.1	10.9	5.2	49.1	11.3	24.3	60.8	173.8	91.8	41.6	37.6

4) Evaluation metrics: All evaluations are based on the one-pass evaluation (OPE), which includes two metrics, *i.e.*, precision and success rate [40]. The precision is measured by the center location error (CLE) between the estimated and the ground truth bounding boxes. In practice, the percentage of frames whose CLE below 20 pixels (as commonly set) in the precision plot (PP) is used to rank the trackers. As for the success rate, the intersection over union (IoU) between the ground truth and estimated results is reported in the success plot (SP), as usual, this work uses the area under the curve (AUC) of SP to rank the trackers. Besides, frames/s is used to rank the tracking speed of the approaches.

B. Comparison with CPU-based Trackers

1) Overall performance: Fig. 4 demonstrates the overall performance of MSCF with other trackers on UAV123@10fps [20], DTB70 [21], and UAVDT [22] benchmarks. The proposed MSCF tracker has outperformed all other trackers with handcrafted features on three benchmarks. On UAV123@10fps, MSCF ranks first in precision (0.648) and AUC (0.471). While on UAVDT, MSCF (0.713) has an advantage of 1.1% over the second-best ARCF-H (0.705) on precision. In terms of AUC, MSCF (0.455) has an advantage of 5.4% over the second-best BACF (0.432). On DTB70 benchmark, MSCF has also achieved the best performance in precision (0.664) and AUC (0.459). Overall evaluation of performance on all three benchmarks in terms of precision,

AUC, and FPS is demonstrated in Table I. Against baseline BACF, MSCF has an advancement of 10.1% in precision and 11.6% in AUC. Besides satisfactory tracking results, the speed of MSCF (37.6 FPS) is adequate for real-time UAV tracking applications.

2) Attribute-based performance: SPs of all trackers on UAV-specific attributes of three benchmarks are demonstrated in Fig. 5. The figure illustrates that mutations caused by challenges including fast motion, camera motion, *etc.*, are commendably sensed and repressed by the MSCF tracker. In case of camera motion, MSCF demonstrates an improvement of performance by 14.6% in UAVDT, and 13.1% in DTB70 against BACF. The ability to cope with object motion has also been promoted by 20.7% in UAV123@10fps, 7.3% in UAVDT and 17.2% in DTB70 from BACF. For explication, mutations of object or camera motion commonly bring sub-peaks in response maps, and the proposed MTF can instantly correct the hybrid label for filter training. Taking the neighborhood of the centre into consideration, the label can adapt to the mutation of the appearance smoothly. Some qualitative comparisons of the top 8 trackers are displayed in Fig. 6.

C. Comparison with Deep-Based Trackers

To better illustrate the comprehensive evaluation with state-of-the-art trackers, the performance of 11 deep-based methods are also represented in comparison with MSCF

(a) Fast motion on UAV123@10fps.   **(b) Camera motion on UAVDT.**   **(c) Object motion on UAVDT.**   **(d) Fast camera motion on DTB70.**

(a) Scale variation on UAV123@10fps.   **(b) Large occlusion on UAVDT.**   **(c) Deformation on DTB70.**   **(d) Motion blur on DTB70.**
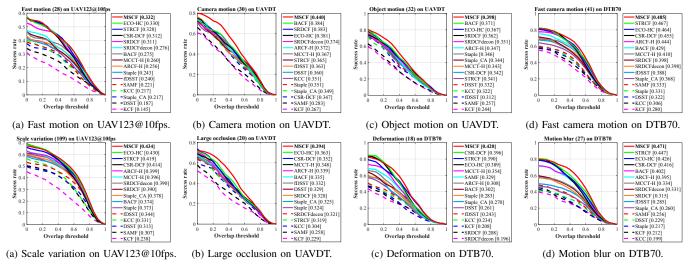
Fig. 5. Attribute-based performance evaluation. Success plots of MSCF and other state-of-the-art trackers on UAV-specific attributes from three benchmarks. MSCF outperforms other trackers on all challenges above.
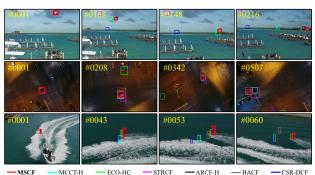

Fig. 6. Qualitative evaluation of 7 top-ranked trackers. From top to down are $Gull2$ from DTB70, $S1301$ from UAVDT, and $wakeboard7$ from UAV123@10fps. More tracking videos can be found here: https://youtu.be/4EPGBCuatxU.

TABLE II

PERFORMANCE COMPARISON OF MSCF WITH OTHER 11

STATE-OF-THE-ART DEEP-BASED TRACKERS ON UAVDT BENCHMARK.

NOTE THAT THE SUPERSCRIPT * DENOTES GPU SPEED.

Trackers	Prec.	Succ.	FPS	Trackers	Prec.	Succ.	FPS
CCOT	0.656	0.406	1.1*	DeepSTRCF	0.667	0.437	6.8*
CoKCF	0.605	0.319	20.1*	TADT	0.677	0.431	32.3*
ECO	**0.700**	*0.454*	16.4*	UDT	0.674	**0.442**	*73.3**
IBCCF	0.603	0.389	3.0*	UDT+	0.696	0.415	*56.9**
MCPF	0.675	0.403	0.6*	ASRCF	*0.700*	0.437	22.2*
MCCT	0.671	0.437	7.9*	**MSCF**	**0.713**	**0.455**	**37.6**

TABLE III

ABLATION STUDY ON AVERAGE PERFORMANCE OF THE PROPOSED

MSCF TRACKER.

Settings	BACF	BACF+AL	BACF+AHL	**MSCF**
Prec.	0.613	**0.668**	*0.673*	**0.675**
Succ.	0.414	**0.456**	*0.458*	**0.462**

in TABLE II. Owing to heavy reliance on high-end GPUs for deep-based computation, while the MSCF tracker is merely in need of a single CPU, the prime cost of MSCF is accordingly lower than deep-based trackers. Moreover, the proposed MSCF tracker is superior to deep-based trackers on both precision and success rate.

D. Albation Study

TABLE III demonstrates the availability of every amendment for the baseline tracker. As the adaptive label (AL), *i.e.*, the first two terms in Eq. (4), is applied in the objective function, where the predefined Gaussian label is replaced by a trained label, the precision and success rate have achieved an improvement by 9.0% and 10.1%, respectively. Taking the adaptive hybrid label (AHL) into consideration, *i.e.*, appending the spatial regularization term to the objective function, the performance has a further improvement. Moreover, after utilizing the AHL in the previous frame, *i.e.*, appending the temporal regularization term, the complete MSCF has been promoted by 0.3% in precision and 0.9% in success rate. To sum up, the MSCF has gained superior competence on the baseline.

V. CONCLUSION

In this work, a mutation sensitive tracker with an adaptive hybrid label is proposed for real-time UAV tracking. Meanwhile, a novel evaluation of the response map is introduced, namely MTF, which indicates the severity of tracking mutations. The whole work-flow also adopts the genetic idea to reach optimal solutions. Comprehensive experiments have corroborated the favorable performance of MSCF. Furthermore, employing the visual tracking adaptability of MSCF for UAV tracking in the real environment is also taken into consideration to conduct in the future. We believe that DCF based framework for real-time UAV tracking can be further improved with the proposition of our MSCF tracker.

REFERENCES

[1] A. A. Meera, M. Popović, A. Millane, and R. Siegwart, "Obstacle-Aware Adaptive Informative Path Planning for UAV-based Target Search," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 718–724.

[2] M. Elloumi, R. Dhaou, B. Escrig, H. Idoudi, and L. A. Saidane, "Monitoring Road Traffic with a UAV-Based System," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, 2018, pp. 1–6.

[3] B. Patel, T. D. Barfoot, and A. P. Schoellig, "Visual Localization with Google Earth Images for Robust Global Pose Estimation of UAVs," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6491–6497.

[4] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning Background-Aware Correlation Filters for Visual Tracking," in *Proceeding of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1144–1152.

[5] C. Fu, B. Li, F. Ding, F. Lin, and G. Lu, "Correlation Filter for UAV-Based Aerial Tracking: A Review and Experimental Evaluation," *arXiv:2010.06255*, 2020.

[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[7] C. Fu, J. Ye, J. Xu, Y. He, and F. Lin, "Disruptor-Aware Interval-Based Response Inconsistency for Correlation Filters in Real-Time Aerial Tracking," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2020.

[8] F. Lin, C. Fu, Y. He, F. Guo, and Q. Tang, "Learning Temporary Block-Based Bidirectional Incongruity-Aware Correlation Filters for Efficient UAV Object Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–14, 2020.

[9] C. Fu, F. Ding, Y. Li, J. Jin, and C. Feng, "Learning Dynamic Regression with Automatic Distractor Repression for Real-Time UAV Tracking," *Engineering Applications of Artificial Intelligence*, vol. 98, p. 104116.

[10] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual Object Tracking Using Adaptive Correlation Filters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2544–2550.

[11] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative Scale Space Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561–1575, 2017.

[12] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Weijer, "Adaptive Color Attributes for Real-Time Visual Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1090–1097.

[13] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 472–488.

[14] A. Bibi, M. Mueller, and B. Ghanem, "Target Response Adaptation for Correlation Filter Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 419–433.

[15] R. Han, W. Feng, and S. Wang, "Fast Learning of Spatially Regularized and Content Aware Correlation Filter for Visual Tracking," *IEEE Transactions on Image Processing*, vol. 29, pp. 7128–7140, 2020.

[16] M. Wang, Y. Liu, and Z. Huang, "Large Margin Object Tracking with Circulant Feature Maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4800–4808.

[17] Y. Han, C. Deng, B. Zhao, and D. Tao, "State-Aware Anti-Drift Object Tracking," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 4075–4086, 2019.

[18] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, vol. 3, pp. 1–122, 2010.

[19] S. W. J. Morrison, "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix," *Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 124–127, 1950.

[20] M. Mueller, N. Smith, and B. Ghanem, "A Benchmark and Simulator for UAV Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 445–461.

[21] S. Li and D.-Y. Yeung, "Visual Object Tracking for Unmanned Aerial Vehicles: A Benchmark and New Motion Models," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017, pp. 4140–4146.

[22] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 375–391.

[23] Y. Li and J. Zhu, "A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration," in *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2014, pp. 254–265.

[24] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate Scale Estimation for Robust Visual Tracking," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2014, pp. 65.1–65.11.

[25] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. Torr, "Staple: Complementary Learners for Real-Time Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1401–1409.

[26] C. Wang, L. Zhang, L. Xie, and J. Yuan, "Kernel Cross-Correlator," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 4179–4186.

[27] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative Correlation Filter with Channel and Spatial Reliability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6309–6318.

[28] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical Convolutional Features for Visual Tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3074–3082.

[29] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4904–4913.

[30] Danelljan, Martin and Bhat, Goutam and Shahbaz Khan, Fahad and Felsberg, Michael, "ECO: Efficient Convolution Operators for Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6638–6646.

[31] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue Correlation Filters for Robust Visual Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4844–4853.

[32] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, "Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 2891–2900.

[33] L. Zhang and P. N. Suganthan, "Robust Visual Tracking via Co-Trained Kernelized Correlation Filters," *Pattern Recognition*, vol. 69, pp. 82–93, 2017.

[34] T. Zhang, C. Xu, and M.-H. Yang, "Multi-Task Correlation Particle Filter for Robust Object Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 4335–4343.

[35] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, and H. Li, "Unsupervised Deep Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1308–1317.

[36] F. Li, Y. Yao, P. Li, D. Zhang, W. Zuo, and M.-H. Yang, "Integrating Boundary and Center Correlation Filters for Visual Tracking With Aspect Ratio Variation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2001–2009.

[37] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual Tracking via Adaptive Spatially-Regularized Correlation Filters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4670–4679.

[38] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-Aware Deep Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1369–1378.

[39] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[40] Y. Wu, J. Lim, and M.-H. Yang, "Object Tracking Benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.