IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 24, 2018, accepted September 21, 2018, date of publication September 28, 2018, date of current version October 25, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2872691

Complementary Tracking Via Dual Color Clustering and Spatio-Temporal Regularized Correlation Learning

JIAQING FAN[1], HUIHUI SONG[1], KAIHUA ZHANG[1], QINGSHAN LIU[1], AND WEI LIAN[2]
[1]B-DAT, CICAEET, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2]Department of Computer Science, Changzhi University, Changzhi 046011, China

Corresponding author: Kaihua Zhang (zhkhua@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61872189, Grant 61876088, Grant 61532009, and Grant 61773002, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20170040, and in part by the Applied Basic Research Project in Shanxi Province under Grant 201601D011007.

ABSTRACT Recently, a simple, yet effective and efficient tracker named Staple has achieved promising performance in terms of efficiency and accuracy on a series of visual tracking benchmarks. Staple is equipped with complementary learners of discriminative correlation filters (DCFs) and color histograms, which are robust to both color changes and deformations. However, it has some drawbacks: 1) Staple only employs standard color histograms with the same quantization step for all sequences, which does not consider the specific structural information of target in each sequence, thereby affecting its discriminative capability to separate target from background. 2) The standard DCFs are efficient but suffer from unwanted boundary effects, leading to failures in some challenging scenarios. To address these issues, we present a dual color clustering and spatio-temporal regularized correlation regressions-based complementary tracker (CSCT). The proposed CSCT includes two components with complementary merits to adaptively deal with significant color variations and deformations for each sequence: First, we design a novel color clustering-based histogram model that first adaptively divides the colors of the target in the 1st frame into several cluster centers, and then the cluster centers are taken as references to construct adaptive color histograms for targets in the coming frames, which enable to adapt significant target deformations. Second, we propose to learn spatio-temporal regularized CFs, which not only enable to avoid boundary effects but also provides a more robust appearance model than the discriminative CFs in Staple in the case of large appearance variations. Compared to Staple, our CSCT with handcrafted features achieves a gain of 5.9%, 3.4%, and 1.5% on OTB100, Temple-Color, and VOT2016 benchmarks in terms of AUC and EAO scores, respectively. Moreover, our CSCT performs favorably against several state-of-the-art trackers, including the deep learning-based trackers.

INDEX TERMS Visual tracking, correlation filter, color histograms, spatio-temporal regularization.

I. INTRODUCTION

Visual tracking is a hot research topic in the filed of computer vision with numerous applications such as video surveillance, motion analysis, and autonomous driving, to name a few [1]–[5]. Despite much progress in recent years, it remains challenging to develop a robust tracking algorithm due to significant target appearance variations caused by the factors such as illumination changes, fast motions, pose variations, partial occlusions and background clutters. Hence, a robust representation plays a key role in a successful tracking, thereby attracting much attention in the past decades.

Recently, much attention has been paid to learn discriminative CF (DCF) based representations for visual tracking [6]–[10]. However, since the learned CFs strongly depend on the spatial layout of the tracked target, they are sensitive to deformations [11]. To address this issue, a simple yet effective tracker named Staple [11] has been proposed, which marries the merits of statistical color information and template learning by CF to favorably handle deformation and color changes simultaneously. However, there exist two main issues to be addressed: (i) Staple only employs a stan-

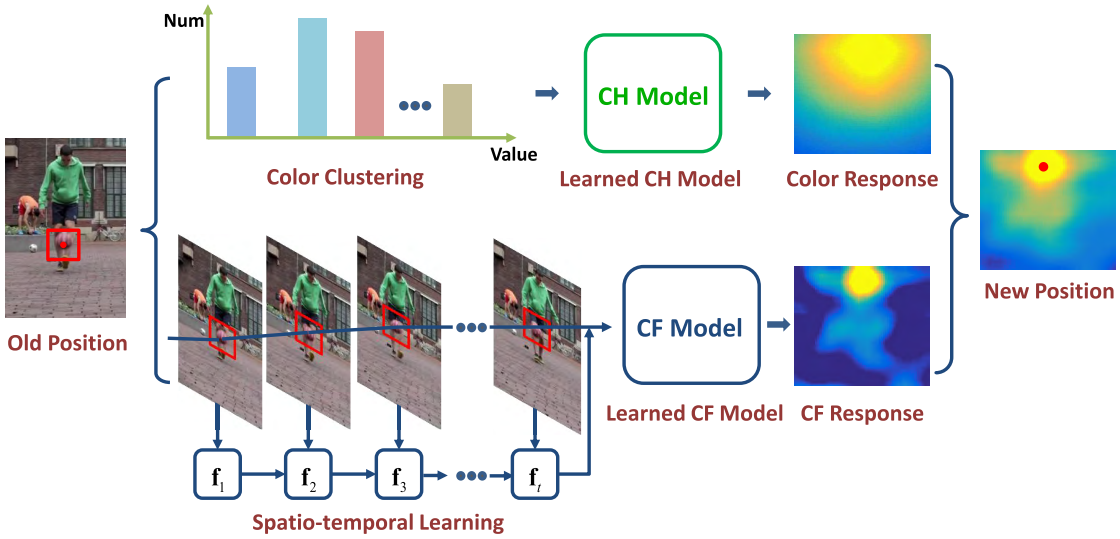


FIGURE 1. Basic flow of our CSCT. Here, CH denotes color histogram and CF indicates correlation filter. On the top, we employ the K-means algorithm to get several clustering centers of the colors of the target object in the 1st frame. Then, we assign each pixel the index of its nearest clustering center. Meanwhile, on the bottom, we learn a spatio-temporal regularized CF model with multiple samples from the historical tracking results, and the learned CFs emphasize more to the recent samples to adapt target appearance variations over time. Finally, the color clustering histogram response is merged with the spatio-temporal CF response to yield the final response.

dard color histogram with the same quantization step for all sequences, which does not consider the specific structural information of the target in each sequence, thereby it may reduce their description capability. (ii) Staple resorts to the standard DCFs that are efficient but suffer from unwanted boundary effects, leading to failures when target appearance varies significantly in some challenging scenarios.

To alleviate the above issues, in this paper, we present a dual Color clustering and Spatio-temporal regularized correlation regressions based Complementary Tracker (CSCT). The principle of CSCT is shown in Figure 1. Specifically, CSCT leverages the color clustering of the target appearance at the 1st frame to construct a color histogram representation with a set of data-adaptive bins that can effectively encode the structural information of target. Meanwhile, to improve the discriminative capacity of the CFs to be learned, a spatio-temporal regularization term is introduced to regularize the CF model to be learned, which is more robust to drastic appearance variations than the traditional CFs. Extensive evaluations on the OTB100 [12], Temple-Color [13] and VOT2016 [14] datasets demonstrate that the proposed CSCT achieves obviously improved performance against the baseline Staple [11], and performs favorably against several state-of-the-art trackers [7], [11], [15]–[21].

The main contributions of this paper are summarized as follows:

1) An effective color clustering histogram model is proposed that is more robust than the standard color histogram model, yielding a more reliable color tracking model.
2) A novel spatio-temporal regularization term is introduced to regularize the CFs to be learned, which combines the useful spatial and temporal information, leading to a more robust CF tracking model.
3) The proposed CSCT with handcrafted features achieves promising performance on the OTB100, Temple-Color and VOT2016 benchmarks in terms of both accuracy and efficiency.

II. RELATED WORK

A. COLOR-BASED TRACKING

In the earlier proposed object tracking approaches, color information is widely employed to increase their tracking robustness. Pérez *et al.* [22] introduce a multi-part color model to capture a rough spatial layout ignored by global histograms. Adam *et al.* [23] represent object template by multiple image fragments or patches to achieve a more robust local histogram-based representation. Recently, instead of simple color representations, Danelljan *et al.* [24] leverage more sophisticated attribute-based color features to describe object appearance, achieving promising tracking performance. Possegger *et al.* [15] exploit an adaptive object color histogram model to suppress nearby similar regions, resulting in a robust and reliable tracking result.

B. DISCRIMINATIVE CORRELATION FILTER TRACKING

Recently, DCFs have attracted much attention in visual tracking for their advantages in terms of efficiency and robustness. Using DCFs for visual tracking starts with MOSSE [6], which learns CFs with a few samples in the frequency domain, thereby facilitating fast Fourier transforms (FFTs) for efficient computation that runs at 669 frames per second (FPS). Henriques *et al.* [25] first explore the circulant structure of dense samples with kernel embedding that facilitates learning

CFs for fast tracking. Henriques *et al.* [7] further improve the CF tracker in [25] by extending its feature representation from image intensities to histogram of gradients (HOGs). Ma *et al.* [26] exploit complementary traits of different layers of deep features and use a coarse-to-fine search strategy to learn more effective CFs for visual tracking, significantly boosting performance in the OTB100 dataset. Recently, Danelljan *et al.* propose a series of spatially regularized CF-based trackers [8], [27], [28] with impressive performance. The spatially regularized DCF (SRDCF) tracker [8] tries to suppress the boundary effects of the learned CFs with Gaussian shaped spatial regularization weights. Based on [8], [27] proposes an adaptive decontamination scheme to learn more effective CFs, which adaptively learns the reliability of each training sample and eliminates the influence of contaminated ones. In [28], learning CFs is conducted in the continuous spatial domain of various feature maps, which is able to achieve sub-pixel accuracy of the tracking locations.

C. INTEGRATE MULTIPLE ESTIMATES

A widely adopted strategy to mitigate inaccurate predictions is to combine the estimations of an ensemble of methods, so that the weakness of the trackers are reciprocally compensated. Kwon *et al.* [29], [30] make use of complementary basic trackers, built by combining different observation models and motion models, and then integrate their estimates in a sampling framework. Wang and Yeung [31] combine several independent trackers via a factorial HMM, modelling both the object trajectory and the reliability of each tracker across time. Rather than using different kinds of trackers, the Multi-Expert Entropy Minimisation (MEEM) tracker [16] maintains a collection of past models and chooses the prediction of one according to an entropy criterion. Bertinetto *et al.* [11] combine two common ridge regression scores directly, where the local representations (with HOGs) and global representations (with color histograms) work together.

III. METHODOLOGY

As the DCF based tracker [7], the proposed CSCT adopts the tracking-by-detection paradigm. Here, our main concerns are how to effectively take advantage of the color distribution of the target object in the 1st frame and how to merge the clustering color response with the CF response, which will be depicted in detail as below.

A. COLOR CLUSTERING FOR HISTOGRAM-BASED MODEL

The top branch in Figure 1 illustrates how we design the clustering color histogram in our algorithm. In the 1st frame, we cluster all the RGB colors of the whole target area, which contains both object and background areas, into n categories by the K-means algorithm, yielding the clustering center set $\{\mathbf{c}_i \in \mathbb{R}^3\}_{i=1}^n$. Then, for each pixel of RGB color $\mathbf{u} \in \mathbb{R}^3$, its feature representation has a special form as

$$\boldsymbol{\psi}[\mathbf{u}] = \mathbf{e}_{k[\mathbf{u}]}, \tag{1}$$

where $\mathbf{e}_i = [\underbrace{0, 0, \ldots, 0}_{i-1}, 1, 0, \ldots, 0]^\top$. The index $k[\boldsymbol{u}]$ in (1) is obtained by

$$k[\mathbf{u}] = \arg\min_i \{\|\mathbf{u} - \mathbf{c}_i\|_2^2\}_{i=1}^n, \tag{2}$$

where $\mathbf{c}_i$ denotes the i-th clustering center yielded by the K-means algorithm. Afterwards, for each pixel feature representation, we employ a linear regression objective over the object and background regions Ω_o and $\Omega_b \subset \mathbb{R}^2$ as

$$\mathcal{E}(\boldsymbol{\beta}) = \frac{1}{|\Omega_o|}\sum\nolimits_{\boldsymbol{u}\in\Omega_o}\left(\boldsymbol{\beta}^\top\boldsymbol{\psi}[\boldsymbol{u}] - 1\right)^2 + \frac{1}{|\Omega_b|}\sum\nolimits_{\boldsymbol{u}\in\Omega_b}\left(\boldsymbol{\beta}^\top\boldsymbol{\psi}[\boldsymbol{u}]\right)^2, \tag{3}$$

where $\boldsymbol{\beta}$ denotes the feature weight vector.

Replacing (1) in (3), we have

$$\mathcal{E}(\boldsymbol{\beta}) = \sum\nolimits_{i=1}^n \left[\frac{N^i(\Omega_o)}{|\Omega_o|}\left(\beta^i - 1\right)^2 + \frac{N^i(\Omega_b)}{|\Omega_b|}\left(\beta^i\right)^2\right], \tag{4}$$

where $N^i(\Omega_l) = |\{\mathbf{u} \in \Omega_l | k[\mathbf{u}] = i\}|$, $l \in \{o, b\}$. Setting $\partial\mathcal{E}(\boldsymbol{\beta})/\partial\beta^i = 0$, the solution of minimizing $\mathcal{E}(\boldsymbol{\beta})$ can be achieved as

$$\widetilde{\beta}^i = \frac{N^i(\Omega_o)}{N^i(\Omega_o) + \frac{|\Omega_o|}{|\Omega_b|}N^i(\Omega_b)}. \tag{5}$$

To adapt target appearance variations over time, we leverage a simple online update strategy

$$\boldsymbol{\beta}_t = (1 - \eta_{cc})\boldsymbol{\beta}_{t-1} + \eta_{cc}\widetilde{\boldsymbol{\beta}}_t, \tag{6}$$

where η_{cc} is a learning factor and $\widetilde{\boldsymbol{\beta}}_t$ is the vector of $\widetilde{\beta}_t^i$ calculated by (5) using the tracking results at frame t. Finally, as for each pixel $\mathbf{u}$ at frame t, after calculating its color clustering feature representation $\boldsymbol{\psi}[\mathbf{u}]$ via (1), its response can be obtained by

$$\mathbf{r}_{cc}(\mathbf{u}) = \boldsymbol{\beta}_t{}^\top\boldsymbol{\psi}[\mathbf{u}] \tag{7}$$

Note that our improvement can significantly relieve the adverse impacts of the color noises due to the fixed quantization step in the standard color histogram model in Staple [11] and obtain a more reliable color response, finally improving the tracking performance.

B. SPATIO-TEMPORAL REGULARIZED CORRELATION TRACKING

The bottom branch in Figure 1 illustrates how we build the spatio-temporal regularized CFs in our method. Let $\mathbf{f}_{t-1}^d$ denote the d-th CF channel at the $(t-1)$-th frame, $\mathbf{w}$ denote the Gaussian-shaped spatial weight function, and $\{\mathbf{f}^d\}_{d=1}^D$ denote the set of the CF channels to be learned. Then, as [10], we introduce an effective spatio-temporal regularization term $\sum_{d=1}^{D}\left\|\mathbf{w}\cdot\mathbf{f}^d - \mathbf{w}\cdot\mathbf{f}_{t-1}^d\right\|_2^2$, where $\mathbf{w}$ is the spatial weight to

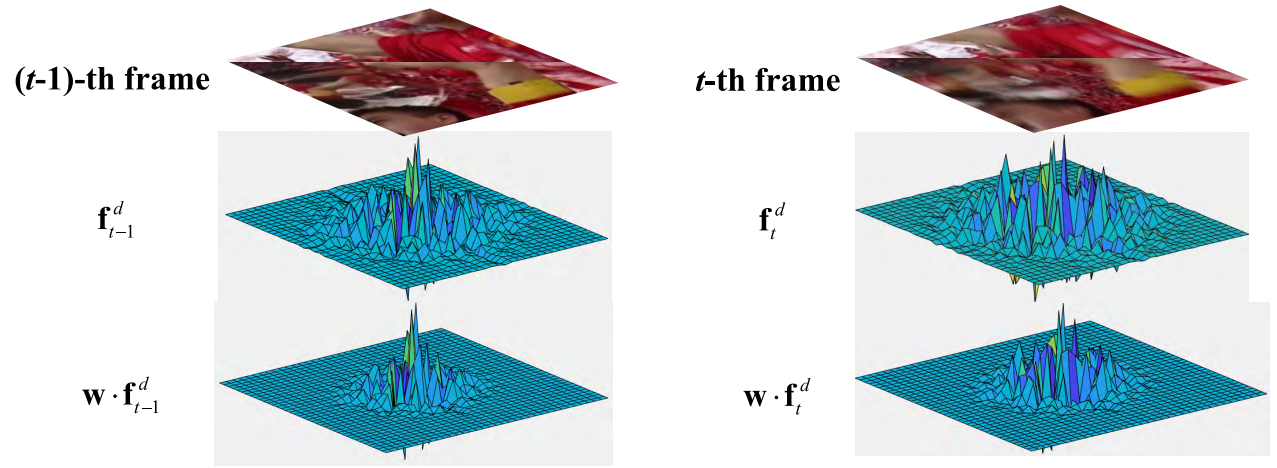


FIGURE 2. Visualization of the spatio-temporal regularized term in (8). When the object suffers from drastic variations, the difference between $\mathbf{w}\cdot\mathbf{f}_{t-1}^d$ and $\mathbf{w}\cdot\mathbf{f}_t^d$ will become large, leading to learning more stable CFs between two frames. Here, the w is the fixed gaussian-shaped weights borrow from SRDCF [8].

regularize each frame's learned d-th channel correlation filter $\mathbf{f}^d$, measuring the difference between the current frame's regularized CFs $\mathbf{w}\cdot\mathbf{f}^d$ and the former learned CFs $\mathbf{w}\cdot\mathbf{f}_{t-1}^d$. As shown in Figure. 2, the purpose of introducing the spatio-temporal regularized term is to make our learned CFs more self-adaptive to the environment changes. For example, our CF model will be more sensitive to these sudden appearance variations (like motion blur, illumination changes), owing to the spatio-temporal regularized term becoming large and constraining the learned filter $\mathbf{f}_t$. While suffering from the slow variations (like partial occlusion), thanks to our CF update strategy, the learned CF will be close to the former learned CF, which will not degrade in several occluded frames. In summary, with the introduction of the spatio-temporal regularization, our method can provide a more robust appearance model than the standard CFs, leading to superior performance.

Then, optimize the objective function as

$$\arg\min_{\{\mathbf{f}^d\}_{d=1}^D} \sum_{d=1}^{D} \left\| \mathbf{x}_t^d * \mathbf{f}^d - \mathbf{y} \right\|_2^2 + \lambda \sum_{d=1}^{D} \left\| \mathbf{w}\cdot\mathbf{f}^d - \mathbf{w}\cdot\mathbf{f}_{t-1}^d \right\|_2^2, \tag{8}$$

where the symbol $*$ denotes correlation operator, the symbol $\cdot$ denotes element-wise multiplication, $\{\mathbf{x}_t^d\}_{d=1}^D$ is a set of D channel features, $\mathbf{y}$ is a desired Gaussian-shaped response map, and λ controls the impact of the spatio-temporal regularization term.

Setting $\mathbf{g}^d = \mathbf{f}^d - \mathbf{f}_{t-1}^d$, (8) can be rewritten as

$$\arg\min_{\{\mathbf{g}^d\}_{d=1}^D} \sum_{d=1}^{D} \left\| \mathbf{x}_t^d * \mathbf{g}^d - \left(\mathbf{y} - \mathbf{x}_t^d * \mathbf{f}_{t-1}^d\right) \right\|_2^2 + \lambda \sum_{d=1}^{D} \left\| \mathbf{w}\cdot\mathbf{g}^d \right\|_2^2, \tag{9}$$

then we set $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{x}_t^d * \mathbf{f}_{t-1}^d$, and reformulate (9) as

$$\arg\min_{\{\mathbf{g}^d\}_{d=1}^D} \sum_{d=1}^{D} \left\| \mathbf{x}_t^d * \mathbf{g}^d - \tilde{\mathbf{y}} \right\|_2^2 + \lambda \sum_{d=1}^{D} \left\| \mathbf{w}\cdot\mathbf{g}^d \right\|_2^2. \tag{10}$$

For notation clarity, we assume that only a single channel in the following derivation, and drop the channel index $(\cdot)^d$, since the filter learning is independent across channels. Then, the objective function is simplified to

$$\arg\min_{\mathbf{g}} \left\| \mathbf{x}_t * \mathbf{g} - \tilde{\mathbf{y}} \right\|_2^2 + \lambda \left\| \mathbf{w}\cdot\mathbf{g} \right\|_2^2. \tag{11}$$

The objective function in (11) is convex with respect to $\mathbf{g}$, and hence the minimization problem has a globally optimal solution that can be achieved via ADMM [32]. To this end, we introduce a dual variable $\mathbf{h}$ and the constraint $\mathbf{h}-\mathbf{w}\cdot\mathbf{g}=\mathbf{0}$, yielding the following Augmented Lagrangian objective

$$\begin{aligned}\mathcal{L}\left(\hat{\mathbf{h}}, \mathbf{g}, \hat{\mathbf{s}}\right) &= \left\| \hat{\mathbf{h}}^H diag\left(\hat{\mathbf{x}}_t\right) - \hat{\tilde{\mathbf{y}}} \right\|_2^2 + \lambda \left\| \mathbf{g}_w \right\|_2^2 \\ &+ \left[\hat{\mathbf{s}}^H \left(\hat{\mathbf{h}} - \hat{\mathbf{g}}_w\right) + \overline{\hat{\mathbf{s}}^H \left(\hat{\mathbf{h}} - \hat{\mathbf{g}}_w\right)} \right] + \mu \left\| \hat{\mathbf{h}} - \hat{\mathbf{g}}_w \right\|_2^2, \end{aligned} \tag{12}$$

where $\hat{\mathbf{s}}$ is a complex Lagrange multiplier, $\mu > 0$, and we define $\mathbf{g}_w = \mathbf{w}\cdot\mathbf{g}$ for compact notation. Moreover, the equivalence in (12) follows from the Parseval theorem, where the operator $()^H$ is conjugate transpose, the operator $\hat{\mathbf{a}} = vec\left(\mathcal{F}(\mathbf{a})\right)$ is a DFT transform and reshape into a column vector, i.e., $\mathbf{a} \in \mathbb{R}^{D\times 1}$, with $D = w\times h$, while $\overline{()}$ is conjugation operation.

For the purposes of derivation we rewrite (12) into a fully vectorized form

$$\begin{aligned}\mathcal{L}\left(\hat{\mathbf{h}}, \mathbf{g}, \hat{\mathbf{s}}\right) &= \left\| \hat{\mathbf{h}}^H diag\left(\hat{\mathbf{x}}_t\right) - \hat{\tilde{\mathbf{y}}} \right\|_2^2 + \lambda \left\| \mathbf{g}_w \right\|_2^2 \\ &+ \left[\hat{\mathbf{s}}^H \left(\hat{\mathbf{h}} - \sqrt{D}\mathbf{F}\mathbf{W}\mathbf{g}\right) + \overline{\hat{\mathbf{s}}^H \left(\hat{\mathbf{h}} - \sqrt{D}\mathbf{F}\mathbf{W}\mathbf{g}\right)} \right] \end{aligned}$$

$$+\mu \left\| \hat{\mathbf{h}} - \sqrt{D}\mathbf{F}\mathbf{W}\mathbf{g} \right\|_2^2, \tag{13}$$

where $\mathbf{F}$ denotes $D \times D$ orthonormal matrix of Fourier coefficients, such that the Fourier transform is defined as $\hat{\mathbf{x}} = \mathcal{F}(\mathbf{x}) = \sqrt{D}\mathbf{F}\mathbf{x}$ and $\mathbf{W} = diag(\mathbf{w})$. For simplicity, we denote the (13) into four terms as

$$\mathcal{L}\left(\hat{\mathbf{h}}, \mathbf{g}, \hat{\mathbf{s}}\right) = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4, \tag{14}$$

where

$$\begin{aligned}
\mathcal{L}_1 &= \left(\hat{\mathbf{h}}^H diag\left(\hat{\mathbf{x}}_t\right) - \hat{\bar{\mathbf{y}}}\right) \overline{\left(\hat{\mathbf{h}}^H diag\left(\hat{\mathbf{x}}_t\right) - \hat{\bar{\mathbf{y}}}\right)}^{\top}, \\
\mathcal{L}_2 &= \|\mathbf{g}_w\|_2^2, \\
\mathcal{L}_3 &= \left[\hat{\mathbf{s}}^H \left(\hat{\mathbf{h}} - \sqrt{D}\mathbf{F}\mathbf{W}\mathbf{g}\right) + \overline{\hat{\mathbf{s}}^H \left(\hat{\mathbf{h}} - \sqrt{D}\mathbf{F}\mathbf{W}\mathbf{g}\right)}\right], \\
\mathcal{L}_4 &= \mu \left\| \hat{\mathbf{h}} - \sqrt{D}\mathbf{F}\mathbf{W}\mathbf{g} \right\|_2^2.
\end{aligned} \tag{15}$$

We then employ the ADMM algorithm that alternatively solves the following subproblems,

$$\begin{cases}
\hat{\mathbf{h}}^{i+1} = \arg\min_{\mathbf{h}} \mathcal{L}\left(\hat{\mathbf{h}}, \mathbf{g}^i, \hat{\mathbf{s}}^i\right), \\
\mathbf{g}^{i+1} = \arg\min_{\mathbf{g}} \mathcal{L}\left(\hat{\mathbf{h}}^{i+1}, \mathbf{g}, \hat{\mathbf{s}}^i\right), \\
\hat{\mathbf{s}}^{i+1} = \hat{\mathbf{s}}^i + \mu\left(\hat{\mathbf{h}}^{i+1} - \mathbf{g}^{i+1}\right).
\end{cases} \tag{16}$$

Subproblem $\hat{\mathbf{h}}$: Minimizer of $\hat{\mathbf{h}}$ is derived by setting its complex gradient of the augmented Lagrangian to zero as

$$\frac{\partial \mathcal{L}_1}{\partial \bar{\hat{\mathbf{h}}}} + \frac{\partial \mathcal{L}_2}{\partial \bar{\hat{\mathbf{h}}}} + \frac{\partial \mathcal{L}_3}{\partial \bar{\hat{\mathbf{h}}}} + \frac{\partial \mathcal{L}_4}{\partial \bar{\hat{\mathbf{h}}}} = \mathbf{0}, \tag{17}$$

where the partial complex gradients are:

$$\begin{aligned}
\frac{\partial \mathcal{L}_1}{\partial \bar{\hat{\mathbf{h}}}} &= diag\left(\hat{\mathbf{x}}_t\right) diag\left(\hat{\mathbf{x}}_t\right)^H \hat{\mathbf{h}} - diag\left(\hat{\mathbf{x}}_t\right) \hat{\mathbf{Q}}_{\mathbf{y}}^{\mathbf{H}}, \\
\frac{\partial \mathcal{L}_2}{\partial \bar{\hat{\mathbf{h}}}} &= \mathbf{0}, \\
\frac{\partial \mathcal{L}_3}{\partial \bar{\hat{\mathbf{h}}}} &= \hat{\mathbf{s}}, \\
\frac{\partial \mathcal{L}_4}{\partial \bar{\hat{\mathbf{h}}}} &= \mu\left(\hat{\mathbf{h}} - \sqrt{D}\mathbf{F}\mathbf{W}\mathbf{g}\right),
\end{aligned} \tag{18}$$

where $\sqrt{D}\mathbf{F}\mathbf{W}\mathbf{g} = \hat{\mathbf{g}}_w$ according to our original definition of $\hat{\mathbf{g}}_w$. Putting (18) into (17), we have

$$\hat{\mathbf{h}} = \frac{\hat{\mathbf{x}}_t \cdot \hat{\bar{\mathbf{y}}} + \mu \hat{\mathbf{g}}_w - \hat{\mathbf{s}}}{\hat{\mathbf{x}}_t \cdot \bar{\hat{\mathbf{x}}}_t + \mu}, \tag{19}$$

whose iteration form is

$$\hat{\mathbf{h}}^{i+1} = \frac{\hat{\mathbf{x}}_t \cdot \hat{\bar{\mathbf{y}}} + \mu \hat{\mathbf{g}}_m^i - \hat{\mathbf{s}}^i}{\hat{\mathbf{x}}_t \cdot \bar{\hat{\mathbf{x}}}_t + \mu^i}, \tag{20}$$

where the constraint penalty $\mu^{i+1} = \beta\mu^i$.

Subproblem $\mathbf{g}$: Similarly, we set its complex gradient equal to zero

$$\frac{\partial \mathcal{L}_1}{\partial \bar{\mathbf{g}}} + \frac{\partial \mathcal{L}_2}{\partial \bar{\mathbf{g}}} + \frac{\partial \mathcal{L}_3}{\partial \bar{\mathbf{g}}} + \frac{\partial \mathcal{L}_4}{\partial \bar{\mathbf{g}}} = \mathbf{0}, \tag{21}$$

and the partial gradients are

$$\begin{aligned}
\frac{\partial \mathcal{L}_1}{\partial \bar{\mathbf{g}}} &= \mathbf{0}, \\
\frac{\partial \mathcal{L}_2}{\partial \bar{\mathbf{g}}} &= \lambda \mathbf{W}\mathbf{g}, \\
\frac{\partial \mathcal{L}_3}{\partial \bar{\mathbf{g}}} &= -\sqrt{D}\mathbf{W}\mathbf{F}^H \hat{\mathbf{s}}, \\
\frac{\partial \mathcal{L}_4}{\partial \bar{\mathbf{g}}} &- \mu\left(\sqrt{D}\mathbf{W}\mathbf{F}^H \hat{\mathbf{h}} - D\mathbf{W}\mathbf{g}\right).
\end{aligned} \tag{22}$$

Putting (22) into (21), yielding

$$\begin{aligned}
\lambda \mathbf{W}\mathbf{g} - \sqrt{D}\mathbf{W}\mathbf{F}^H \hat{\mathbf{s}} - \mu\sqrt{D}\mathbf{W}\mathbf{F}^H \hat{\mathbf{h}} + \mu D\mathbf{W}\mathbf{g} = \mathbf{0}, \\
\mathbf{W}\mathbf{g} = \mathbf{W}\frac{\sqrt{D}\mathbf{F}^H\left(\hat{\mathbf{s}} + \mu\hat{\mathbf{h}}\right)}{\lambda + \mu D}.
\end{aligned} \tag{23}$$

According to the definition of the inverse DFT, i.e., $\mathcal{F}^{-1}\left(\hat{\mathbf{x}}\right) = \frac{1}{\sqrt{D}}\mathbf{F}^H\hat{\mathbf{x}}$, and the values in $\mathbf{w}$ are not zeros, so the solution to (23) is

$$\mathbf{g} = \frac{\mathcal{F}^{-1}\left(\hat{\mathbf{s}} + \mu\hat{\mathbf{h}}\right)}{\frac{\lambda}{D} + \mu}, \tag{24}$$

and the iteration form is

$$\mathbf{g}^{i+1} = \frac{\mathcal{F}^{-1}\left(\hat{\mathbf{s}}^i + \mu^i \hat{\mathbf{h}}_c^{i+1}\right)}{\frac{\lambda}{D} + \mu^i}. \tag{25}$$

After achieving the convergence of $\mathbf{g}$ as $\mathbf{g}^n$, the solution of the learned CF is $\mathbf{f} = \mathbf{g}^n + \mathbf{f}_{t-1}$.

Finally, we have the response map of the learned CFs $\{\mathbf{f}^d\}_{d=1}^D$ as

$$\mathbf{r}_{cf} = \sum_{d=1}^{D} \mathcal{F}^{-1}(\hat{\mathbf{x}}_t \cdot \hat{\mathbf{f}}^d) \tag{26}$$

Different from the standard CF learning in Staple [11] that only employs the standard CF tracker, we combine the spatial and temporal regularizers into one term to constrain the learned CFs. When the target object suffers from occlusion, Staple will learn a corrupted CF, but our method can alleviate this over-fitting problem via spatio-temporally regluarizing the CFs to keep it close to the previous models.

C. MERGING STRATEGY

Finally, combining the color histogram response and the spatio-temporal regularized correlation filter response in a linear way, we obtain the final location response

$$\mathbf{r} = \eta \mathbf{r}_{cc} + (1 - \eta)\mathbf{r}_{cf}, \tag{27}$$

where $\mathbf{r}_{cc}$ is the clustering color histogram response, $\mathbf{r}_{cf}$ is the spatio-temporal regularized CF response and η is a merging factor, and the tracked location is determined by maximizing $\mathbf{r}$.

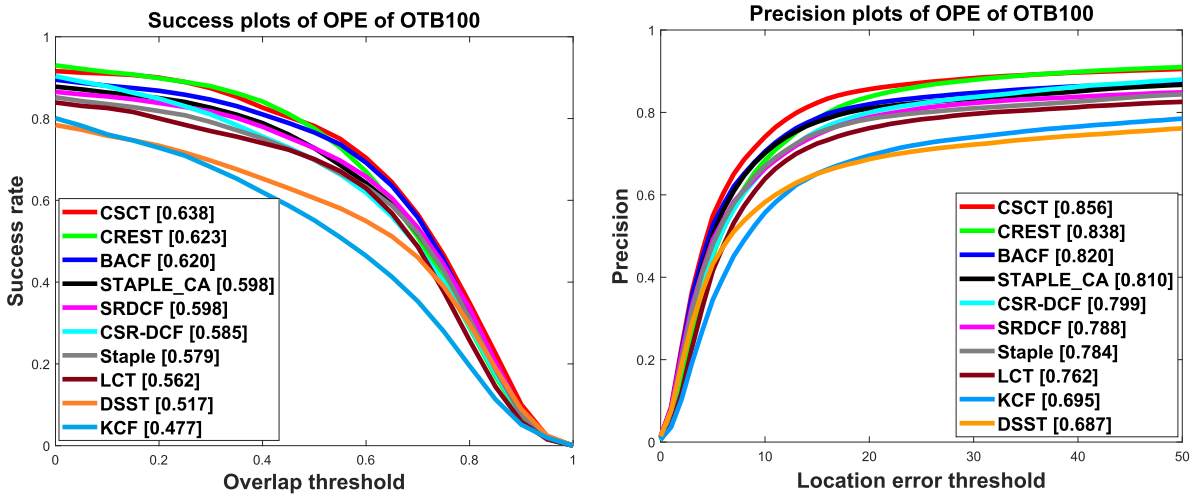


FIGURE 3. Success and precision plots on OTB100. The legend of the success plot reports the AUC scores and the legend of the precision plot reports the distance precision scores for the threshold at 20 pixels.

IV. EXPERIMENTAL RESULTS

A. IMPLEMENTATION DETAILS

We set the color cluster learning factor $\eta_{cc} = 0.04$, the cluster center number $n = 16$,which is chosen according to a simple experiment on the clustering center numbers, and the merge factor $\eta = 0.3$, and $s = 42$-dimensional HOG+color naming (CN) features are used to represent the targets. The features are further weighted by a cosine window to reduce the boundary discontinuities. As for the ADMM algorithm, we set the hyper-parameter in (8) to $\lambda = 0.01$ and the step size parameter in (20) and (25) $\beta = 2$, $\mu^0 = 1$ throughout all the experiments. We extensively evaluate our method on 3 popular benchmarks including OTB100 [12], Temple Color [13] and VOT2016 [14]. The proposed CSCT is implemented in MATLAB 2015a and runs at 18 fps on a PC with Intel i7-4790 CPU (3.6 GHz) and 16 GB RAM memory.

B. QUANTITATIVE EVALUATION

1) OTB-100 DATASET

a: OVERALL PERFORMANCE

Figure 3 shows the success and precision plots of 10 representative state-of-the-art trackers on OTB100, including the proposed CSCT, Staple (CVPR2016) [11], DSST (BMVC2014) [33], LCT (CVPR2015) [18], SRDCF (ICCV2015) [8], BACF (ICCV2017) [34], CREST (ICCV2017) [35], CSR-DCF (CVPR2017) [17], KCF (T-PAMI2015) [7] and STAPLE_CA (CVPR2017) [9]. We exploit one-pass evaluation (OPE) for all trackers and report both the precision and success plots for comparison. Following [12], in the precision plots, we use the distance precision rate at threshold 20 for ranking, while in the success plots, we use the area under curve (AUC) score for ranking.

Figure. 3 shows the results of all compared trackers. Among them, the proposed CSCT achieves the best overall performance in terms of both success and precision rates with an AUC score of 63.8% and a precision score of 85.6%, outperforming the second-best method CREST by 1.5% and 1.8%. It is noted that different from our method which only uses the HOG and CN features, CREST employs several layers of deep CNN-based features, leading to a slow implementation, but still performs worse than our method, demonstrating the effectiveness of the complementary advantages between the color clustering model and the spatio-temporal regularized model in our CSCT. Moreover, the comparison shows that CSCT outperforms the other real-time trackers by a large margin. As an instance, in the precision plots, our method improves the second best real-time tracker BACF by 3.6%, while in the success plots, our tracker has a gain of 1.8% compared to BACF.

b: ATTRIBUTE-BASED PERFORMANCE

The targets in OTB100 mainly suffer from 11 attributes of challenging factors including illumination variation (IV), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), out-of-view (OV), background clutters (BC), in-plane rotation (IPR), out-of-plane rotation (OPR) and low resolution (LR). To facilitate analyzing the strength and weakness of the proposed approach, we further evaluate the trackers on videos with these 11 attributes. Figure. 4 shows the success plots of videos with various attributes, while Figure. 5 shows the corresponding precision plots. Among them, CSCT ranks within top 3 on all 11 attributes in the success plots, and outperforms the others on 7 out of 11 attributes. In the precision plots, CSCT ranks top 1 on 7 out of 11 attributes. Since the AUC score of the success plot is much more accurate than the score at one position in the precision plot, as in [36], in the following we mainly analyze the ranked results based on the success plots.

On the videos with attributes such as SV, OCC, DEF, BC, MB, FM and OV, CSCT ranks 1st among all the evaluated

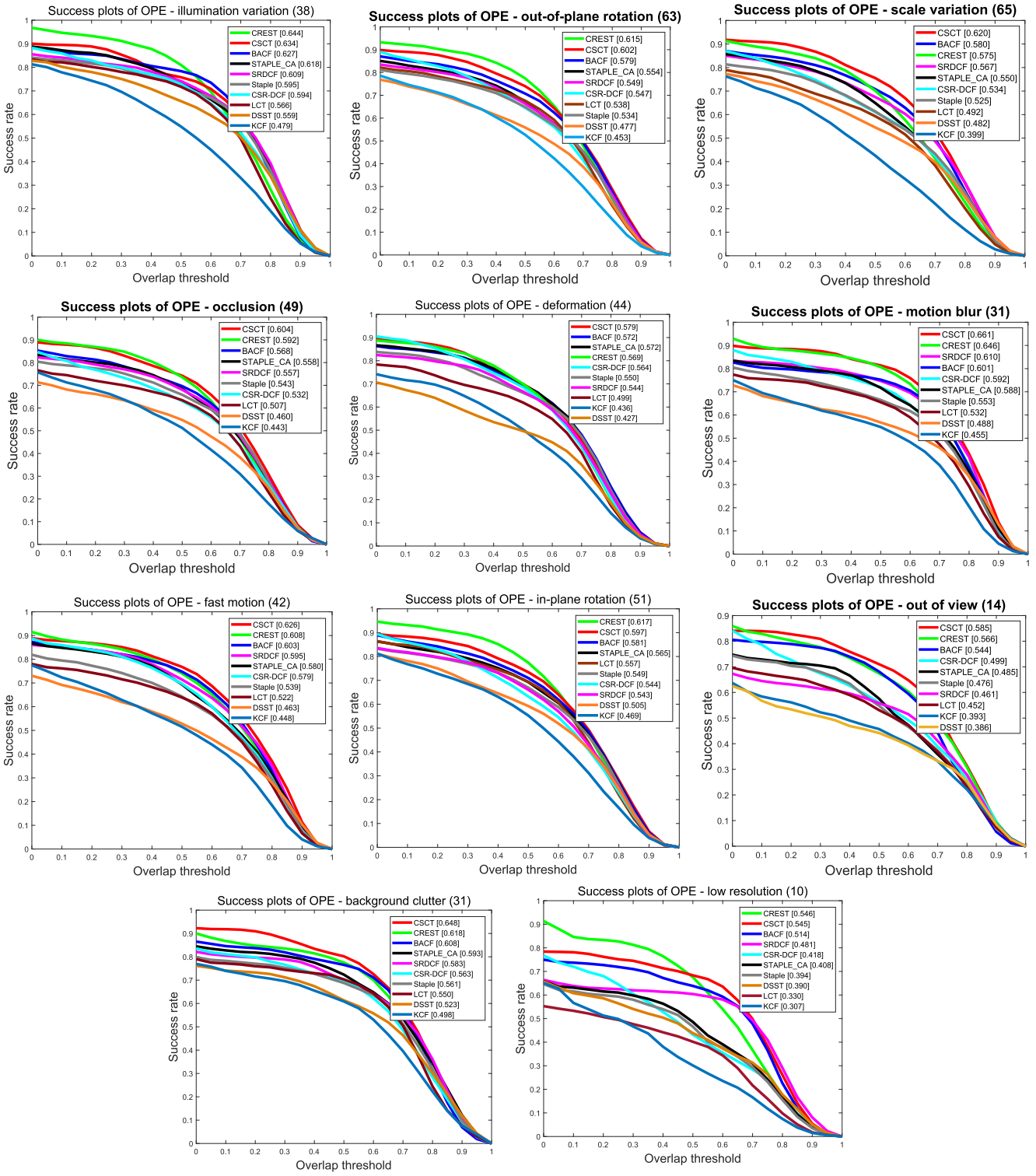


FIGURE 4. Success plots of videos with various attributes.

trackers. All these methods ignore the structure information of the color-based representation and the variations among a serious of learned CFs in consecutive frames. However, to enhance the representation capability of the color-based model, our CSCT leverages color clustering strategy to learn a data-adaptive color histogram, leading to a more robust appearance model. Moreover, to improve the discriminative capacity of the learned CFs, CSCT employs the variations

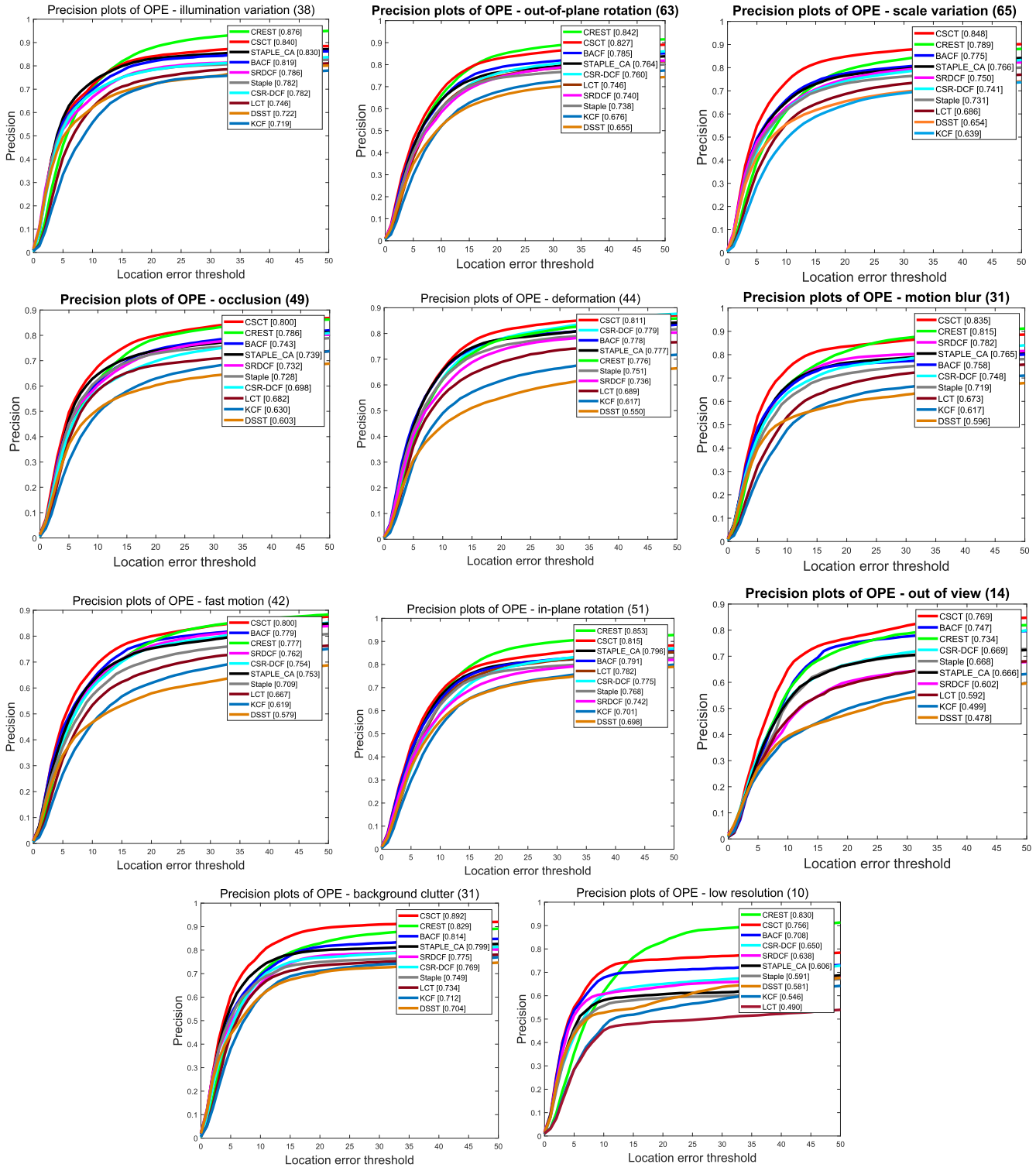


FIGURE 5. **Precision plots of videos with various attributes.**

between the CFs learned from consecutive frames as regularization to learn a more reliable CF model. The experimental results demonstrate that the improved color clustering model along with the modified spatio-temporal regularized CF have a positive effect on handling the challenging factors with the above-mentioned attributes.

On the videos with LR, IPR, OPR and IV, CSCT ranks 2nd among all evaluated algorithms with a narrow margin to the

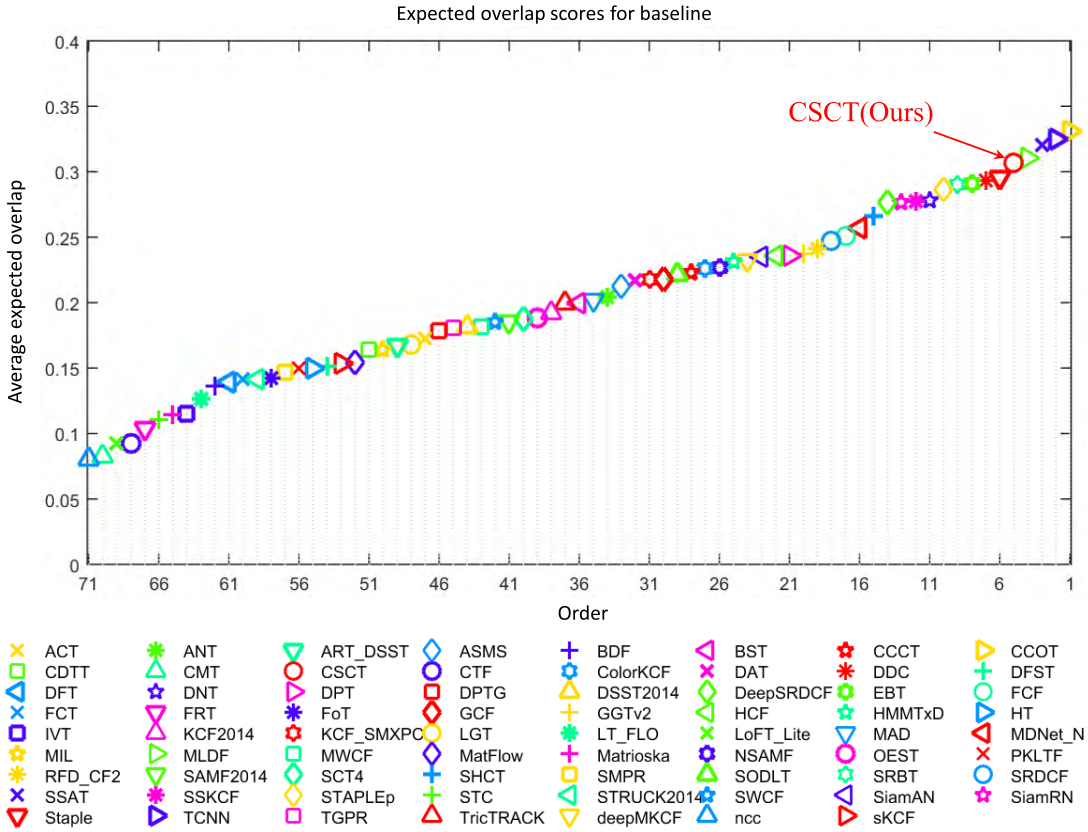


FIGURE 6. Expected Average Overlap (EAO) order plot on VOT2016. The better trackers are located at the right. The EAO measure, computed as the average EAO over typical sequence lengths, is displayed in the legend (see [14] for details).

TABLE 1. Average AUC of the 10 trackers in terms of 11 attributes: top three results are in red, blue and green fonts. Best viewed in color model.

Attri.	BACF	CSR-DCF	STAPLE_CA	DSST	CREST	LCT	Staple	KCF	SRDCF	CSCT
LR	**0.514**	0.418	0.408	0.390	**0.546**	0.330	0.394	0.307	0.481	**0.545**
IPR	**0.581**	0.544	0.565	0.505	**0.617**	0.557	0.549	0.469	0.543	**0.597**
OPR	**0.579**	0.547	0.554	0.477	**0.615**	0.538	0.534	0.453	0.549	**0.602**
SV	**0.580**	0.534	0.550	0.482	**0.575**	0.492	0.525	0.399	0.567	**0.620**
OCC	**0.568**	0.532	0.558	0.460	**0.592**	0.507	0.543	0.443	0.557	**0.604**
DEF	**0.572**	0.564	**0.572**	0.427	0.569	0.499	0.550	0.436	0.544	**0.579**
BC	**0.608**	0.563	0.593	0.523	**0.618**	0.550	0.561	0.498	0.583	**0.648**
IV	**0.627**	0.594	0.618	0.559	**0.644**	0.566	0.595	0.479	0.609	**0.634**
MB	0.601	0.592	0.588	0.488	**0.646**	0.532	0.553	0.455	**0.610**	**0.661**
FM	**0.603**	0.579	0.580	0.463	**0.608**	0.522	0.539	0.448	0.595	**0.626**
OV	**0.544**	0.499	0.485	0.386	**0.566**	0.452	0.476	0.393	0.461	**0.585**

top tracker. Due to the sudden appearance variations (i.e. the illumination changes) limit the color clustering learning and spatio-temporal CF learning, CSCT cannot employ a more robust appearance model than the deep CNN-based appearance in CREST. Over time, it will cause the object drift, making CSCT unable to perform well on these attributes.

Finally, Table 1 reports the AUC scores of all compared trackers under these attributes. Among them, our CSCT achieves the highest AUC scores in 7 out of the total 11 attributes. In the attributes of IPR, OPR and IV, CSCT achieves favorable performance against CREST, BACF and STAPLE_CA that achieve the top-3 best performance.

However, when the target undergoes LR, CSCT cannot perform as well as CREST. This is mainly because the fact that the HOG along with CN features cannot well represent the texture information of the target with low resolution. In general, CSCT significantly outperforms the others on most attributes listed by Table 1.

2) TEMPLE-COLOR DATASET

We perform comparative experiments on Temple-Color dataset [13] which consists of 128 color sequences. We compare CSCT with the state-of-the-art trackers mentioned above including Staple (CVPR 2017) [11],

TABLE 2. Comparing CSCT with the 7 trackers using EAO, accuracy rate (Acc), robustness rate (Rob), no-reset average overlap (Ao) and average speed (*fps*) on the VOT2016 benchmark. The top three results are in red, blue and green fonts. Best viewed in color display.

Tracker	Struck	CSR-DCF	SRDCF	EBT	Staple	KCF	CREST	CSCT
Where	ICCV11	CVPR17	ICCV15	CVPR16	CVPR16	PAMI15	ICCV17	Ours
EAO	0.142	**0.338**	0.247	0.291	**0.295**	0.192	0.283	**0.310**
Acc	0.439	0.510	**0.535**	0.441	**0.545**	0.491	0.514	**0.524**
Rob	3.37	**0.85**	1.50	**0.92**	1.35	2.01	**1.08**	1.10
AO	0.242	0.376	**0.397**	0.370	0.388	0.301	**0.435**	**0.420**
Speed	6	13	4	6	**52**	**107**	2	**18**

fDSST (T-PAMI 2017) [37], LCT (CVPR 2015) [18], SRDCF (ICCV 2015) [8], MEEM (ECCV 2014) [16], BACF (ICCV 2017) [34], HCFT (ICCV 2015) [26], KCF (T-PAMI 2015) [7] and STAPLE_CA (CVPR 2017) [9]. Fig. 7 shows the comparisons of overlap success plots for different trackers. We note that CSCT outperforms STAPLE_CA and Staple by 2.0% and 3.4%, respectively, further demonstrating the effectiveness of the color clustering histogram and spatio-temporal regularization strategies.

3) VOT2016 DATASET

In Table 2, we compare our CSCT with several top trackers on the VOT 2016 [14] benchmark, including CREST [35], EBT [38], CSR-DCF [17], Staple [11], SRDCF [8], Struck [39] and KCF [7]. As suggested by [14], Table 2 reports the results of the evaluated trackers in terms of EAO (estimated average overlap), Acc (accuracy), Rob (robustness), AO (average overlap) and speed (in *fps*). Among these methods, CSR-DCF achieves the best result under EAO metric. Meanwhile, CSCT achieves the second-best performance with a EAO score of 0.310 and performs at a speed of 18 *fps* that is faster than CSR-DCF with 13 *fps*. According to the analysis of the VOT report [14], the EAO score of CSCT is 0.310 that outperforms the definition of the strict state-of-the-art bound 0.251 by 5.9%, and thus it can be regarded as state-of-the-art.

Figure. 6 displays the EAO ranking orders of the compared 70 tracking methods that participate in the VOT2016 challenge. Our proposed CSCT ranks 5th in the ranking order plot and outperforms the other real-time trackers (like Staple and DAT) by a large margin. Among the four trackers that perform better than ours, CCOT and TCNN both apply complex continuous or tree-structured convolution operator to improve their tracking accuracy, resulting in an inferior speed compared to our CSCT. Moreover, it is noted that SSAT is an extended version of MDNet tracker, which utilizes OTB and VOT benchmark for pre-training. And MLDF also combines low, mid and high-level features from the pre trained VGG networks, which is complex and slower than our CSCT.

C. QUALITATIVE EVALUATION

Fig. 8 qualitatively compare the results of the top performing trackers: CSR-DCF (CVPR2017) [17], Staple (CVPR2016) [11], BACF (ICCV2017) [34], CREST (ICCV2017) [35] and CSCT on 10 challenging sequences. In a majority of these sequences, CSR-DCF fails to locate the targets or

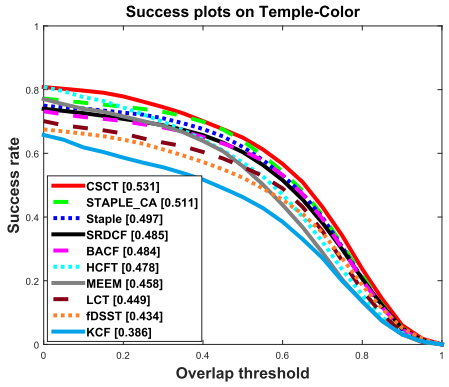


FIGURE 7. Overlap success plots of different trackers on Temple-Color. Only 10 trackers are displayed for clarity.

estimates scale incorrectly because of the limited performance of the spatial regularized CF framework. CREST reformulates DCFs as a one-layer convolutional neural network and integrates feature extraction, response map generation as well as model update into neural networks for an end-to-end training. It performs well on the attributes of LR, IPR, OPR and IV. However, the classifier of CREST is trained to focus on the residual of appearance changes, which may lead to overfitting in presence of severe deformation and heavy occlusion. As a result, it does not perform well in handling deformation (e.g. *Bolt2*) and occlusion (e.g. *Box*). The CF based trackers (e.g. Staple and BACF) improve conventional CFs by leveraging both the strength of CF-based model and color-based model, or modeling foreground and background of the object over time, respectively. Nevertheless, they do not take full advantage of the target color information and ignore the temporal variation of the learned CFs in consecutive frames. In contrast, our CSCT leverages target color distribution to learn a more reliable color-based model and incorporates both spatial and temporal regularization into the CF framework, as a result, achieving a robust performance against a variety of challenging factors in visual tracking.

D. ABLATION STUDIES

To verify the effectiveness of our color-based improvement (refer to Section III-A) and CF-based improvement (refer to Section III-B), we report the resluts using either only color clustering-based improvement (CCI) or only CF-based improvement (CFI), demonstrating how much the CCI or CFI

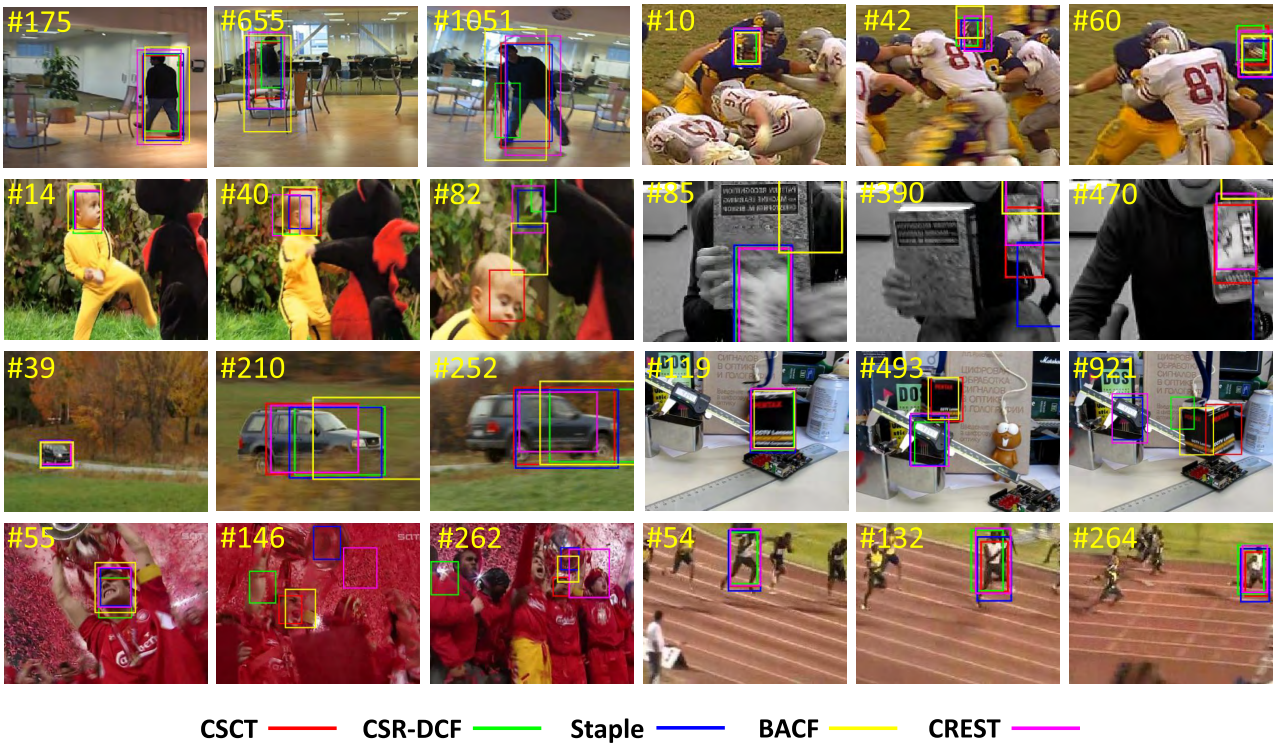


FIGURE 8. Qualitative evaluations of our CSCT, CSR-DCF [17], Staple [11], BACF [34], CREST [35] on 8 challenging sequences (from left to right and top to down: *Human2*, *Football1*, *Dragonbaby*, *Cliffbar*, *Carscale*, *Box*, *Soccer* and *Bolt2*, respectively). Our CSCT performs favorably against state-of-the-art.

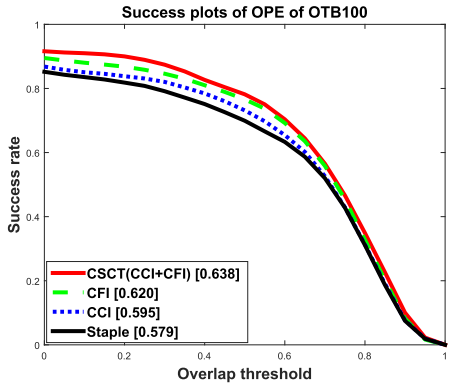


FIGURE 9. Ablative experiments on OTB100. CCI, CFI denote only color clustering-based improvement and only CF-based improvement, respectively. Staple [11] is the baseline tracker, while CSCT is the combination of CCI and CFI. The AUC score for each tracker is shown in the legend.

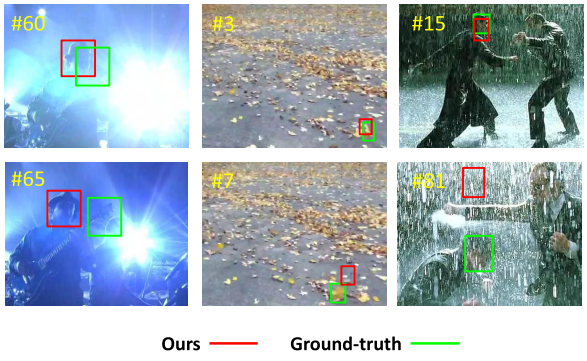


FIGURE 10. Failure cases of the proposed tracker, where we utilize red and green bounding boxes to denote our results and ground-truths.

contributes to the overall performance of CSCT. We employ OTB100 for test and compare with CSCT, CCI, CFI, and the baseline Staple [11]. Figure 9 shows the comparative results in terms of the AUC scores of OPE of success rates. Specifically, CFI outperforms Staple by 4.1%, demonstrating the effectiveness of the spatio-temporal regularized term in CFI. Besides, CCI cannot perform favorably as CFI mainly because of the limitation of the color-based feature. Finally, the CSCT (CCI+CFI) boosts the performance significantly with an AUC score of 0.638 that outperforms the baseline Staple by 5.9%, demonstrating the complementary advantages of these two mechanisms in CSCT.

E. FAILURE CASES

We show some failure cases of the proposed method in Fig 10. In the first column, the glaring foreground and background regions contain few colors, limiting the discriminative ability of our color clustering model. Furthermore, in the second video, our method fails to track the leaf as it suffers from low resolution. When calculating the clustering centers, the low resolution-based frame cannot contain enough color

information to train an effective color histogram model. Although the spatio-temporally regularized CF is learned, in some low resolution sequences, only the CF model cannot solve the severe appearance variation problem. Namely, our CSCT loses the color complementary merit. In the third sequence, the target person has drastic motion and undergoes severe illumination variation, the failure of our method is due to the lack of enough motion information (like the optical flow), which will be taken into account in our future work.

V. CONCLUSION

In this paper, we have proposed a novel complementary tracking algorithm via dual color clustering and spatio-temporal regularized correlation regressions. Specifically, to overcome the noisy interference of the standard color histogram for visual tracking, we cluster the color channels from the ground-truth target in the first frame by the K-means algorithm, yielding a data-adaptive non-uniform quantizer to design a robust color histogram, resulting in a more robust color-based model. Moreover, to alleviate the drift problem from sudden appearance variations and make better use of the frame-wise temporal information, we propose a novel spatio-temporal regularization to learn robust CFs. Finally, both the color clustering histogram based model and the spatio-temporal regularized CF model are linearly combined to yield a robust appearance model for our CSCT. Extensive experimental results on three popular benchmarks demonstrate that the proposed CSCT achieves favourable performance against several state-of-the-art tracking algorithms. In the future work, we will merge two or several simple trackers to obtain a more powerful tracking algorithm, while running even faster. Also, more motion information (eg. the optical flow) is encouraged to our future improvements of the proposed CSCT.

REFERENCES

[1] S. Zhang, H. Zhou, F. Jiang, and X. Li, "Robust visual tracking using structurally random projection and weighted least squares," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1749–1760, Nov. 2015.

[2] S. Zhang, H. Yao, X. Sun, and X. Lu, "Sparse coding based visual tracking: Review and experimental comparison," *Pattern Recognit.*, vol. 46, no. 7, pp. 1772–1788, Jul. 2013.

[3] K. Zhang, L. Zhang, and M. Yang, "Fast compressive tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2002–2015, Oct. 2014.

[4] K. Zhang, Q. Liu, Y. Wu, and M.-H. Yang, "Robust visual tracking via convolutional networks without training," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1779–1792, Apr. 2016.

[5] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 127–141.

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

[7] 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.

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

[9] 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.

[10] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang. (2018). "Learning spatial-temporal regularized correlation filters for visual tracking," [Online]. Available: https://arxiv.org/abs/1803.08679

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

[12] 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.

[13] P. Liang, E. Blasch, and H. Ling, "Encoding color information for visual tracking: Algorithms and benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5630–5644, Dec. 2015.

[14] M. Kristan *et al.*, "The visual object tracking VOT2016 challenge results," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2016, pp. 777–823.

[15] H. Possegger, T. Mauthner, and H. Bischof, "In defense of color-based model-free tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2113–2120.

[16] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 188–203.

[17] A. Lukezic, T. Vojíř, L. Č. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 8, Jun. 2017, pp. 6309–6318.

[18] 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.

[19] J. Ning, J. Yang, S. Jiang, L. Zhang, and M.-H. Yang, "Object tracking via dual linear structured SVM and explicit feature map," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4266–4274.

[20] J. Valmadre, L. Bertinetto, J. A. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5000–5008.

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

[22] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2002, pp. 661–675.

[23] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2006, pp. 798–805.

[24] 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. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1090–1097.

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

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

[27] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1430–1438.

[28] 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.* Springer, 2016, pp. 472–488.

[29] J. Kwon and K. Mu Lee, "Visual tracking decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1269–1276.

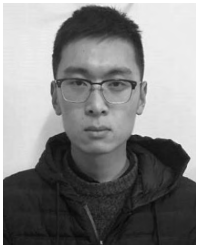
[30] J. Kwon and K. Mu Lee, "Tracking by sampling trackers," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1195–1202.

[31] N. Wang and D.-Y. Yeung, "Ensemble-based tracking: Aggregating crowdsourced structured time series data," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1107–1115.

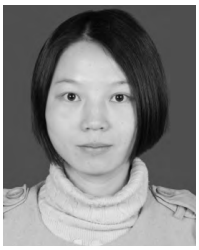
[32] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[33] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, Nottingham, U.K., Sep. 2014, pp. 1–11.

[34] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jun. 2017, pp. 1–9.
[35] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M.-H. Yang, "Crest: Convolutional residual learning for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2555–2564.
[36] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 2411–2418.
[37] 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.
[38] G. Zhu, F. Porikli, and H. Li, "Beyond local search: Tracking objects everywhere with instance-specific proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 943–951.
[39] 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.



JIAQING FAN is currently pursuing the M.S. degree with the School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China. His current research interests include visual object tracking algorithms.



HUIHUI SONG received the B.S. degree in technology and science of electronic information from the Ocean University of China in 2008, the master's degree in communication and information system from the University of Science and Technology of China in 2011, and the Ph.D. degree in geography and resource management from The Chinese University of Hong Kong in 2014. She is currently a Professor with the Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing, China. Her research interests include remote sensing image processing and image fusion.

KAIHUA ZHANG received the B.S. degree in technology and science of electronic information from the Ocean University of China in 2006, the M.S. degree in signal and information processing from the University of Science and Technology of China in 2009, and the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, in 2013. From 2009 to 2010, he was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University. He is currently a Professor with the School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China. His research interests include image segmentation, level sets, and visual tracking.

QINGSHAN LIU received the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academic of Science, Beijing, China, in 2003, and the M.S. degree from the Department of Auto Control, Southeast University, Nanjing, in 2000. He was an Associate Professor with the National Laboratory of Pattern Recognition, Chinese Academic of Science, and an Associate Researcher with the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong, from 2004 and 2005. He was an Assistant Research Professor with the Department of Computer Science, Computational Biomedicine Imaging and Modeling Center, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA, from 2010 to 2011. He is currently a Professor with the School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China. He was a recipient of the President Scholarship of the Chinese Academy of Sciences in 2003. His current research interests are image and vision analysis, including face image analysis, graph and hypergraph-based image and video understanding, medical image analysis, and event-based video analysis.

WEI LIAN received the B.S. degree in automation from the Taiyuan University of Technology, Taiyuan, China, in 2000, and the M.S. and Ph.D. degrees in automatic control theory and engineering from Northwestern Polytechnical University, Xi'an, China, in 2003 and 2007, respectively. He was a Research Assistant/Associate with the Department of Computing, The Hong Kong Polytechnic University, from 2007 to 2016. He is currently an Associate Professor with the Department of Computer Science, Changzhi University, Changzhi, China.