Complementary Discriminative Correlation Filters Based on Collaborative Representation for Visual Object Tracking

Xue-Feng Zhu, Xiao-Jun Wu, *Member, IEEE* Tianyang Xu, Zhen-Hua Feng, *Member, IEEE* and Josef Kittler, *Life Member, IEEE*

***Abstract*—In recent years, discriminative correlation filter (DCF) based algorithms have significantly advanced the state of the art in visual object tracking. The key to the success of DCF is an efficient discriminative regression model trained with powerful multi-cue features, including both hand-crafted and deep neural network features. However, the tracking performance is hindered by their inability to respond adequately to abrupt target appearance variations. This issue is posed by the limited representation capability of fixed image features. In this work, we set out to rectify this shortcoming by proposing a complementary representation of a visual content. Specifically, we propose the use of a collaborative representation between successive frames to extract the dynamic appearance information from a target with rapid appearance changes, which results in suppressing the undesirable impact of the background. The resulting collaborative representation coefficients are combined with the original feature maps using a spatially regularised DCF framework for performance boosting. The experimental results on several benchmarking datasets demonstrate the effectiveness and robustness of the proposed method, as compared with a number of state-of-the-art tracking algorithms.**

***Index Terms*—Visual object tracking, discriminative correlation filter, feature representation, collaborative representation**

I. Introduction

Visual object tracking is a fundamental research topic in computer vision and pattern recognition, with many practical applications in CCTV surveillance, human-computer interaction and robot vision. Given the initial state of a target of interest, the task of visual object tracking is to detect and localise the target in the subsequent video frames automatically. Through decades of research, a significant progress has been made in this field. However, due to a number of challenging factors such as background clutter, illumination variation, scale variation, camera motion, partial occlusion and blur, it is still a very challenging task to achieve efficient and robust tracking in unconstrained scenarios.

During the past decades, a variety of tracking algorithms have been developed [1], [2], [3], [4]. These approaches can be divided into two main categories: generative methods and discriminative methods. A generative method tends to establish a target appearance model, with which the tracker finds the best matching region as the target location. In contrast, discriminative methods cast the tracking task as a classification or regression problem. The classifier is trained by extracting training samples from the target and background, with the objective of achieving separability between them.

Recently, owing to the superior speed and performance, Discriminative Correlation Filters (DCF) has been widely studied as a general framework for online visual tracking [5], [6], [7], [4]. Generally, DCF realises object tracking by designing a discriminative filter, which produces high response in the area of the target and low response in the background. The tracking performance of DCF has recently been further improved by using spatio-temporal appearance modelling [8], [9], [10] and multi-dimensional features, including deep neural network features.

In spite of the potential of hand-crafted and deep neural network features to provide a powerful representation of the target, their effectiveness is still restricted, due to the rigidity of the process used in modelling the target visual appearance. In particular, the capability and descriptive power of such feature descriptors tend to degenerate in the presence of large appearance variations. Moreover, existing DCF-based methods utilise a standard tracking-learning-detection paradigm with a pre-defined updating rate to formulate the online tracking problem, without considering the spatio-temporal continuity of content in natural video sequences. To address the above issues, we propose a novel mechanism for extracting an effective target model by introducing collaborative representation into the DCF formulation, enabling the learning of adaptive contextual features for robust visual tracking.

Regarding the first issue, existing feature extraction mechanisms inadvertently impose limitations on the performance of trained models, especially in the scenarios encountering large appearance changes. The pioneering solutions favoured plain models, for instance using grayscale features [11], [12], which achieve high tracking speed, but seriously compromise robustness. Later, Henriques *et al.* advocated HOG features for visual tracking [13], which help to learn a more discriminative

DOI: 10.1109/TCSVT.2020.2979480

X.-F Zhu and X.-J Wu (corresponding author) are with the School of Internet of Things Engineering, Jiangnan University, Wuxi, P.R. China. (e-mail: xuefeng_zhu95@163.com, wu_xiaojun@jiangnan.edu.cn)

T. Xu is with the School of Internet of Things Engineering, Jiangnan University, Wuxi, P.R. China and the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. (e-mail: tianyang_xu@163.com)

Z.-H. Feng and J. Kittler are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. (e-mail: {z.feng; j.kittler}@surrey.ac.uk)

appearance model in a kernel space. Danelljan *et al.* [14] proposed the use of Colour Names (CN) for tracking and achieved superior performance compared to the trackers exploiting purely grayscale features. The combined use of HOG and CN features by Li *et al.* demonstrated further enhancements in performance [15]. However, the generic nature of hand-crafted features like HOG and CN compromises their ability to tune the extraction of discriminatory information to the target appearance. Moreover, they are not invariant with rotations. As a result, filters learnt using only hand-crafted features always suffer from the lack of discriminative information, especially when the appearance of a target undergoes a rapid variation.

Recently, motivated by the impressive results achieved in the field of classification and detection, deep convolutional networks have been investigated in the context of visual object tracking and found to be very effective [16], [17]. Deep representations are more robust and discriminative, especially in challenging tracking scenarios. However, as deep descriptors are of low resolution, they lose some detail appearance information, which makes the target hard to locate precisely and to estimate its accurate scale. Moreover, not all deep features are necessarily informative, and the presence of irrelevant dimensions degrades the discriminatory power of DCF, not mentioning that extracting deep features in each frame and training filters over deep representations is expensive computationally.

To address these complex issues and their negative impact on the DCF tracking performance, we propose a novel generative learning method to extract a dynamic target representation by adaptively exploiting complementary sources of information conveyed by multi-cue features. We propose to use a collaborative representation [18] between two consecutive frames for that purpose. The collaborative representation coefficients are then applied as features (CR features) to learn the target detection and localisation model. As a result, the learnt representation adapts effectively to any target variations, and ignores the impact of the background clutter at the same time. In consequence, the proposed generative feature extractor via collaborative representation overcomes the limitations of the conventional, fixed feature representation, especially when the target appearance changes dramatically.

While DCF's agility is essential to be able to respond to rapid changes in target visual appearance, it is equally important for it to exhibit temporal consistency. This requirement reflects that the prevailing motion of the target is smooth and continuous. As the visual target representation often contains redundant information, without imposing temporal consistency constraint, the filter coefficients may arbitrarily switch between multiple alternative solutions. Such instability is undesirable. More importantly, ignoring temporal consistency impacts directly on the speed of filter convergence.

Note that the limited spatio-temporal continuity stems from the DCF based trackers adopting fixed patterns of spatial and temporal modelling. Traditional DCF-based trackers extract numerous feature representations in each single frame to train the model and update it at a fixed rate in every frame [13], [19], [20], [21], [22], [23]. As the core method lacks the ability to represent the target dynamics, by the same token it is unable to promote temporal consistency. In order to capture the spatial context, the so called attention mechanism has been widely investigated and a significant success has been achieved in many tasks of computer vision [24], [25], [26], [27], [28], [29], [30]. In DCF-based tracking, attention modelling is usually realised by employing spatial regularisation. Danelljan *et al.* first proposed a spatially regularised model for tracking with a fixed negative Gaussian-shaped spatial weight vector to penalise the filter coefficients which significantly alleviates boundary effect [31]. Subsequently, Kiani *et.al* [22] proposed to use a pre-defined binary matrix to generate real negative and positive samples by a method of cropping, which effectively suppresses the inclusion of spatial background and boundary region as training samples. However, these models usually try to address boundary effect by using a fixed penalty vector to regularise spatially, which lacks adaptability to target changes between successive frames.

To address these problems, we propose a structural spatio-temporal modelling formulation for tracking. The generated collaborative representation features (CR features) are the transformation coefficients of successive frames, which are able to represent dynamic appearance variations of the target during tracking and support temporal continuity. The notable characteristics of the collaborative representation coefficients is their ability to track the appearance changes of the target, while ignoring the background, which adaptively suppresses the spatial boundary effect, and can be regarded as complementary spatial regularisation. Combining the original feature representations extracted in the current frame with the CR features generated between successive frames, the model is endowed with the capability to adapt to appearance changes through the proposed learning and updating online scheme.

In this paper, we propose Collaborative Representation based Complementary Discriminative Correlation Filters for visual tracking (CRCDCF). The proposed use of generative collaborative representation learning approach enhances both, the discriminatory, as well as the target transformation information captured by traditional feature extraction methods, which is important for coping with dramatic object appearance changes. Inspired by the superior performance of deep features extracted from CNNs pretrained on the large scale ImageNet dataset [32] and the effective spatial regularisation of BACF [22], we adopt the paradigm of BACF as a baseline, but equip it with hand-crafted features and deep features from pretrained ResNet [33] and the corresponding CR features. The framework of BACF effectively alleviates the boundary effect of DCFs, which improves the performance while maintaining favourable speed. Embedding discriminative deep features and complementary CR features in the framework of BACF, an adaptive structural spatio-temporal model is designed to improve the tracking accuracy and robustness.

The main contributions of the proposed CRCDCF method are summarised as follows:

- We propose to use an adaptive generative learning method for extracting a complementary feature representation. It is conveyed by features extracted by collaborative representation between successive frames (CR features), which yield learned filters that are more discriminative and robust in the presence of abrupt appearance variations.

- We propose a structural spatio-temporal model for tracking. The generated CR features are able to represent temporal appearance change of targets and impose adaptive spatial regularisation. The CR feature channels and the original feature maps are used jointly within the spatially regularised BACF framework to achieve complementary discrimination and adaptive spatio-temporal continuity.
- We extensively evaluate our method on a number of well-known benchmarking datasets such as OTB2013 [2], Temple-Colour128 [5], UAV123 [34], VOT2018 [35] and LaSOT [36] datasets. The experimental results demonstrate the effectiveness and robustness of our method compared with the state-of-the-art trackers.

II. Related Work

For comprehensive review of existing tracking methods the reader can refer to recent surveys [1], [2], [3], [4]. In this section, we briefly discuss the techniques that are most relevant to our methods.

Generative methods typically consider the most similar region as the target in each frame, based on specific similarity functions. Early work included Kalman filtering [37], mean-shift [38] and incremental learning [39] methods. Later, Mei *et al.* introduced sparse representation to visual tracking [40]. In the sparse tracking paradigm, target candidates are represented by a sparse linear combination of target templates (dictionary atoms). The representation coefficients are then used to measure similarities by calculating the reconstruction error of each target candidate. Both sparse and low-rank representation based trackers have achieved appreciable success [41], [42], [43], [44], [45]. Recently, Zhang *et al.* propose a novel circulant sparse tracker (CST) [46], which exploits circulant target templates via embedding high dimensional HOG features.

On the other hand, discriminative methods cast a tracking task as a classification problem that aims at distinguishing the target from background. Since the proposal of the Minimum Output Sum of Squared Error (MOSSE) filter [12], DCF based trackers have been received extensive attention. Specifically, Henriques *et al.* exploited the circulant structure in a kernel space with HOG features [11], [13]. More advanced DCF based trackers concentrate on feature representation, scale detection and spatial regularisation. As for feature representation, HOG, CN and CNN features have been demonstrated to provide superior performance for visual tracking [15], [17], [47], [48], [49]. For scale detection, DSST [50] and fDSST [51] propose a successful framework to handle the problem of target scale variation. As to spatial regularisation, SRDCF [31], CSRDCF [52], BACF [22] and LADCF [8] incorporate effective strategies to solve the issue of spatial boundary effect.

Recently, Peng *et al.* proposed a novel densely connected DCFs framework named DCDCF for visual tracking [53], which exploits the densely connected structure of multiple DCFs. In DCDCF, the translation response maps from correlation filters of previous layers are used as feature inputs to correlation filters of the next layer. All the feature maps and interim response maps from various filters are reused in DCDCF, which is able to effectively capture the appearance variations of target and can significantly enhance discrimination of the learned model.

In our work, we employ collaborative representation to extract the coefficients of the target between successive frames as complementary feature cues. The CR features are essentially similar to the translation response maps of DCFs. After obtaining CR features through generative learning, we adopt these coefficients together with the original features for discriminative filters learning, and show it enhances the discriminative power of the trained model.

III. DCF and Background-Aware CF

In this section, we first briefly revisit the classical Discriminative Correlation Filter, which has received wide attention because of its outstanding performance in recent benchmarks [7], [4]. In the *t*-th frame, the multi-channel discriminative correlation filters *h* are designed by optimising the following objective function:

$$\min_{h} \frac{1}{2}\sum_{j=1}^{mn} \|y(j) - \sum_{d=1}^{D} h_d^T f_d[\Delta\tau_j]\|_2^2 + \frac{\lambda}{2}\sum_{d=1}^{D}\|h_d\|_2^2, \quad (1)$$

where $f_d \in \mathbb{R}^{mn\times 1}$ is the *d*-th channel feature map of an $m \times n$ image patch centred on the position of the target, and $[\Delta\tau_j]$ denotes the circular shift operator. $y \in \mathbb{R}^{mn\times 1}$ is the corresponding Gaussian shaped label with $y(j)$ denoting the *j*-th element of y. D is the number of feature channels and λ is the weighting parameter of the regularisation term. $h_d \in \mathbb{R}^{mn\times 1}$ stands for the filter of the *d*-th channel. The optimisation problem in Eq. (1) can efficiently be solved in the Fourier domain.

Next, we briefly overview the BACF [22] tracker, which has been shown to be effective in addressing the issue of boundary effect by achieving remarkable results on several tracking benchmarks. Multi-channel background-aware correlation filters are learned by the BACF tracker through minimising the following objective function:

$$\min_{h} \frac{1}{2}\sum_{j=1}^{mn} \|y(j) - \sum_{d=1}^{D} h_d^T \mathbf{P} f_d[\Delta\tau_j]\|_2^2 + \frac{\lambda}{2}\sum_{d=1}^{D}\|h_d\|_2^2, \quad (2)$$

where $\mathbf{P}$ is a binary matrix which crops the central patch of each shifted image. $\mathbf{P}f_d[\Delta\tau_j]$ returns all possible patches cropped from the entire frame.

Instead of being generated by cyclic shifting of the positive sample, negative training examples are densely extracted from the background via the cropping operator, $\mathbf{P}$, in BACF. Learning from such real patches significantly improves the discrimination of the filters so that the robustness and accuracy of the BACF tracker are correspondingly boosted. For more details the reader can refer to [22].

In view of the remarkable effectiveness of its spatial regularisation and real-time speed, we adopt BACF as our framework and equip it with a complementary collaborative representation learning mechanism to advance the state-of-the-art further.

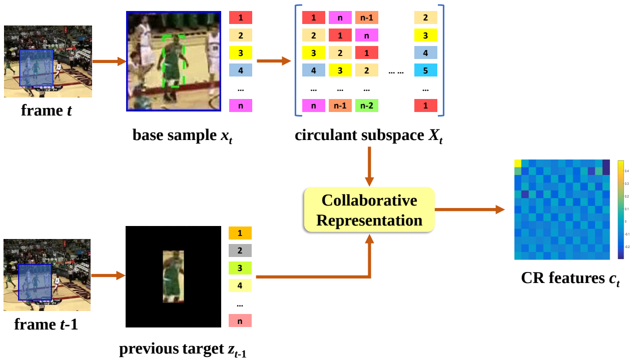


Fig. 1. Illustration of the extraction process of collaborative representation features. The target z_{t-1} in the previous frame and the padded image patch (search window) x_t in the current frame share the same scale. z_{t-1} is filtered by applying a pre-defined mask to the padded image patch centred at the position of the target in the last frame. Each channel the original feature set enables the system to generate the corresponding collaborative representation feature channel through Eq. (3), Eq. (4) and Eq. (5).

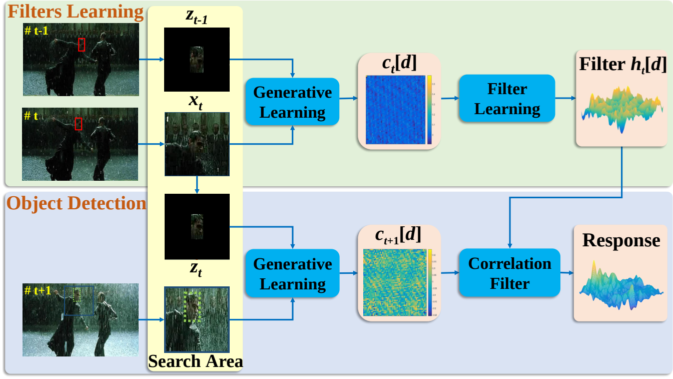


Fig. 2. Illustration of collaborative representation features in DCF-based tracking. The model obtained by joint discriminative filter learning and generative collaborative representation learning exhibits enhanced discriminative power, as it has the capacity to reflect appearance variations and alleviate the negative impact of the target background.

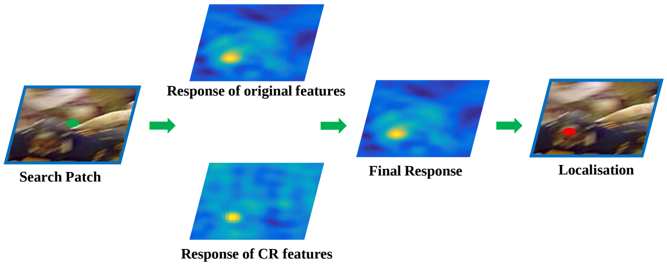


Fig. 3. Visualisation of the response maps in the target location step. The final object location response map is obtained by fusing the response map of the original features and the response map of the collaborative representation features.

IV. Our Approach

In this section, we present the details of the proposed complementary discriminative correlation filters based on collaborative representation for visual tracking. First, we extract transform coefficients via a generative collaborative representation method. Second, we train a model in terms of background-aware filters using hand-crafted, deep and complementary CR features. The details of the framework and our proposed tracker are furnished in Section IV-C.

A. Generative collaborative representation learning

Given a $m \times n$ padded image patch in the t-th frame, containing the target in its centre, we compute multi-channel features $x_t \in \mathbb{R}^{mn \times D}$. Then cyclic shift samples of x_t are used to construct a subspace $X_t \in \mathbb{R}^{mn \times Dmn}$, which is circulant. The subspace X_t is employed to obtain a collaborative representation [18] of the target $z_{t-1} \in \mathbb{R}^{mn \times D}$ using high dimensional features of the previous frame:

$$\min_{c_t[d]} \|X_t[d]c_t[d] - z_{t-1}[d]\|_2^2 + \lambda_1 \|c_t[d]\|_2^2, \tag{3}$$

where d is channel index, λ_1 is the weight of regularisation term and c_t is an array of representation coefficients, coined as CR features. The target z_{t-1} is extracted from the previous frame by applying a pre-defined mask with only the target region activated. Formally, this objective function defines a standard ridge regression problem that has a closed-form solution:

$$c_t[d] = (X_t^T[d]X_t[d] + \lambda_1 I)^{-1} X_t^T[d] z_{t-1}[d], \tag{4}$$

where I is an identity matrix. Due to the circulant property of the subspace X_t, Eq. (4) can be solved by converting it into equivalent form in the Fourier domain as follows:

$$\hat{c}_t[d] = \frac{\hat{x}_t^*[d] \odot \hat{z}_{t-1}[d]}{\hat{x}_t^*[d] \odot \hat{x}_t[d] + \lambda_1}, \tag{5}$$

where $\hat{x}_t^*$ is the complex conjugate of $\hat{x}_t$ in the frequency domain, $\hat{x}_t$ is the Fourier representation of x, and $\odot$ denotes the element-wise multiplication operator. The solution of c_t needs Fast Fourier Transform (FFT) and inverse FFT, which can be solved in $\mathcal{O}(Dmnlog(mn))$. The computational cost of the element-wise multiplication is $\mathcal{O}(1)$. Therefore, the total complexity of Eq. (3) is $\mathcal{O}(Dmnlog(mn))$. The extraction process of the CR features is visualised in Fig. 1.

After extracting the CR features, they are concatenated with the original features across all channels:

$$f_t = [x_t[1]; x_t[2]; ...; x_t[D]; c_t[1]; c_t[2]; ...; c_t[D]], \tag{6}$$

where f_t denotes all the features collected for filter learning and target localisation.

In Fig. 2, we attempt to intuitively illustrate the roles of the CR features in the respective stages of filter learning and object detection. In the learning stage, the CR features are the representation coefficients of the target in frame $t-1$ using the original features of frame t as a dictionary. This representation has the propensity to reflect appearance variations of the target. In the subsequent stage of target localisation, the CR features essentially serve as the target transformation coefficients, which roughly estimate the change in the target position in frame t. Thanks to the generative representation learning, discriminative filter learning is able to moderate the impact of the background and locate the target more accurately.

The significance of the proposed CR features is conveyed visually in Fig. 3, where we give an example of the filter response to the original features and to the CR features. In tracking, given a new frame, a search image patch is cropped by centring the target mask at the target position in the previous frame. The extracted image patch in frame $t-1$ is collaboratively represented by shifted samples of the search window in frame t to extract CR features and the corresponding DCF. In frame t the response maps are calculated by the filters updated in the previous frame and the features extracted in the current frame. Note that the response map of the original features is accurate enough for target localisation when the tracking scenes undergo small variations. However, it may be limited in scenarios involving abrupt target appearance variations and background clutter. With the additional response of the CR features, the final target response becomes more robust.

B. Discriminative Background-aware Filter Learning

Although the conventional DCF based trackers perform efficient appearance model learning, they suffer from the spatial boundary effect. The recent efforts addressing this issue has witnessed a considerable progress in suppressing the boundary information of the filters. Examples include the BACF tracker, briefly described in Section III. In Eq. (2), BACF uses a binary matrix to crop real positive and negative samples from the image, forcing the filters to learn as if the background was set to zero. Optimising the objective function in Eq. (2), BACF learns discriminative background-aware filters with HOG features. However, the model with background-aware filters learnt only from HOG features lacks of the ability to discriminate robustness, which results in tracking failures when some complicated cases occur.

In our work, we propose to learn a more robust and discriminative appearance model with background-aware filters using HOG, CN, deep features from pretrained CNN and corresponding CR features through the paradigm of BACF. HOG features emphasise the gradient information of the image, while CN features focus on the colour characteristics. As for the deep features, we employ *Res4e* of pretrained ResNet50 as layers for feature extraction. The deep features pay more attention to semantic information and have the ability to extract powerful appearance representation. The proposed CR features are sensitive to appearance variations and in this sense provide complementary information for visual object representation. Thus, the model with background-aware correlation filters trained with these features is much more discriminative and robust, with improved tracking performance.

C. Tracking Framework

The framework for the proposed DCF tracking based on collaborative representation is summarised in Algorithm 1.

Detection: Following fDSST [51] and BACF [22], the position and scale of the target are detected simultaneously. In our tracker, multiple scales a^s of the search region $\{f_s\}_{s\in\{\lfloor\frac{1-S}{2}\rfloor,\dots,\lfloor\frac{S-1}{2}\rfloor\}}$ are analysed to extract HOG, CN, deep and corresponding CR features. The S denotes the number of scales. Then a filter, h_s, is designed for each f_s in the Fourier domain to obtain multi-channel response maps as:

$$\hat{R}_s = \hat{h}_s \odot \hat{f}_s, \tag{7}$$

Then the interpolation strategy in SRDCF [31] is applied to the response map corresponding to each f_s. The position and scale of target p_t and s_t is predicted according to the maximum of the final response map.

In addition, in view of the fact that a direct summing of all the response maps of different features may cause some interference, as not all the features in every frame are discriminative enough, we fuse these response maps by an adaptive weighting strategy. Specifically, for the channels from the same feature category, we sum the corresponding response maps directly. Then we compute a weighted average of the response maps produced by different feature maps. The weights are defined by the peak-to-side ratio (PSR) of each response map. The PSR is computed as:

$$PSR_i = \frac{R_i^{max} - \mu_i}{\sigma_i}, \tag{8}$$

where R_i^{max} is the maximum value of the response map corresponding to the *i*-th features, μ_i and σ_i are the mean and the standard deviation of the *i*-th response map. A large value of PSR indicates that the response map is more reliable. The final response map can be calculated as:

$$R = \sum_{i=1}^{K} PSR_i \cdot R_i. \tag{9}$$

The position of the target coincides with the point of maximum value of the final response map.

Learning: In the learning stage, we adopt the HOG, CN and deep features of the padded image patch centred at the position of the target, p_t. The CR features are generated by using Eq. (3) and Eq. (4) in Section IV-A. Then, the original features and the CR features of the image patch are used in Eq. (2) in Section III to obtain the multi-channel background-aware filters h. Specifically, in the first frame, we use the padded target patch cut-out by a pre-defined mask uncovering the target region of the previous frame, z_{t-1}, to generate CR features by using Eq. (5).

Updating: We utilise the same online model updating strategy as other DCF based trackers:

$$h_t = (1-\eta)h_{t-1} + \eta h, \tag{10}$$

where η is the online updating rate, and h denotes the filters learned with image patch in *t*-th frame using Eq. (2). Based on this updating strategy, we use filters h_t to locate the target position in (*t*+1)-th frame.

V. Experiments

We perform extensive comparative experiments on several well-known databases and achieve the state-of-the-art performance. In this section, we describe the implementation details of our tracker, including the experimental platform and the parameters setup. Next, the benchmarking datasets and the

Algorithm 1 CRCDCF algorithm

```
Input: Image frame $I_t$; Filters $h_{t-1}$; Target position $p_{t-1}$ and
    target scale $s_{t-1}$ from frame $t-1$; Target region $z_{t-1}$ in
    the $t-1$ frame;
Extract search image patch with $S$ scales from $I_t$ at $p_{t-1}$;
Extract corresponding features representations $[x_s]_{s=1}^{S}$ and
    generate CR features $[c_s]_{s=1}^{S}$ through Eq. (3);
Concatenate original features and CR features to $[f_s]_{s=1}^{S}$;
Calculate response maps $[R_s]_{s=1}^{S}$ using Eq. (7) and
    Eq. (9);
Locate target $p_t$ and estimate scale $s_t$ from the maximal
    value of response maps;
Obtain multi-cue features $x_t$ based on current $p_t$ and $s_t$;
Generate corresponding CR features and concatenate orig-
    inal features with CR features: $f_t$.
Training filters $h$ with $f_t$ using Eq. (2);
Preserve target region $z_t$ in current frame;
Update filters $h_t$ using Eq. (10).
Output: Target position $p_t$ and scale $s_t$; Updated filters $h_t$
    for frame $t$; Target region $z_t$ in frame $t$.
```

criteria for evaluation are introduced, as well as the state-of-the-art trackers we compared with. Next, the sensitivity of the parameters to our method is analysed, and the effectiveness of each component of the proposed CRCDCF is demonstrated. Finally, we analyse the experimental results on various datasets and discuss the merits of our CRCDCF method.

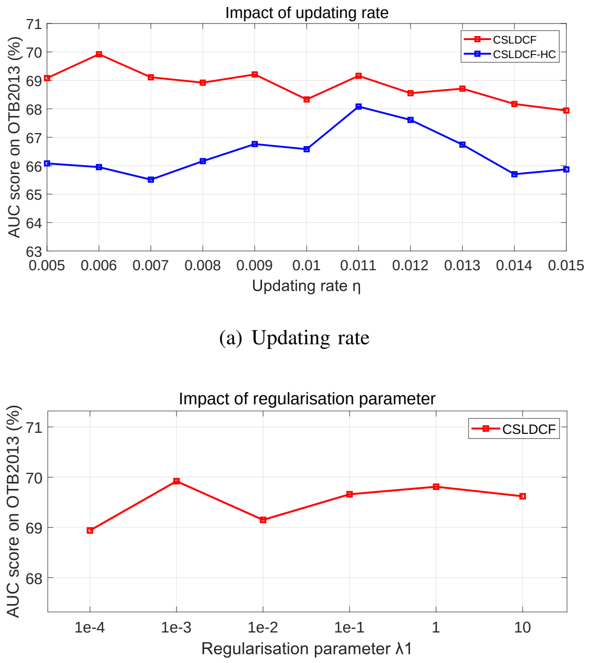
A. Implementation Details

The proposed CRCDCF algorithm was implemented in MATLAB 2018a on an Intel i9 3.00GHz CPU. We adopt both hand-crafted and deep features in our method. As for visual tracking, it has been demonstrated that robust and discriminative feature representation plays a dominant role. We equip the proposed CRCDCF-HC method with hand-crafted features such as HOG and CN, and embed both hand-crafted and deep features in CRCDCF. The deep features we adopted are extracted from the *Res4e* layers of pretrained ResNet-50. The tracking speed of CRCDCF-HC is about 21*fps* and speed of CRCDCF is 5*fps* approximately.

In the proposed CRCDCF, the padding parameter of the search region is set to 4.5, the regularisation parameters λ in Eq. (2) [22] and λ_1 in Eq. (3) take values $\lambda = 0.01$ and $\lambda_1 = 0.001$. The updating rate η is chosen to be 0.013 for hand-crafted and CR features while it is set to 0.006 for deep features. We generate $S = 5$ scales S of the feature data, with the scale factor a of $a = 1.01$ [51]. The parameter settings are fixed for all experiments.

B. Experimental Setup

Datasets: We extensively evaluate our algorithm on five benchmarks: OTB2013 [2], Temple-Color128 [5], UAV123 [34], VOT2018 [35] and LaSOT [36] datasets to compare with the state-of-the-art trackers. OTB2013 dataset is one of most popular datasets in visual object tracking containing 50 video sequences with 11 challenging attributes. Temple-Colour 128 (TC128) consists of 128 colour video sequences and UAV123 is composed of 123 challenging sequences. The VOT2018 dataset is provided by the VOT2018 challenge, which contains 60 video sequences with 5 challenging attributes including size change, occlusion, motion change, illumination change, camera motion. The LaSOT dataset is a high-quality benchmark for Large-scale Single Object Tracking, which consists of 1400 sequences in 70 categories with more than 2500 frames in each sequence. In this paper, we make a comparison using a simplified version of LaSOT that contains 280 video sequences.



(a) Updating rate

(b) Regularisation parameter

Fig. 4. The experimental results obtained by our CRCDCF on OTB2013 for (a) different updating rates and (b) different regularisation parameters.

TABLE I
PERFORMANCE EVALUATION ON OTB2013 FOR ABLATION ANALYSIS OF THE PROPOSED CRCDCF.

Configurations	DP Score	AUC Score
HOG	84.9%	64.5%
HOG + CR	85.6%	65.8%
CN	72.4%	51.1%
CN + CR	74.1%	52.3%
HOG + CN	86.7%	66.5%
HOG + CN + CR	89.1%	68.1%
HOG + CN + CNN	91.7%	69.0%
HOG + CN + CNN + CR	92.7%	69.9%

Evaluation metrics: As for OTB2013, TC128, UAV123 and LaSOT, the One Pass Evaluation (OPE)[2] is employed to evaluate the performance of various trackers. The precision and success plots respectively, based on centre location error and bounding box overlap ratio, are used for the comparison of various trackers. In addition, three criteria, namely Distance Precision (DP), Overlap Precision (OP), Area Under

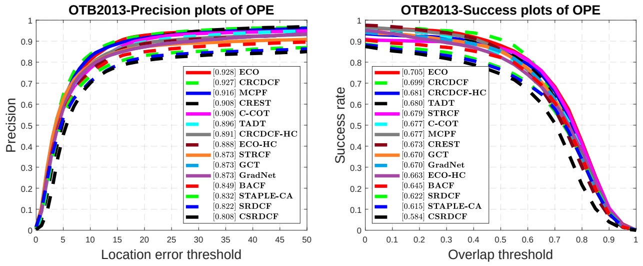


Fig. 5. The experimental results of our CRCDCF compared with 13 trackers on OTB2013 dataset. This figure shows the precision and success plots in terms of the OPE protocol. The DP and AUC score of each tracker is shown in the legend.

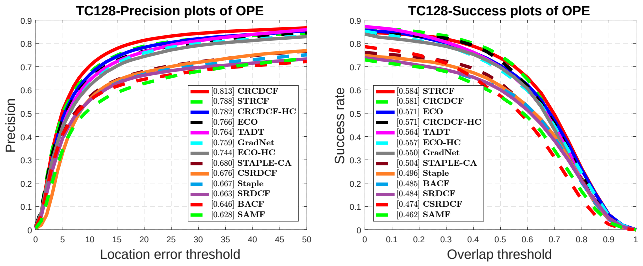


Fig. 6. The experimental results of our CRCDCF compared with other 11 trackers on TC128 dataset. This figure shows the precision and success plots in terms of the OPE protocol. The DP and AUC score of each tracker is shown in the legend.

Curve (AUC) are employed to quantitatively compare the performance of the various algorithms. The DP is defined as the percentage of location errors within a threshold of 20 pixels. The OP is the percentage of overlap ratios surpassing a threshold of 0.5. The AUC is the area under the curve of success plot. As for the VOT2018 dataset, we use the Expected Average Overlap (EAO), Average Overlap (AO) and Failures as criteria. All the three evaluation metrics for VOT2018 are in the reset-based protocols. The EAO is the main evaluation standard in VOT challenges which accounts for both accuracy and robustness. The AO in VOT2018 indicates the mean accuracy of a tracker. The Failures stands for the robustness of a tracker, and a lower value means better robustness.

State-of-the-art competitors: We compare our CRCDCF with 21 state-of-the-art trackers including SAMF [15], SRDCF [31], Staple [20], BACF [22], STAPLE-CA [21], CSRDCF [52] , ECO-HC [9], STRCF [54], CREST [55], CFWCR [56], SiamFC [57], CFNet [58], MCPF [59], C-COT [60] and ECO [9], SRDCF-deep [17], TRACA [61], MCCT [48], GCT [62], TADT [63], and GradNet [64]. For a fair comparison, all the results of these trackers were obtained by re-running these algorithms on the datasets using the source codes published by the original authors with the provided parameter settings.

C. Self Analysis

In this part, we first analyse the sensitivity of the algorithm to parameters. Then an ablation analysis is conducted to demonstrate the effectiveness of each component of our proposed CRCDCF method.

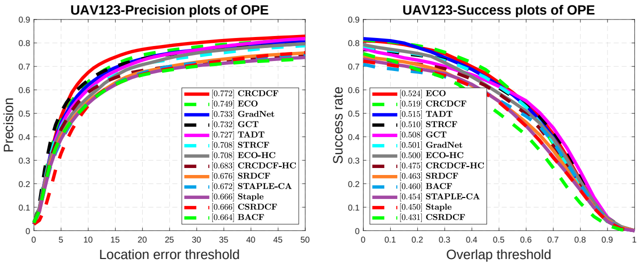


Fig. 7. The experimental results of our CRCDCF compared with other 11 trackers on UAV123 dataset. This figure shows the precision and success plots in terms of the OPE protocol. The DP and AUC score of each tracker is shown in the legend.

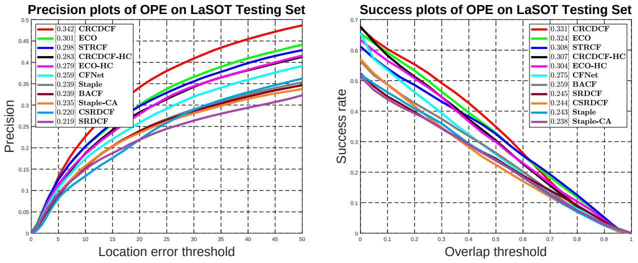


Fig. 8. The experimental results of our CRCDCF compared with other 10 trackers on LaSOT dataset. This figure shows the precision and success plots in terms of the OPE protocol. The DP and AUC score of each tracker is shown in the legend.

In Fig. 4, we present the impact of the updating rate, η, and the regularisation parameter, λ_1, on the performance of our method. As can be seen from the figure, the tracking performance varies smoothly with the changes of the updating rate, η, which confirms that CRCDCF is stable, thanks to the effect of the proposed CR features. In Fig. 4(b), we analyse the impact of the regularisation parameter λ_1 in Eq. (3). It is evident that the choice of parameter λ_1 is not critical.

Additionally, we carry out an ablation analysis on OTB2013 to demonstrate the effectiveness of each component of our CRCDCF. In Table. I, we present the results in terms of the DP and AUC scores of CRCDCF using different feature configurations including HOG, CN, deep CNN features and the proposed CR features. As shown in the table, our method improves, in terms of the DP and AUC scores, by 7.8% and 5.4% respectively on the OTB2013 dataset. It can be seen obviously that, with the proposed CR features, the tracking performance is improved considerably.

D. Results and Analysis

Overall Performance: In Fig. 5, we report the precision and success rate plots of the experimental results on OTB2013 with DP and AUC scores in the figure legend. Compared with other trackers, our CRCDCF-HC achieves 89.1% in DP and 68.1% in AUC. These results are much better than those achieved by all the tracking methods using only hand-crafted features. They are even better than some trackers using deep features/structures. In terms of DP and AUC, our CRCDCF achieves the second best performance of 92.7% and 69.9% , falling behind ECO by only 0.1% in DP and 0.6% in AUC.

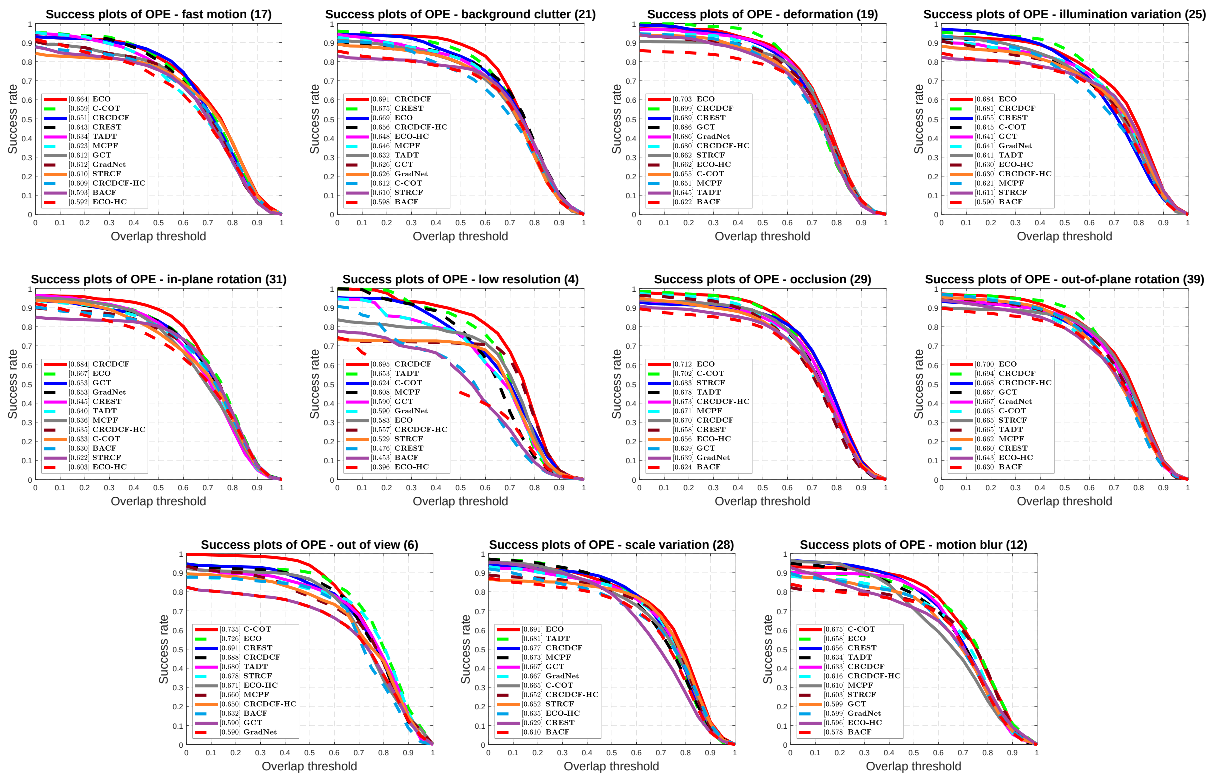


Fig. 9. Success plots obtained on the OTB2013 dataset in terms of 11 challenging factors, including fast motion,background clutter, deformation, illumination variation, in-plane rotation, low resolution, occlusion, out-of-plane rotation, scale variation, out of view and motion blur. For clarity, only the results of the top 12 trackers in each attributes are provided.

TABLE II
THE OP RESULTS OF 10 TRACKERS ON THE OTB2013, TC128, UAV123 AND LASOT DATASETS. THE TOP THREE RESULTS ARE RESPECTIVELY SHOWN IN RED, BLUE AND GREEN.

		Staple	CSRDCF	STAPLE-CA	SRDCF	BACF	ECO-HC	STRCF	ECO	CRCDCF-HC	CRCDCF
Average OP(%)	OTB2013	73.8	76.1	76.5	79.0	82.2	83.4	86.6	87.9	85.2	90.8
	TC128	61.7	57.4	63.0	60.1	60.8	68.5	73.7	70.4	72.6	74.8
	UAV123	54.7	55.1	55.1	55.1	55.5	58.5	60.9	63.1	58.2	64.2
	LaSOT	24.0	22.4	23.8	24.5	26.3	29.9	32.5	32.9	30.4	34.9

TABLE III
EVALUATION ON THE VOT2018 DATASET IN TERMS OF EAO, AO, AND FAILURES OF 11 TRACKERS. THE TOP THREE TRACKERS ARE SHOWN IN COLOUR.

	Staple	GradNet	CSRDCF	SRDCF-Deep	SiamFC	C-COT	TRACA	MCCT	ECO	GCT	CRCDCF
EAO	0.169	0.247	0.256	0.154	0.187	0.267	0.139	0.273	0.281	0.275	0.312
AO	0.523	0.507	0.485	0.485	0.498	0.485	0.431	0.526	0.476	0.485	0.521
Failures	44.02	24.11	23.57	46.00	34.03	20.41	53.68	19.22	17.66	21.42	15.46

The reason why our method dose not exceed ECO in AUC is that ECO employs many samples from historical frames to train the model for every frame, which improves its accuracy and robustness.

The experimental results given in terms of the precision and success rate plots, obtained on the TC128 dataset, are presented in Fig. 6. On TC128, our CRCDCF achieves the best score of 81.3% in terms of DP and performs better than the second best tracker STRCF by 2.5%. With the proposed CR features, our CRCDCF is able to locate the target more precisely, leading to better tracking accuracy. In terms of AUC, our CRCDCF achieves 58.1%, ranking the second best, and it is lower than the best tracker STRCF by only 0.3%. The objective function of STRCF is equipped with an additional temporal regularisation term, besides a spatial regularisation term, which results in preventing model degeneration and in achieving better robustness. Using hand-crafted features only, our CRCDCF-HC achieves 78.2% in DP, which exceeds ECO with deep features. The AUC score of CRCDCF-HC of 57.1% is the same as that achieved by ECO.

As for dataset of UAV123, Fig. 7 displays the precision and success rate plot over all the 123 video sequences. Compared to other trackers, our CRCDCF achieves the bset score in DP with 77.2% and performs better than the second and third best trackers ECO and GradNet by 2.3% and 3.9% respectively, mainly owing to the proposed use of complementary genera-

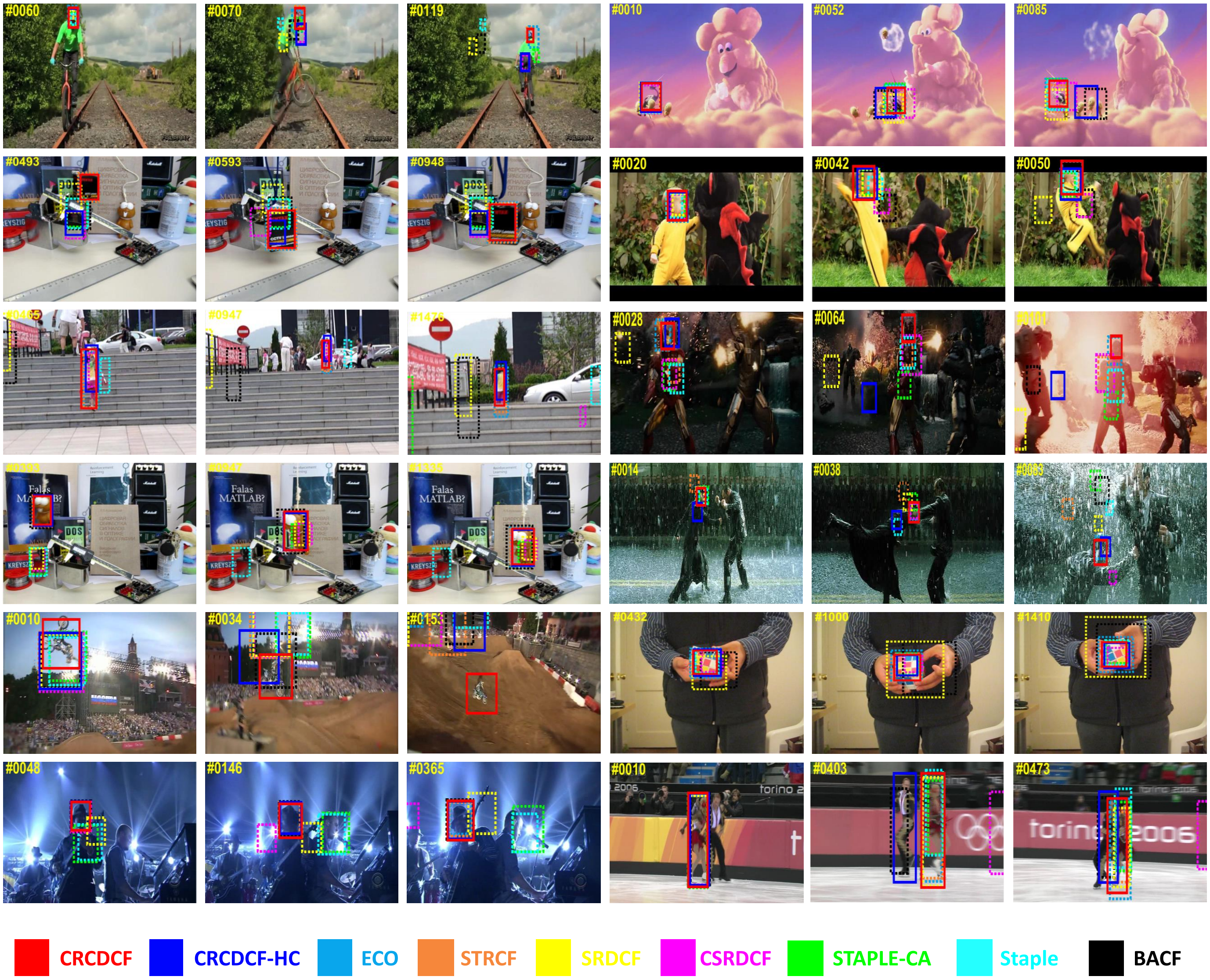


Fig. 10. Qualitative performance comparison on challenging video sequences including *Biker, Bird2, Box, DragonBaby, Girl2, Ironman, Lemming, Matrix, MotorRolling, Rubik, Shaking, Skating2-1*. The colour bounding boxes are the corresponding results of CRCDCF, CRCDCF-HC, ECO, STRCF, SRDCF, CSRDCF, STAPLE-CA, Staple and BACF respectively.

tive learning which benefits the accuracy of target localisation. In terms of AUC, our CRCDCF achieves score of 51.9%, falling only 0.5% behind the best tracker ECO. Our CRCDCF-HC achieves scores of 68.3% and 47.5% respectively in DP and AUC.

In Fig. 8, we provide the results in terms of precision and success plots on LaSOT with DP and AUC. From the plots, our CRCDCF is the best, achieving 34.2% in DP and 33.1% in AUC, which beats ECO by 4.1% and 0.7% respectively. The main reason why our CRCDCF outperforms other trackers is that the sequences of LaSOT are generally longer, and our tracker is enabled to perform better on these long video sequences by using more robust deep CNN features and the complementary spatio-temporal information from the proposed CR features. Additionally, our CRCDCF-HC also achieves superior performance with 28.3% and 30.7% in DP and AUC compared with the other trackers without using deep features.

Finally, we evaluate our proposed CRCDCF on OTB2013, TC128, UAV123 and LaSOT in terms of the OP metric. Table II shows the results, compared to several competitive trackers. On all these four datasets, our CRCDCF achieves the best score with gains of 2.9%, 1.1%, 1.1%, 2.0% over the second best tracker respectively. Additionally, we report the evaluation results on VOT2018 in Table III. Our CRCDCF achieves 0.312 in terms of EAO which gains about 11% over ECO.

Attribute-Based Evaluation: We present the results of the attribute-based evaluation of 10 state-of-the-art trackers on OTB2013 in Fig. 9. We show the success plots with AUC scores for 11 video attributes including fast motion, background clutter, deformation, illumination variation, in-plane rotation, low resolution, occlusion, out-of-plane rotation, scale variation, out of view and motion blur. The results show that the proposed CRCDCF outperforms other trackers in 3 attributes, *i.e.*, background clutter, in-plane rotation and low resolution. In our method, the filters learning and updating with the proposed collaborative representation feature model is equipped to capture the appearance variations. This property helps to alleviate the impact of the background. Besides, the adopted deep features from ResNet-50 are more robust, especially when the target exhibits rotation and abrupt variation. In addition, compared to other trackers using only hand-crafted features, our CRCDCF-HC achieves the best performance in 9 attributes, which demonstrates the effectiveness and robustness of the CR features we proposed.

Qualitative Results: In Fig. 10, we show the qualitative tracking results of the state-of-the-art trackers, *i.e.*, SRDCF, Staple, CSRDCF, Staple_CA, BACF, STRCF, ECO and our CRCDCF-HC and CRCDCF on some challenging sequences. These videos all contain serious target appearance variation. Still, our CRCDCF performs well on these videos as the model with filters trained using CR features is agile enough to respond to rapid appearance variation during tracking. The generative collaborative representation between successive two frames is instrumental in extracting more discriminative information for tracking.

Tracking Failures: In Fig. 11, we provide some samples of failed cases, and take these cases to analyse the demerits of the proposed method. As can be seen from the figure, our method is unable to handle situations where the target suffers from long-term absence, which leads to tracking failures in videos such as *Bird1* and *Soccer*. Thus, mechanisms to prevent model degradation will be required by our method. Besides, if the scale of the target changes rapidly, especially when the aspect ratio of the target encounters significant variation, our method will not be able to precisely estimate the size of the target. As can be seen in video sequences *Diving* and *Jump*, the aspect ratio of target changes abruptly. The bounding boxes predicted by our method are inaccurate. Moreover, the interference from the same object as the target may cause tracking failures, especially when the deep CNN features are used by the tracker, as shown in video sequence *Coupon*. This is mainly because deep CNN features focus on semantic information and ignore the texture information, which results in a high response at the position of a similar object.

VI. Conclusions

In this paper, a collaborative representation mechanism for visual tracking using discriminative correlation filters has been proposed. The key ingredient is a novel method for generating a collaborative representation feature cue as a complementary target representation for use in conjunction with the DCF framework. The collaborative representation feature provides valuable additional discriminative information and robustness to create an effective DCF model, especially when the target undergoes dramatic appearance variations. The experimental results on OTB2013, TC128, UAV123, LaSOT and VOT2018 demonstrate the competitive performance of our CRCDCF tracker.

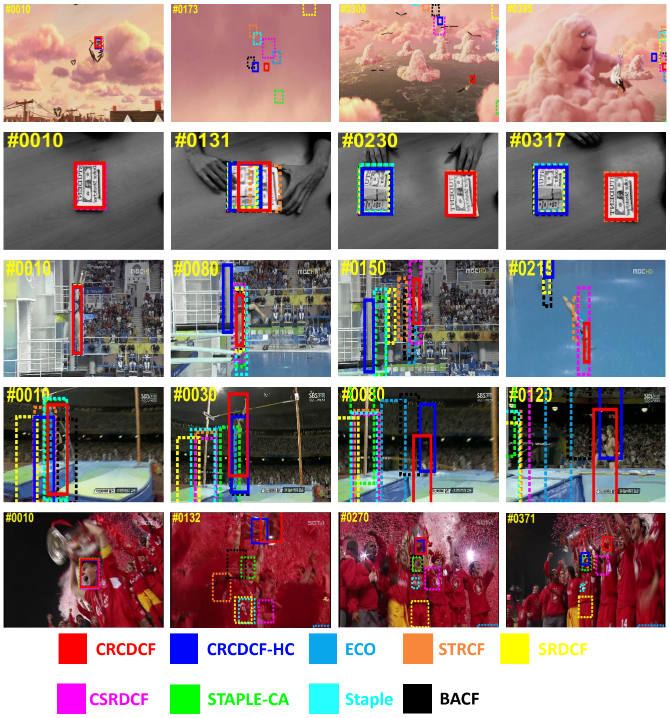


Fig. 11. Some failed cases on challenging video sequences including *Bird1, Coupon, Diving, Jump, Soccer*. The colour bounding boxes are the corresponding results of CRCDCF, CRCDCF-HC, ECO, STRCF, SRDCF, CSRDCF, STAPLE-CA, Staple and BACF respectively.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (61672265, U1836218, 61902153), the 111 Project of Ministry of Education of China (B12018), the EPSRC Programme Grant (FACER2VM) EP/N007743/1 and the EPSRC/MURI/Dstl project under Grant EP/R013616/1.

References

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm computing surveys (CSUR)*, vol. 38, no. 4, p. 13, 2006.

[2] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2411–2418.

[3] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1442–1468, 2014.

[4] 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.

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

[6] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder, "The visual object tracking vot2015 challenge results," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 1–23.

[7] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, and R. Pflugfelder, "The visual object tracking vot2016 challenge results," in *Proceedings of the IEEE international conference on computer vision workshops*, 2016, pp. 191–217.

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

[9] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, “Eco: Efficient convolution operators for tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6638–6646.

[10] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, “Joint group feature selection and discriminative filter learning for robust visual object tracking,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7950–7960.

[11] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting the circulant structure of tracking-by-detection with kernels,” in *European conference on computer vision*. Springer, 2012, pp. 702–715.

[12] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, “Visual object tracking using adaptive correlation filters,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2544–2550.

[13] 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, 2014.

[14] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer, “Adaptive color attributes for real-time visual tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1090–1097.

[15] Y. Li and J. Zhu, “A scale adaptive kernel correlation filter tracker with feature integration,” in *European conference on computer vision*. Springer, 2014, pp. 254–265.

[16] 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*, 2015, pp. 3074–3082.

[17] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, “Convolutional features for correlation filter based visual tracking,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 58–66.

[18] L. Zhang, M. Yang, and X. Feng, “Sparse representation or collaborative representation: Which helps face recognition?” in *2011 International conference on computer vision*. IEEE, 2011, pp. 471–478.

[19] T. Xu and X.-J. Wu, “Fast visual object tracking via distortion-suppressed correlation filtering,” in *2016 IEEE International Smart Cities Conference*. IEEE, 2016, pp. 1–6.

[20] 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*, 2016, pp. 1401–1409.

[21] M. Mueller, N. Smith, and B. Ghanem, “Context-aware correlation filter tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1396–1404.

[22] H. Kiani Galoogahi, A. Fagg, and S. Lucey, “Learning background-aware correlation filters for visual tracking,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1135–1143.

[23] T. Xu, X.-J. Wu, and J. Kittler, “Non-negative subspace representation learning scheme for correlation filter based tracking,” in *International Conference on Pattern Recognition*. IEEE, 2018, pp. 1888–1893.

[24] M. Jian, K.-M. Lam, J. Dong, and L. Shen, “Visual-patch-attention-aware saliency detection,” *IEEE transactions on cybernetics*, vol. 45, no. 8, pp. 1575–1586, 2014.

[25] Y. Rao, J. Lu, and J. Zhou, “Attention-aware deep reinforcement learning for video face recognition,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3931–3940.

[26] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, “Multi-context attention for human pose estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1831–1840.

[27] M. Jian, W. Zhang, H. Yu, C. Cui, X. Nie, H. Zhang, and Y. Yin, “Saliency detection based on directional patches extraction and principal local color contrast,” *Journal of Visual Communication and Image Representation*, vol. 57, pp. 1–11, 2018.

[28] L. Jiang, X.-J. Wu, and J. Kittler, “Dual attention mobdensenet (damdnet) for robust 3d face alignment,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[29] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, “Towards universal object detection by domain attention,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7289–7298.

[30] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, “An accelerated correlation filter tracker,” *Pattern Recognition*, p. 107172, 2019.

[31] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, “Learning spatially regularized correlation filters for visual tracking,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4310–4318.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[33] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[34] M. Mueller, N. Smith, and B. Ghanem, “A benchmark and simulator for uav tracking,” in *European conference on computer vision*. Springer, 2016, pp. 445–461.

[35] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Bhat, A. Lukezic, A. Eldesokey *et al.*, “The sixth visual object tracking vot2018 challenge results,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[36] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, “Lasot: A high-quality benchmark for large-scale single object tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5374–5383.

[37] H. T. Nguyen and A. W. Smeulders, “Fast occluded object tracking by a robust appearance filter,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 1099–1104, 2004.

[38] D. Comaniciu, V. Ramesh, and P. Meer, “Real-time tracking of non-rigid objects using mean shift,” in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, vol. 2. IEEE, 2000, pp. 142–149.

[39] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, “Incremental learning for robust visual tracking,” *International journal of computer vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[40] X. Mei and H. Ling, “Robust visual tracking using 1 minimization,” in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 1436–1443.

[41] T. Zhang, S. Liu, N. Ahuja, M.-H. Yang, and B. Ghanem, “Robust visual tracking via consistent low-rank sparse learning,” *International Journal of Computer Vision*, vol. 111, no. 2, pp. 171–190, 2015.

[42] X. Mei and H. Ling, “Robust visual tracking and vehicle classification via sparse representation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 11, pp. 2259–2272, 2011.

[43] S. Zhang, H. Zhou, F. Jiang, and X. Li, “Robust visual tracking using structurally random projection and weighted least squares,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 11, pp. 1749–1760, 2015.

[44] Y. Yang, W. Hu, W. Zhang, T. Zhang, and Y. Xie, “Discriminative reverse sparse tracking via weighted multitask learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 5, pp. 1031–1042, 2015.

[45] T. Zhang, C. Xu, and M.-H. Yang, “Robust structural sparse tracking,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 473–486, 2019.

[46] T. Zhang, A. Bibi, and B. Ghanem, “In defense of sparse tracking: Circulant sparse tracker,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3880–3888.

[47] F. Du, P. Liu, W. Zhao, and X. Tang, “Joint channel reliability and correlation filters learning for visual tracking,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[48] 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*, 2018, pp. 4844–4853.

[49] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, “Learning low-rank and sparse discriminative correlation filters for coarse-to-fine visual object tracking,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[50] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, “Accurate scale estimation for robust visual tracking,” in *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press, 2014.

[51] 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.

[52] 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*, 2017, pp. 6309–6318.

[53] C. Peng, F. Liu, J. Yang, and N. Kasabov, "Densely connected discriminative correlation filters for visual tracking," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 1019–1023, 2018.

[54] 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*, 2018, pp. 4904–4913.

[55] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. Lau, and M.-H. Yang, "Crest: Convolutional residual learning for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2555–2564.

[56] Z. He, Y. Fan, J. Zhuang, Y. Dong, and H. Bai, "Correlation filters with weighted convolution responses," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1992–2000.

[57] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*. Springer, 2016, pp. 850–865.

[58] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[59] 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*, 2017, pp. 4335–4343.

[60] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 472–488.

[61] J. Choi, H. Jin Chang, T. Fischer, S. Yun, K. Lee, J. Jeong, Y. Demiris, and J. Young Choi, "Context-aware deep feature compression for high-speed visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 479–488.

[62] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4649–4659.

[63] 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*, 2019, pp. 1369–1378.

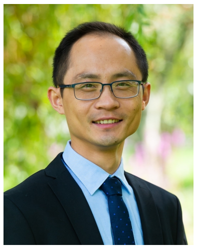
[64] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang, and H. Lu, "Gradnet: Gradient-guided network for visual object tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6162–6171.

Xue-Feng Zhu received the B.Eng. degree in internet of things engineering from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2017. He is a PhD student at the School of Internet of Things Engineering, Jiangnan University, Wuxi, China. His research interests include computer vision and machine learning.

Xiao-Jun Wu received the B.Sc. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1991. He received the M.S. degree and the Ph.D. degree in pattern recognition and intelligent systems from Nanjing University of Science and Technology, Nanjing, China, in 1996 and 2002, respectively.

He is currently a Professor in artificial intelligent and pattern recognition at the Jiangnan University, Wuxi, China. His research interests include pattern recognition, computer vision, fuzzy systems, neural networks and intelligent systems.

Tianyang Xu received the B.Sc. degree in electronic science and engineering from Nanjing University, Nanjing, China, in 2011. He received the PhD degree at the School of Internet of Things Engineering, Jiangnan University, Wuxi, China, in 2019. He is currently a research fellow at the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, United Kingdom. His research interests include visual tracking and deep learning. He has published several scientific papers, including International Conference on Computer Vision, IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, Pattern Recognition etc. He achieved top 1 tracking performance in the VOT2018 public dataset.



Zhen-Hua Feng (S'13-M'16) received the Ph.D. degree from the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, U.K. in 2016. He is currently a Senior Research Fellow at CVSSP, the University of Surrey. His research interests include computer vision, machine learning and pattern recognition.

He has published more than 40 scientific papers in top-ranking conferences and journals, including IJCV, CVPR, ICCV, IEEE TIP, IEEE TIFS, IEEE TCSVT, Pattern Recognition, Information Sciences etc. He has received the 2017 European Biometrics Industry Award from the European Association for Biometrics (EAB) and the 2018 AMDO Best Paper Award for Commercial Application.

Josef Kittler (M'74-LM'12) received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is a distinguished Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He published the textbook Pattern Recognition: A Statistical Approach and over 700 scientific papers. His publications have been cited more than 66,000 times (Google Scholar).

He is series editor of Springer Lecture Notes on Computer Science. He currently serves on the Editorial Boards of Pattern Recognition Letters, Pattern Recognition and Artificial Intelligence, Pattern Analysis and Applications. He also served as a member of the Editorial Board of IEEE Transactions on Pattern Analysis and Machine Intelligence during 1982-1985. He served on the Governing Board of the International Association for Pattern Recognition (IAPR) as one of the two British representatives during the period 1982-2005, President of the IAPR during 1994-1996.