Learning Low-rank and Sparse Discriminative Correlation Filters for Coarse-to-Fine Visual Object Tracking

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

***Abstract*—Discriminative correlation filter (DCF) has achieved advanced performance in visual object tracking with remarkable efficiency guaranteed by its implementation in the frequency domain. However, the effect of the structural relationship of DCF and object features has not been adequately explored in the context of the filter design. To remedy this deficiency, this paper proposes a Low-rank and Sparse DCF (LSDCF) that improves the relevance of features used by discriminative filters. To be more specific, we extend the classical DCF paradigm from ridge regression to lasso regression, and constrain the estimate to be of low-rank across frames, thus identifying and retaining the informative filters distributed on a low-dimensional manifold. To this end, specific temporal-spatial-channel configurations are adaptively learned to achieve enhanced discrimination and interpretability. In addition, we analyse the complementary characteristics between hand-crafted features and deep features, and propose a coarse-to-fine heuristic tracking strategy to further improve the performance of our LSDCF. Last, the augmented Lagrange multiplier optimisation method is used to achieve efficient optimisation. The experimental results obtained on a number of well-known benchmarking datasets, including OTB2013, OTB50, OTB100, TC128, UAV123, VOT2016 and VOT2018, demonstrate the effectiveness and robustness of the proposed method, delivering outstanding performance compared to the state-of-the-art trackers.**

***Index Terms*—Visual object tracking, discriminative correlation filter, lasso regression.**

I. Introduction

VISUAL object tracking is an important research topic in computer vision and pattern recognition, with practical uses in CCTV surveillance, robotics, medical image analysis and human-computer interactive applications. The visual tracking community, following the development of mathematical theories and modelling techniques in recent decades, has reported an impressive spectrum of achievements, ranging from template matching [1], statistical learning theory [2], particle filter [3], subspace learning [4], discriminative correlation filter [5], to deep neural networks [6]. Despite the existing success, it still remains a challenge to construct robust tracking algorithms due to the difficulties arising from complicated appearance variations of a target as well as its backgrounds, such as non-rigid deformation, illumination change, background clutter and occlusion. The increasingly challenging tracking benchmarks with corresponding evaluation methodologies, *e.g.*, OTB2013 [7], OTB100 [8], UAV123 [9], TC128 [10] and VOT [11], help to ensure a sustained vitality of the research area.

Among the existing state-of-the-art tracking approaches, the discriminative correlation filter (DCF) paradigm [5] has been demonstrated to achieve outstanding performance [12], [13]. The advantages of DCF derive from its spatial appearance model, circulant matrix structure [14] and efficient optimisation in the frequency domain. By considering the tracking task as ridge regression, DCF-based trackers are inherently computationally efficient and enjoy performance gains achieved from augmenting the set of training samples [15]. The basic DCF approach has been enhanced by several improvements achieved by considering more complex components, *e.g.*, scale detection [16], spatial regularisation [17], continuous domain mapping [18] and multi-response fusion [19]. A common characteristic of these approaches is that they employ the ℓ_2-norm to construct the objective function, which is a reasonable choice if one wishes to balance bias and variance of the estimate. However, the ℓ_2-norm sacrifices the model interpretability, especially when dealing with high-dimensional deep features. To obviate this issue, we propose to formulate the DCF paradigm in a lasso regression form, enforcing the estimate to be sparse. In addition, considering the fact that soft-thresholding operation is not stable, which may create excessive variation in the magnitude of the prediction signal, we constrain the estimate to be low-rank across frames. Consequently, we can obtain highly correlated filters among successive frames, distributed on a low-dimensional manifold. Fig. 1 depicts the learning scheme of the proposed LSDCF method. Given labelled training samples, the combination of sparse and low-rank constraints adaptively identify a specific temporal-spatial-channel configuration for the learning of discriminative filters.

Besides the mathematical formulation, another consensus amongst the advanced trackers is to use robust deep neural network features. With the fast development in deep neural networks, many computer vision and pattern recognition

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

Z. 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)

X.-J. Wu is with the School of Internet of Things Engineering, Jiangnan University, Wuxi, P.R. China. (e-mail:wu_xiaojun@jiangnan.edu.cn)

 DOI: 10.1109/TCSVT.2019.2945068

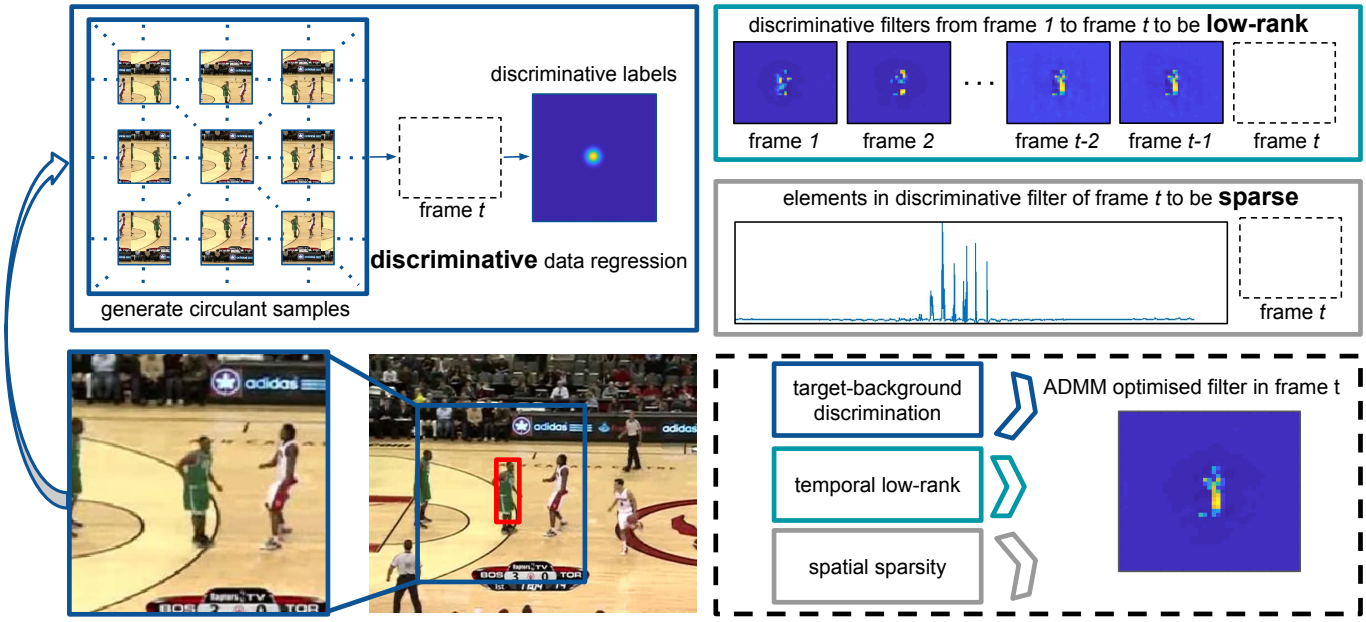


Fig. 1. Illustration of the learning framework of our LSDCF. Besides the discriminative data fitting, we propose to learn low-rank and sparse correlation filters. The sparsity is enforced to reduce less relevant features with enhanced discrimination and interpretation. The low-rank constraint is employed across temporal frames to improve the stability and smoothness. The proposed ADMM optimisation scheme iterates between updating corresponding sub-problems.

tasks[20], [21], [22], [23], including visual tracking[24], [25], have been shown to benefit from powerful deep discriminative features [26], [27], [28], [29]. The top-ranked trackers in recent competitions, *e.g.*, OTB100 [8], VOT2017 [30], VOT2018 [31], are all based on deep neural network features. It should be noted that the pooling operation in convolutional networks enhances the discrimination while sacrifices resolution, impeding accurate tracking. To simultaneously maximise discrimination and avoid spatial compression, existing state-of-the-art trackers usually fuse hand-crafted features and deep features for performance boosting. To explore the fusion of diverse features further, we analyse the characteristics of these two feature categories and propose a two-stage detection strategy. As deep features are expert in recognising a specific target in an image patch, we employ them to perform coarse target detection in the first stage. Then, a new search window is placed on the coarse detection result, and we perform fine-grained localisation by adding hand-crafted features. The proposed coarse-to-fine strategy enables robust tracking in challenging videos, devising an effective approach, combining the complementary characteristics of hand-crafted and deep features.

In this paper, we propose learning low-rank and sparse discriminative correlation filters (LSDCF) with a coarse-to-fine tracking strategy for robust visual tracking. The effectiveness of the proposed method has been extensively evaluated on a number of benchmarks such as OTB2013 [7], OTB50 [8], OTB100 [8], UAV123 [9], TC128 [10], VOT2016 [13], and VOT2018 [31]. The results demonstrate that LSDCF outperforms the state-of-the-art trackers. The main innovations of the proposed method include:

- A new discriminative filter learning formulation based on lasso regression and low-rank constraint. By design, specific elements are identified in the high-dimensional feature representations to learn a parsimonious filter system. The stability is achieved by enforcing the filters to be robust to temporal appearance variations.
- A coarse-to-fine heuristic tracking strategy that enhances the effectiveness of feature fusion. The complementary characteristics of hand-crafted and deep features are exploited to improve the accuracy and robustness of localisation and detection, respectively.
- A detailed analysis of each ingredient of LSDCF, *i.e.*, sparsity, low-rank, and coarse-to-fine strategy. The experimental results confirm the merit and effectiveness of each part of the entire tracker.

The paper is organised as follows: In Section II, we introduce related studies to the proposed tracking method. A detailed mathematical formulation and heuristic tracking strategy of the proposed LSDCF method are presented in Section III and furnished with an efficient optimisation approach. The experimental results and their analysis are reported in Section IV. Last, the conclusions are drawn in Section V.

II. Related Work

Many methods have been proposed to develop basic mathematical formulations and additional controlling systems for online visual tracking. In this section, we briefly review pertinent theories and techniques related to our work. More detailed introductions have been provided in recent surveys [32], [8], [13], [33].

A. Mathematical Formulation

One basic element in constructing an object appearance model is the mathematical formulation, through which a vision task can be transformed to an algorithmic problem.

The Lucas-Kanade algorithm [1], [34] is the seminal generative learning method that achieves tracking by template matching, with the assumption that the flow is essentially constant in a local neighbourhood of the pixel under consideration. But this assumption is fragile due to the complicated appearance variations of an object. To create a better model, the mean shift method [35] has been applied to visual object tracking by iteratively locating the maximal mode of a density function that describes the target appearance. Though global appearance is reflected in a density function, mean shift trackers suffer from the presence of local minima. In the framework of Bayesian statistical inference, particle filter method [3] was developed to estimate the posterior distribution of a tracking target, alleviating the local minima problem by sampling operations. Such a prediction-updating paradigm satisfies the non-linear and dynamic properties of a changing target, especially in high-dimensional feature spaces. Consequently, particle filter has been a popular modelling approach in the visual tracking field for decades, even successfully integrated with the tides of compress sensing [36] and deep learning [37].

Compress sensing [38] results in the development of sparse and low-rank trackers. For sparse-representation-based trackers [39], [40], the reconstruction error between a candidate and a dictionary provides a robust appearance similarity metric. In contrast, low-rank trackers [41], [42] enforce a low-rank constraint on the representation matrix corresponding to multiple candidates, highlighting the inherent low-rank structure of candidate representations that can be learned jointly.

Besides the above generative tracking formulations, discriminative learning approaches have also been explored in parallel. Instead of minimising the difference between candidate and appearance density function, discriminative trackers maximise the classification score to distinguish the target from background. The fundamental discriminative trackers emerged via the application of statistical learning theory [2], that aims at learning a classifier given positive and negative labelled samples. To enhance the discrimination in complex scenarios, online-boosting-based trackers [43], [44], [45] have been proposed to train a strong classifier by integrating multiple weak classifiers. The structured output strategy [46] has also been designed to further jointly couple the input and output. In addition, to capitalise on the achievements of deep learning, end-to-end learnt deep neural networks have been developed to train a localisation system using large-scale vision datasets [47], [48], [49].

Recent discriminative approaches focus on correlation filters [5], that formulate a tracking task as ridge regression with guaranteed efficiency provided by circulant matrix [14] and Fast Fourier Transform (FFT). Outstanding performance has been achieved by the combination of discriminative correlation filters and deep features on several benchmarks [13], [31], [50]. But the basic formulation of DCF has not been deeply questioned to see whether further improvements in performance can be gained with deep features. To rectify this, we propose to learn low-rank and sparse correlation filters to achieve enhanced discrimination and robust visual object tracking.

B. Controlling System

Despite the success in developing rigorous mathematical formulations in visual tracking, a practical tracking task can not be modelled by existing theories accurately. To mitigate this issue, additional controlling strategies are necessary to balance the characteristics of the theories reflecting simplifying assumptions about the universe of target tracking and their validity in challenging video sequences.

Considering the temporal dimension, a forgetting factor [4] was proposed and widely used to address robust model updating in tracking, with more importance granted to recently-acquired appearance and less to earlier observations. Such a strategy enables dynamic appearance fusion in tracking, but a fixed updating rate suffers from the risk of potential unrecoverable failures. Some recent studies deal with the temporal correlations via two approaches, *i.e.*, involving more historical samples and exploiting target re-detection. Specifically, by incorporating weighted historical samples into the learning stage, the resulting adaptive decontamination of the training set [51], [52], [53], [54] achieves robust tracking performance. In the same spirit, Efficient Convolution Operators (ECO) [55] significantly improve the computational efficiency by decreasing the number of historical frames via a data clustering technique. Re-detection [56], on the other hand, intuitively alleviates tracking shifts and failures by employing a separate classifier that performs target detection when a pre-specified condition is satisfied.

Besides temporal coherence, spatial configuration is another controlling element that has gained emphasis in visual tracking. Unlike other data, *e.g.*, gene and aurals signal, natural images are the projections of the 3D space onto a 2D plane. Accordingly, significant improvements have been made by identifying effective spatial configurations. Confronted with local appearance variations, the fragments-based tracking method [57], [58] was proposed to estimate the target by fragments, effectively improving tracking performance under partial occlusions. Similar strategies were applied in constructing the structural sparse model [59] and the structural/patch-based correlation filter model [60], [61], [62], [63]. To emphasise the spatial configuration in a global pattern, spatial regularisation [17] was proposed in the DCF paradigm, with more energy concentrated in the central region. The corresponding, context-aware [64] and background-aware [65] DCF trackers have been shown to achieve enhanced discrimination via expanding the appearance information from the surroundings of a target.

To further improve the tracking performance, we propose a coarse-to-fine heuristic tracking strategy as an additional controlling method. This strategy realises adaptive feature fusion by hierarchically highlighting the advantages of deep neural network features and hand-crafted features.

III. The Proposed LSDCF Method

In this section, we introduce our LSDCF method by extending the classical DCF paradigm from ridge regression to lasso regression, with an additional low-rank constraint across temporal frames. A detailed optimisation process is developed by employing the augmented Lagrange method. Additionally, a novel coarse-to-fine heuristic tracking strategy is proposed to unveil complementary characteristics between hand-crafted and deep neural network features.

A. Learning Low-rank and Sparse Discriminative Correlation Filters

1) Ridge Regression Formulation: Given the location of a target at the current frame, the aim of visual object tracking is to predict the location of the target in the next frame. In the learning stage, we aim to train a discriminative filter that obtains high-value responses around the target centre and low-value responses for the background. Following the DCF formulation, a search window centred around the target is extracted with feature representation $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$. Its corresponding circulant matrix is denoted as [14]:

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & x_3 & \ldots & x_n \\ x_n & x_1 & x_2 & \ldots & x_{n-1} \\ x_{n-1} & x_n & x_1 & \ldots & x_{n-2} \\ \vdots & & \ddots & & \vdots \\ x_2 & x_3 & x_4 & \ldots & x_1 \end{bmatrix}, \tag{1}$$

in which each row is considered as an augmented sample, such that $\mathbf{X}$ can be directly employed as a training data matrix [15], [66]. Given labelled training pairs $\{\mathbf{X}, \mathbf{y}\}$, the learning stage is formulated as ridge regression to find an optimal filter $\mathbf{w} \in \mathbb{R}^n$ to distinguish the target region from background. (For simplicity, we focus on single-channel, one-dimensional signals here. Multi-channel, two-dimensional representations are formulated in Subsection III-A4):

$$\mathbf{w} = \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2, \tag{2}$$

where λ is the weight of the regularisation term. Referring to the Fourier theorem, a closed-form solution in the frequency domain can be obtained as:

$$\hat{\mathbf{w}} = \frac{\hat{\mathbf{x}} \odot \hat{\mathbf{y}}^*}{\hat{\mathbf{x}} \odot \hat{\mathbf{x}}^* + \lambda \mathbf{1}}, \tag{3}$$

where $\odot$ is the element-wise multiplication, $\mathbf{1}$ is an all-ones vector with the same size as $\hat{\mathbf{x}}$, $\hat{\cdot}$ denotes Fourier representation and $\cdot^*$ is the complex conjugate.

2) Lasso Regression Formulation: The original ridge regression formulation achieves predominant performance via combining hand-crafted features, *e.g.*, gray pixels, histogram of oriented gradients (HOG) and Colour Names (CN) [5], [16], [19]. However, to realise a bias-variance trade-off, ridge regression enables all the elements in the model activated, without the ability to produce a parsimonious model. To sharpen the focus of the discriminative filters on relevant features, we favour a simpler estimate (filters), with more attention given to the relations between discriminative responses and data correlations. Here, a simpler estimate always implies that the elements of the model are sparse, and most energy is concentrated on specific dimensions. In addition, a sparse estimate is more appropriate for deep features as the dimensionality of deep models, *e.g.*, AlexNet [26], VGG [27], GoogLeNet [28] and ResNet [29], exceeds 10^5 even in one convolutional layer. Therefore, we aim to learn a sparse discriminative filter via ridge regression [67]:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1, \tag{4}$$

such that the parsimony can be achieved for the discriminative filters even in high-dimensional feature spaces.

3) Low-rank Filter Learning: Despite the success of lasso regression in pattern recognition, it suffers from excessive sensitivity and unstable performance [68]. To mitigate this problem, in our lasso regression formulation we improve the robustness by encouraging temporal smoothness. Specifically, a low-rank constraint is imposed on the estimates across video frames, so that the quality of inherent features can be nurtured. We reformulate the objective as:

$$\begin{aligned} &\mathbf{w} = \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1 \\ &\text{s.t.} \quad rank\,(\mathbf{W}_t) - rank\,(\mathbf{W}_{t-1}) < \tau, \end{aligned} \tag{5}$$

where $\mathbf{W}_t = \left[\frac{t}{t-1}(1-\alpha)\mathbf{W}_{t-1}, t\alpha\mathbf{w}\right]$ $(t > 1)$ stores the historical learned filters with a forgetting factor α, and τ is a pre-defined threshold. We omit the frame index for $\mathbf{w}$, as $\mathbf{w} = \mathbf{w}_t$. Here, we impose low-rank across all the frames. as the equal rank constraint is enforced starting from the second frame, *i.e.*, $rank\,(\mathbf{W}_2) = rank\,(\mathbf{W}_1)$. However, it is inefficient to calculate $rank\,(\mathbf{W}_t)$, especially in long-term videos with a large number of frames. Therefore, we use its sufficient statistics as a substitute:

$$\|\mathbf{w} - \boldsymbol{\mu}_{t-1}\| < \tau' \tag{6}$$

where $\boldsymbol{\mu}_{t-1}$ is the mean vector of $\mathbf{W}_{t-1}$, which is the same as the updated filter in the DCF paradigm. The proof of sufficiency is intuitive as we impose a linear relationship among temporal filters.

To emphasise the low-rank constraint, we rewrite the objective to learn the low-rank and sparse discriminative correlation filters as:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w} - \boldsymbol{\mu}_{t-1}\|_2^2. \tag{7}$$

It should be noted that $\boldsymbol{\mu}_{t-1}$ is the weighted mean of historical filters from the first frame to the $(t-1)$-th frame which can be incrementally updated easily, such that it is the exact filter used for tracking in the t-th frame. We realise this low-rank property across temporal frames by employing the ℓ_2-norm and storing the weighted mean only, simplifying storage and computational cost. The proposed low-rank and sparse properties are experimentally verified in Section IV-B2.

4) Multi-channel LSDCF: To achieve a structured combination of multi-channel features in the DCF paradigm, we extend our formulation to multi-channel LSDCF. We denote the multi-channel feature representations, *e.g.*, CN [69], HOG [70] and ResNet [29], as $\mathcal{X} \in \mathbb{R}^{N \times N \times C}$, where $N \times N$ is the spatial resolution and C is the number of channels. We denote $\mathcal{W}$ as the discriminative correlation filters with the same size as

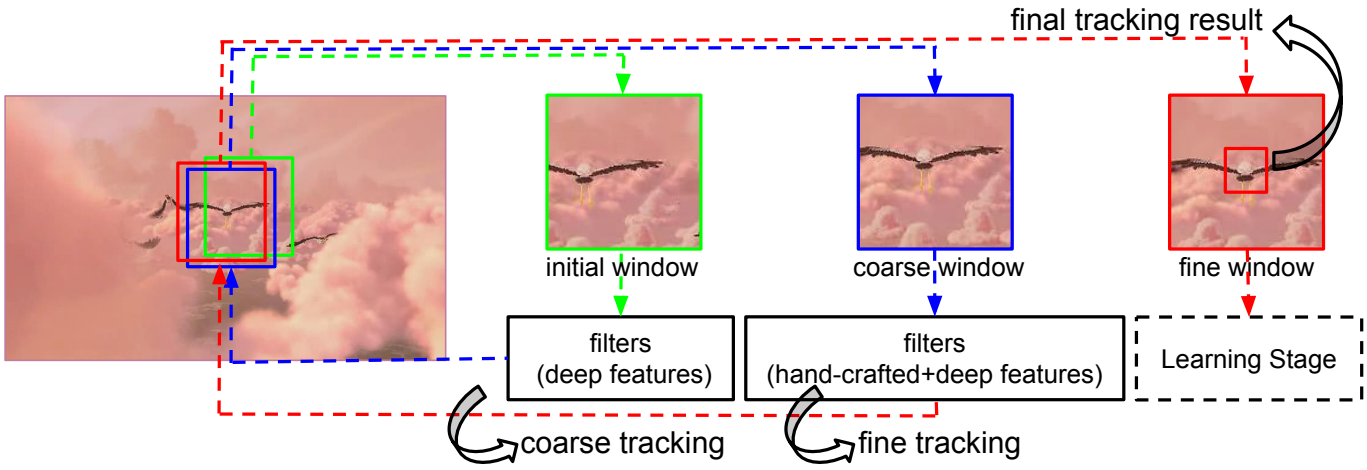


Fig. 2. Illustration of the proposed coarse-to-fine tracking strategy. First, an initial window (*green*) is extracted based on the tracking result from the last frame. Coarse tracking is performed on the initial window with deep features to obtain a coarse result. Second, a coarse window (*blue*) is extracted based on the coarse result. Fine-grained tracking is then performed on the coarse window with both hand-crafted and deep features to obtain the final result. Last, a fine-grained search window (*red*) is extracted based on the final result and is used in the learning stage.

$\mathcal{X}$, and the incremental updated mean of $\mathcal{W}$ as $\mathcal{U}$ (the multi-channel, 2-dimensional extension of $\boldsymbol{\mu}_{t-1}$).

Given $\mathcal{X}$ and the 2-dimensional pre-defined gaussian shaped labels $\mathbf{Y} \in \mathbb{R}^{N\times N}$ [15], we formulate the objective of multi-channel LSDCF as follows:

$$\begin{aligned}\mathcal{W} = \arg\min_{\mathcal{W}} &\left\| \sum_{k=1}^{C} \mathcal{X}^k \circledast \mathcal{W}^k - \mathbf{Y} \right\|_F^2 + \lambda_1 \sum_{k=1}^{C} \left\| \mathcal{W}^k \right\|_{1,1} \\ &+ \lambda_2 \sum_{k=1}^{C} \left\| \mathcal{W}^k - \mathcal{U}^k \right\|_F^2 ,\end{aligned} \tag{8}$$

where $\circledast$ is the circular convolution operator [5], and $\mathcal{W}^k \in \mathbb{R}^{N\times N}$ denotes the k-th channel of $\mathcal{W}$. The Frobenius norm of a matrix $\mathbf{M} \in \mathbb{R}^{N\times N}$ is defined as $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} m_{i,j}^2}$, where $m_{i,j}$ corresponds to the i-th row j-th column element in $\mathbf{M}$. The $\ell_{1,1}$-norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_{1,1} = \sum_{i=1}^{N}\sum_{j=1}^{N} |m_{i,j}|$.

B. Optimisation

Due to the convexity of the proposed formulation, we apply the augmented Lagrange method [71] to optimise Obj. (8). Concretely, we introduce slack variable $\mathcal{W}' = \mathcal{W}$ and construct the following Lagrange function:

$$\begin{aligned}\mathcal{L} = &\left\| \sum_{k=1}^{C} \mathcal{X}^k \circledast \mathcal{W}^k - \mathbf{Y} \right\|_F^2 + \lambda_1 \sum_{k=1}^{C} \left\| \mathcal{W}'^k \right\|_{1,1} \\ &+ \lambda_2 \sum_{k=1}^{C} \left\| \mathcal{W}^k - \mathcal{U}^k \right\|_F^2 + \frac{\nu}{2} \sum_{k=1}^{C} \left\| \mathcal{W}^k - \mathcal{W}'^k + \frac{\boldsymbol{\Gamma}^k}{\nu} \right\|_F^2 ,\end{aligned} \tag{9}$$

where $\boldsymbol{\Gamma}$ is the Lagrange multiplier sharing the same size as $\mathcal{X}$ and ν is the corresponding penalty weighting parameter. Then Alternating Direction Method of Multipliers [72] is employed to achieve iterative optimisation with guaranteed convergence as follows.

1) Optimising $\mathcal{W}$: In order to optimise $\mathcal{W}$, we employ circulant structure [5] and Parseval's theorem to solve the following subproblem in the frequency domain:

$$\begin{aligned}\min &\left\| \sum_{k=1}^{C} \hat{\mathcal{X}}^k \odot \hat{\mathcal{W}}^k - \hat{\mathbf{Y}} \right\|_F^2 + \lambda_2 \sum_{k=1}^{C} \left\| \hat{\mathcal{W}}^k - \hat{\mathcal{U}}^k \right\|_F^2 \\ &+ \frac{\nu}{2} \sum_{k=1}^{C} \left\| \hat{\mathcal{W}}^k - \hat{\mathcal{W}}'^k + \frac{\hat{\boldsymbol{\Gamma}}^k}{\nu} \right\|_F^2 .\end{aligned} \tag{10}$$

A closed-form solution can be derived as [73]:

$$\hat{\mathbf{w}}_{i,j} = \left(\mathbf{I} - \frac{\hat{\mathbf{x}}_{i,j}\hat{\mathbf{x}}_{i,j}^{\top}}{\lambda_2 + \nu/2 + \hat{\mathbf{x}}_{i,j}^{\top}\hat{\mathbf{x}}_{i,j}} \right) \mathbf{q}, \tag{11}$$

where $\mathbf{q} = \left(\hat{\mathbf{x}}_{i,j}\hat{y}_{i,j} + \nu\hat{\mathbf{w}}'_{i,j} - \nu\hat{\boldsymbol{\gamma}}_{i,j} + \lambda_2\hat{\mathbf{u}}_{i,j} \right) / (\lambda_2 + \nu)$, vectors $\hat{\mathbf{w}}_{i,j}$ ($\hat{\mathbf{w}}_{i,j} = \left[\hat{w}_{i,j}^1, \hat{w}_{i,j}^2, \ldots, \hat{w}_{i,j}^C \right] \in \mathbb{C}^C$), $\hat{\mathbf{x}}_{i,j}$, and $\hat{\mathbf{u}}_{i,j}$ denote the i-th row j-th column units of $\hat{\mathcal{W}}$, $\hat{\mathcal{X}}$ and $\hat{\mathcal{U}}$, respectively, across all C channels.

2) Optimising $\mathcal{W}'$: In order to optimise $\mathcal{W}'$, we minimise the following subproblem:

$$\min \lambda_1 \sum_{k=1}^{C} \left\| \mathcal{W}'^k \right\|_{1,1} + \frac{\nu}{2} \sum_{k=1}^{C} \left\| \mathcal{W}^k - \mathcal{W}'^k + \frac{\boldsymbol{\Gamma}^k}{\nu} \right\|_F^2 . \tag{12}$$

The shrinkage operator can be used to realise a closed-form solution for each element separately:

$$w_{i,j}'^k = \operatorname{sign}(p) \max\left(0, |p| - \frac{\lambda_1}{\nu} \right), \tag{13}$$

where $p = w_{i,j}^k + \frac{\gamma_{i,j}^k}{\nu}$, with $w_{i,j}^k$ and $\gamma_{i,j}^k$ being the values corresponding to the elements at the i-th row, j-th column and k-th channel in $\mathcal{W}$ and $\boldsymbol{\Gamma}$, respectively.

3) Optimising other variables: We update the multiplier $\mathbf{\Gamma}$ and the penalty ν in each iteration as:

$$\begin{cases} \mathbf{\Gamma} = \mathbf{\Gamma} + \nu\,(\mathcal{W} - \mathcal{W}') \\ \nu = \min\,(\rho\nu, \nu_{\max}) \end{cases}. \tag{14}$$

where ρ controls the strictness of the penalty and $\nu_{\max}$ is the upper limit of the penalty.

C. Coarse-to-Fine Tracking Strategy

After obtaining the multi-channel filter $\mathcal{W}$, the weighted mean filter $\mathcal{U}$ is updated with the forgetting factor α:

$$\mathcal{U} = \alpha\mathcal{W} + (1-\alpha)\,\mathcal{U}. \tag{15}$$

Given the feature representations, $\mathcal{X}$, of a search window in a new frame, the task of the detection stage for the discriminative correlation filters is to obtain the response map $\mathbf{Y}$, which can be efficiently calculated in the frequency domain:

$$\hat{\mathbf{Y}} = \sum_{k=1}^{C} \hat{\mathcal{X}}^k \odot \hat{\mathcal{U}}^k. \tag{16}$$

Then the maximal value in $\mathbf{Y}$ corresponds to the target location.

Existing tracking algorithms combine hand-crafted and deep features by directly summing the corresponding response maps without considering the diversity between them [18], [55]. We analyse the differences of hand-crafted and deep features with the conclusion that deep features are expert in robust target recognition but perform worse in terms of accurate target localisation. An intuitive explanation is that deep features can select the most similar sample corresponding to a specific object, but can not guarantee the object is in the centre of the image. In contrast, hand-crafted features can achieve a better target localisation if it is already near the ground truth. Based on the above observation, we propose a coarse-to-fine heuristic tracking strategy that achieves effective multi-feature fusion. Fig. 2 illustrates the proposed coarse-to-fine tracking strategy.

1) For Deep Features: We employ deep features to realise coarse target detection. Specifically, we extract an initial search window based on the central position of the tracking result from the last frame, and perform correlation filtering using only the filters corresponding to deep features. At this stage, we expect to roughly locate the target. Then, we extract an updated search window based on the coarse estimate. In addition, as deep features are of high volume, and robust to appearance variations, we execute the learning stage every K frames for deep features, with the mean representations of previous $2K$ frames as training samples.

2) For Hand-crafted Features: We perform fine-grained target detection in the extracted coarse search window. A basic assumption is that the ground truth is already close to the central region in the coarse search window. Then, a combination of hand-crafted and deep filters is employed to obtain the final response map. Note that as hand-crafted features suffer from appearance variations more than deep features, we learn the hand-crafted filters in each frame.

IV. Experiments

A. Implementation Details and Evaluation Methodology

To evaluate the performance of the proposed LSDCF method, we implement LSDCF in MATLAB 2016a on an Intel i5 2.50 GHz CPU and GTX 960 GPU. The MATLAB code is publicly available at github[1]. We use CN [69], HOG [70] as hand-crafted features, with $\lambda_1 = 1.2 \times 10^{-5}$, $\lambda_2 = 40$ and $\alpha = 0.6$. ResNet-50 [29] is employed as deep features with $\lambda_1 = 2 \times 10^{-6}$, $\lambda_2 = 6$, $K = 4$ and $\alpha = 0.2$. Deep feature representations are extracted using MatConvNet Toolbox [2] [74]. The parameters are fixed for all datasets.

We evaluate our LSDCF on a number of well-known benchmarking datasets, including OTB2013 [7], OTB50 [8], OTB100 [8], UAV123 [9], TC128 [10], VOT2016 [13], and VOT2018 [31]. We also compare our LSDCF with a number of state-of-the-art trackers, such as VITAL [75] (CVPR18), MetaT [76] (ECCV18), ECO [55] (CVPR17), MCPF [37] (CVPR17), CREST [49] (ICCV17), BACF [65] (ICCV17), CFNet [48] (CVPR17), STAPLE_CA [64] (CVPR17), ACFN [77] (CVPR17), CSRDCF [78] (CVPR17), C-COT [18] (ECCV16), Staple [19] (CVPR16), SiamFC [47] (ECCV16), SRDCF [17] (ICCV15), KCF [5] (TPAMI15), SAMF [79] (ECCVW14) and DSST [80] (TPAMI17). For VOT2016 and VOT2018, we compare our LSDCF with the corresponding top-ranking trackers (C-COT, TCNN, SSAT, MLDF Staple, DDC, EBT, SRBT, STAPLE+, DNT, ECO, CFWCR, LSART, UPDT, SiamRPN, MFT and LADCF) [3] that participated in the VOT challenges [13], [31].

To measure the tracking performance, we use the precision plot and the success plot [7]. Specially, the precision plot measures the percentage of frames, with the distance of the tracking results from the ground truth less than a certain number of pixels. The success plot measures the proportion of successfully tracked frames with the threshold ranging from 0 to 1 (a result is considered successful if the overlap of the two bounding boxes exceeds a given threshold). Four additional objective criteria, *i.e.*, center location error (CLE), distance precision (DP), overlap precision (OP) and area under curve (AUC), are employed to characterise the performance. DP is the corresponding precision plot value (illustrated in the legend of precision plot) when the threshold is set to 20 pixels. CLE is the average central distance between the predicted and ground truth locations of a target. OP is the corresponding success plot value when the threshold is set to 0.5. AUC is the expected success rate (illustrated in the legend of success plot) in terms of overlap evaluation. For VOT2016 and VOT2018, we use the expected average overlap (EAO), accuracy value (A) and robustness value (R) as the evaluation metrics [13].

B. Component Analysis

1) Impact on Quantitative Performance: We first evaluate the effect of the innovative components in LSDCF, *i.e.*, sparsity (S), low-rank constraint (L) and coarse-to-fine strategy (CF). We consider all the combinations and construct

[1] https://github.com/XU-TIANYANG/LSDCF
[2] http://www.vlfeat.org/matconvnet/
[3] http://www.votchallenge.net

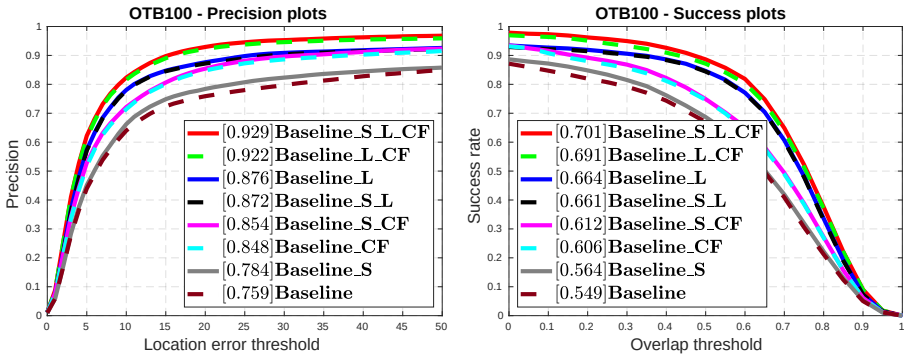


Fig. 3. The performance of different components and their combinations in LSDCF, evaluated on OTB100. The precision plots (*left*) with **DP** and the success plots (*right*) with **AUC** in the legends are presented.

TABLE I

RANK OF DIFFERENT DCF TRACKERS BY CONCATENATING THE FILTERS LEARNT FOR ALL THE FRAMES IN SEQUENCES *Deer*, *Basketball*, *Boy*, *David3*, *Girl*, *Suv*, *Skater* AND *Woman*. THE LOWEST THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE AND BROWN FONTS.

Sequences [#frames]	BACF	STAPLE_CA	SRDCF	C-COT	ECO	KCF	Staple	DSST	SAMF	CSRDCF	**LSDCF**
Deer [71]	7	14	11	**3**	4	71	14	**3**	32	11	**2**
Basketball [725]	42	134	46	**10**	**23**	526	87	16	141	140	**9**
Boy [602]	27	63	39	**8**	**19**	274	61	37	61	46	**4**
David3 [252]	13	53	16	**3**	**8**	252	33	11	23	29	**6**
Girl [500]	26	57	32	**8**	**18**	267	50	36	53	73	**5**
Suv [945]	32	49	36	**4**	**16**	701	49	18	39	78	**6**
Skater [160]	17	38	14	**3**	19	160	23	**12**	18	31	**5**
Woman [597]	29	111	20	**6**	**15**	384	65	35	73	68	**7**

TABLE II

A COMPARISON OF OUR LSDCF WITH STATE-OF-THE-ART METHODS ON OTB2013, OTB50 AND OTB100 IN TERMS OF **OP** AND **CLE**. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE AND BROWN FONTS.

		KCF	SAMF	DSST	SRDCF	SiamFC	Staple	C-COT	CSRDCF	ACFN
Mean **OP/CLE** (%/pixels)	**OTB2013** [7]	60.8/36.3	69.6/29.0	59.7/39.2	76.0/36.8	77.9/29.7	73.8/31.4	83.7/15.6	74.4/31.9	75.0/18.7
	OTB50 [8]	47.7/54.3	59.3/40.5	45.9/59.5	66.1/42.7	68.0/36.8	66.5/32.3	80.9/**12.3**	66.4/30.3	63.2/32.1
	OTB100 [8]	54.4/45.1	64.6/34.6	53.0/49.1	71.1/39.7	73.0/33.2	70.2/31.8	82.3/**14.0**	70.5/31.1	69.2/25.3
Mean **FPS**		**82.7**	11.5	15.6	2.7	12.6	**23.8**	2.2	4.6	13.8
		STAPLE_CA	CFNet	BACF	CREST	MCPF	ECO	MetaT	VITAL	**LSDCF**
Mean **OP/CLE** (%/pixels)	**OTB2013**	77.6/29.8	78.3/35.2	84.0/26.2	86.0/**10.2**	85.8/11.2	**88.7**/16.2	85.6/11.5	**91.4**/**7.4**	**93.9**/**7.8**
	OTB50	68.1/36.3	68.8/36.7	70.9/30.3	68.8/32.6	69.9/30.9	**81.0**/13.2	73.7/17.0	**81.3**/**12.5**	**83.1**/**10.2**
	OTB100	73.0/33.1	73.6/36.0	77.6/28.2	77.6/21.2	78.0/20.9	**84.9**/14.8	79.8/14.2	**86.5**/**9.9**	**88.6**/**9.0**
Mean **FPS**		**18.1**	8.7	16.3	10.1	0.5	8.5	0.8	1.3	6.8

7 trackers. The results on OTB100 are shown in Fig. 3. The Baseline is the original DCF tracker equipped with the same features as our LSDCF. Generally, the proposed sparsity, low-rank constraint and coarse-to-fine strategy produce improvements for the DCF paradigm, both separately and in combinations. Compared with the Baseline, the low-rank constraint (Baseline_L) significantly improves the performance in terms of DP and AUC by 11.7% and 11.5%, respectively. Intuitively explained, a low-rank constraint across temporal frames enables the learned filters to become more invariant to appearance variations. Sparsity and the coarse-to-fine strategy also lead to improvements in the tracking performance. The combination of low-rank and the coarse-to-fine strategy achieves a performance gain from 87.6% to 92.2% in DP and from 66.4% to 69.1% in AUC, compared with Baseline_L. Note that we realise sparsity via the classical lasso method. The gain from sparsity is not as high as that from the other components. But the sparsity alone (Baseline_S) still improves the Baseline from 75.9% to 78.4 in DP and from 54.9% to 56.4%. In addition, the combination of all components (Baseline_S_L_CF) is exactly the proposed LSDCF tracker, which achieves the best performance compared with the other combinations. The above results demonstrate the effectiveness of the proposed low-rank and sparse DCF formulation as well as the coarse-to-fine tracking strategy.

2) Analysis of the Low-rank and Sparsity Constraints: Here, we qualitatively verify the low-rank and sparsity properties of the proposed formulation to gain intuitive understanding

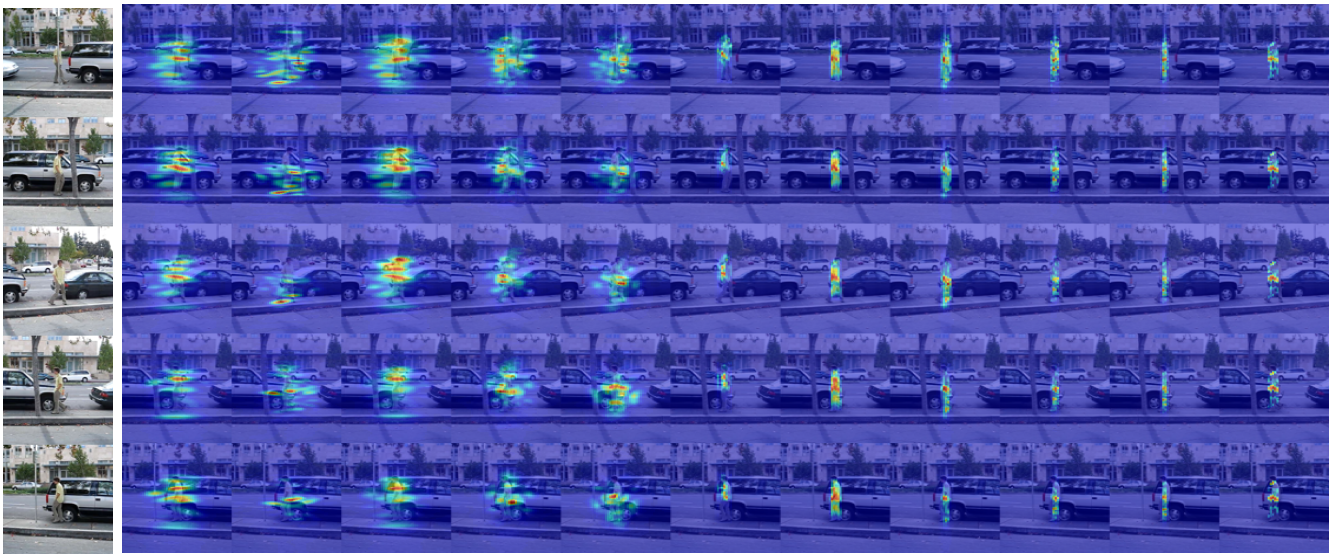


Fig. 4. Illustration of the sparsity of discriminative filters of several methods. We collect the learnt filters from 11 DCF-based trackers, *i.e.*, KCF, DSST, SAMF, Staple, STAPLE_CA, CSRDCF, BACF, SRDCF, ECO, C-COT and our LSDCF, in sequence *David3* [7]. As all these trackers employ HOG features, we only present the filters obtained using HOG feature channels. *David3* contains 252 frames. We visualise the corresponding filters in frame #50 (the 1st row), #100 (the 2nd row), #150 (the 3rd row), #200 (the 4th row) and #250 (the 5th row). To better visualise the sparsity, we present the heat-maps of the filters by accumulating the energy across all the channels.

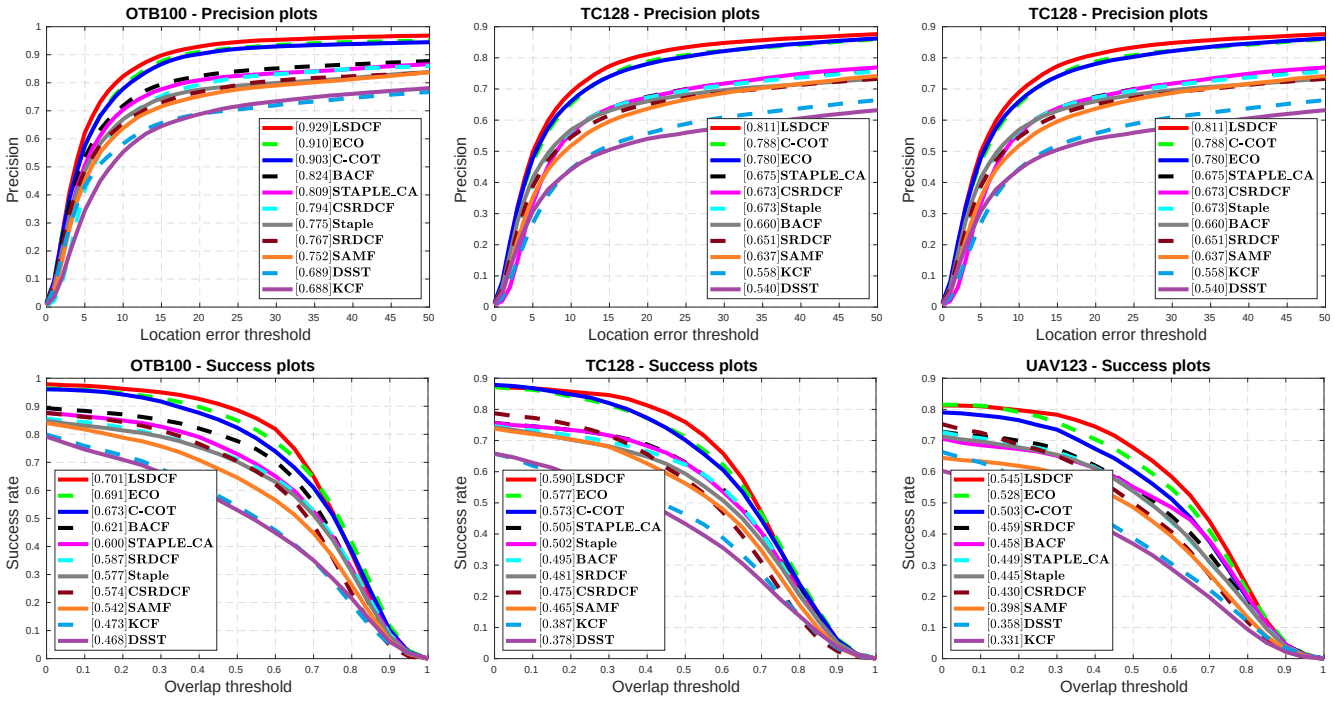


Fig. 5. Evaluations on OTB100, TC128 and UAV123. The precision plots (*first row*) with **DP** in the legend and the success plots (*second row*) with **AUC** in the legend are presented.

of the filter behaviour.

For sparsity, we visualise the obtained discriminative filters of several DCF-based trackers in Fig. 4. We note that the filters of KCF, DSST, SAMF, Staple and STAPLE_CA are densely distributed in the spatial domain. The remaining 6 trackers share the sparsity property. Specifically, CSRDCF and BACF enforce the sparsity by pre-defined masks (only specific filter regions are allowed for filtering) for the filters. SRDCF, ECO and C-COT achieve sparsity by spatial regularisation, with more filter energy concentrating at the image centre. On the other hand, the proposed LSDCF realises sparsity without using a pre-defined mask or weighting scheme. The filters are adaptively shrunk to specific elements by discriminative data fitting and combined low-rank and sparse regularisation terms.

TABLE III
TRACKING RESULTS ON VOT2016. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE AND BROWN FONTS.

	DNT	STAPLE+	SRBT	EBT	DDC	Staple	MLDF	SSAT	TCNN	C-COT	**LSDCF**
EAO	0.269	0.286	0.290	0.291	0.293	0.295	0.311	0.329	**0.327**	**0.331**	**0.407**
Accuracy	0.515	**0.559**	0.497	0.465	0.542	0.547	0.492	**0.579**	0.555	0.541	**0.587**
Robustness	0.33	0.37	0.35	0.25	0.34	0.38	**0.23**	0.29	0.27	**0.24**	**0.18**

TABLE IV
TRACKING RESULTS ON VOT2018. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE AND BROWN FONTS.

	ECO	STAPLE_CA	CFWCR	LSART	UPDT	SiamRPN	MFT	LADCF	**LSDCF**
EAO	0.280	0.286	0.303	0.323	0.378	0.383	**0.385**	**0.389**	**0.387**
Accuracy	0.483	0.509	0.484	0.493	**0.536**	**0.586**	0.505	0.503	**0.523**
Robustness	0.276	0.281	0.267	0.218	0.184	0.276	**0.140**	**0.159**	**0.145**

TABLE V
THE **DP** AND **AUC** RESULTS ON OTB100, PARAMETERISED BY 11 ATTRIBUTES. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE AND BROWN FONTS.

		STAPLE_CA	CFNet	BACF	CREST	C-COT	MCPF	ECO	MetaT	VITAL	**LSDCF**
	BC	78.2/58.6	73.4/56.5	83.0/62.5	82.9/61.8	88.2/65.2	82.3/60.1	**94.2/70.0**	92.6/67.4	**94.6/69.3**	**93.2/70.3**
	DEF	76.0/56.6	69.6/50.8	77.8/58.2	77.6/56.9	85.9/61.4	81.6/57.0	**85.9/63.3**	83.8/62.0	**90.1/65.1**	**90.9/66.6**
	FM	75.3/58.7	71.6/55.8	80.7/60.5	79.2/62.7	**88.3/67.6**	84.5/59.7	87.8/**68.3**	79.3/62.6	**89.2**/67.2	**90.5/69.4**
	IPR	80.6/57.4	76.8/57.2	79.5/58.4	85.3/61.7	87.7/62.7	88.8/62.0	**89.2/65.5**	87.7/63.5	**91.8/66.5**	**92.4/67.3**
Mean	**IV**	81.6/61.3	70.5/54.9	83.0/64.2	87.6/64.4	88.4/68.2	88.1/62.8	**91.4/71.3**	86.4/63.4	**93.4/70.3**	**92.7/71.3**
DP/AUC	**LR**	81.9/44.8	81.0/58.6	79.5/51.4	86.6/47.3	**97.5/62.9**	**96.3**/58.7	88.2/59.1	90.1/47.2	94.2/**64.2**	**99.8/69.0**
(%/%)	**MB**	74.7/57.7	63.3/51.4	76.5/58.5	81.3/65.5	**89.9/70.6**	84.0/59.9	**89.7/70.9**	81.8/65.2	88.0/68.6	**89.7**/**70.3**
	OCC	74.0/56.1	70.3/54.0	74.5/57.6	78.6/59.2	**90.4/67.4**	86.2/62.0	**90.8/68.0**	79.9/61.1	87.3/65.5	**89.2/67.8**
	OPR	75.8/55.2	74.1/54.7	78.7/58.4	84.2/61.5	89.9/65.2	86.7/61.9	**90.7/67.3**	85.2/62.7	**91.3/67.0**	**93.1/69.1**
	OV	69.7/50.9	53.6/42.3	76.5/55.2	73.4/56.6	**89.5**/64.8	76.4/55.3	**91.3/66.0**	72.3/56.0	87.8/**66.0**	**91.4/66.9**
	SV	75.3/54.1	72.6/55.0	77.4/57.6	78.6/57.2	**88.1**/65.4	86.2/60.3	87.9/**66.6**	80.3/58.2	**90.4**/**66.4**	**91.5/68.7**
	All	80.9/60.0	76.9/58.8	82.4/62.1	83.8/62.3	90.3/67.3	87.3/62.8	**91.0/69.1**	85.6/68.2	**91.8/68.2**	**92.9/70.1**

TABLE VI
TRACKING RESULTS WITH DIFFERENT FEATURES ON OTB100.

Feature	OP	DP	FPS
Hand-crafted	78.1%	83.2%	**25.2**
Deep	71.5%	76.5%	8.0
Direct Fusion	85.8%	87.2%	7.1
Coarse-to-Fine Fusion	**88.6%**	**92.9%**	6.8

Therefore, our LSDCF can shrink the elements even within the central region.

For the low-rank constraint, we collect the filters for each frame, concatenate them together in a matrix, and calculate the rank. To guarantee the quality of the selected filters, we only consider some simple sequences where all the trackers successfully track all the frames, *i.e.*, the filters are able effectively to distinguish the target from its surroundings. The results are presented in Table I, which shows that our simplified term, *i.e.*, $\lambda_2 \|\mathbf{w} - \boldsymbol{\mu}_{t-1}\|_2^2$ in Equ. (7) can achieve the low-rank property by only considering the weighted mean filter. Note, C-COT and ECO also exhibit a low-rank property, but this is achieved by modelling historical appearance variations more comprehensively at the expense of increased computational complexity and storage.

3) Fusion Strategy: To achieve a fair comparison of mathematical formulations, we also compare our method with different features and fusion strategies on OTB100. As shown in Table VI, our coarse-to-fine fusion strategy improves the OP/DP from 85.8%/87.2% to 88.6%/92.9%, compared to a direct fusion strategy that just simply sums all the response maps obtained by hand-crated and deep features.

C. Comparison with the State-of-the-art Algorithms

1) Overall Performance: We report the evaluation results of our LSDCF and a number of state-of-the-art trackers on OTB100, TC128 and UAV123 in Fig. 5, using the precision and success plots. Overall, the proposed LSDCF outperforms all the state-of-the-art trackers in DP and AUC on these three datasets. Compared with the second best, LSDCF achieves improvements of 1.9%/1.0%, 2.3%/1.3% and 2.5%/2.7% in terms of DP/AUC on OTB100, TC128 and UAV123, respectively.

Table II presents OP, CLE and FPS of our LSDCF and 17 state-of-the-art trackers on OTB2013, OTB50 and OTB100. On OTB100, LSDCF achieves 88.6% in OP and 9.0 *pixels* in CLE. Compared with the recent VITAL and

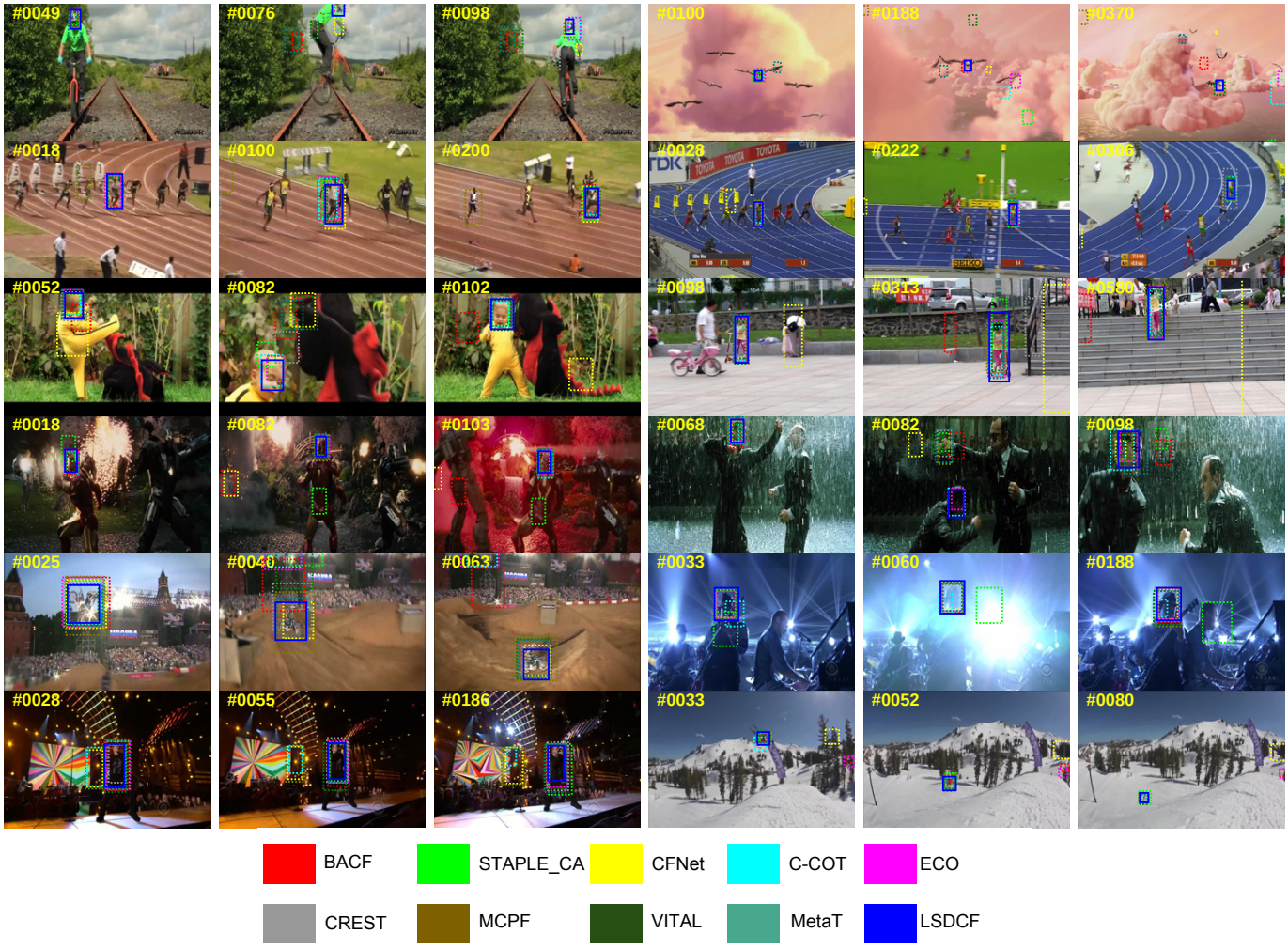


Fig. 6. A qualitative comparison of our LSDCF with state-of-the-art trackers on challenging sequences of OTB100 [8] (left column: *Biker*, *Bolt2*, *Dragonbaby*, *Ironman*, *MotorRolling*, and *Singer2*; right column: *Bird1*, *Bolt1*, *Girl2*, *Matrix*, *Shaking*, and *Skiing*).

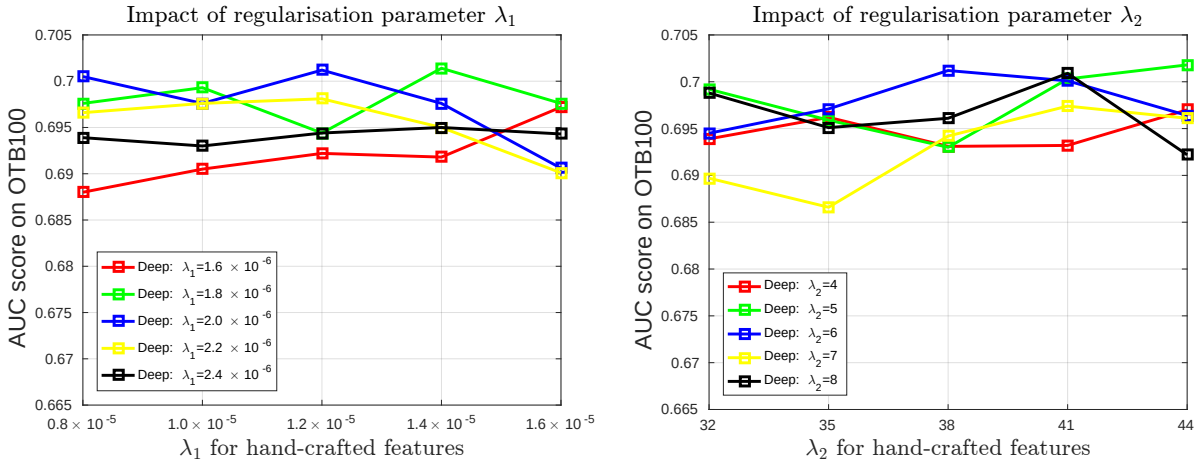


Fig. 7. The experimental results obtained by LSDCF on OTB100, parameterised by different values of λ_1 and λ_2.

MetaT trackers based on deep neural networks, our performance gains are 0.1%/0.9 *pixel* and 8.8%/5.2 *pixels* in terms of OP and CLE, respectively. On OTB2013, LSDCF performs better than VITAL (by 2.5%) in terms of OP but with a lower CLE (by 0.4 *pixel*). In addition, on OTB50, our tracker outperforms recent trackers, *i.e.*, CSRDCF (by16.7%/20.1 *pixels*), STAPLE_CA (by15.0%/26.1 *pixels*), C-COT (by2.2%/2.1 *pixels*), BACF

(by10.0%/18.0 *pixels*), ECO (by 2.1%/3.0 *pixels*) and VITAL (by 1.8%/2.3 *pixels*) in terms of OP/CLE, respectively.

Table III and Table IV show the evaluation results on VOT2016 and VOT2018. According to the tables, our LSDCF method achieves the best tracking results in all metrics on VOT2016. Compared with C-COT, LSDCF realises improvements from 0.331/0.541/0.24 to 0.407/0.587/0.18 in terms of EAO, accuracy and robustness. The tracking results of LSDCF on VOT2018 are also among the top three, compared with more recent published tracking algorithms. Therefore, the proposed LSDCF tracking method achieves outstanding performance against the state-of-the-art trackers with a favourable speed.

2) Performance in Different Attributes: We compiled the tracking results annotated by 11 attributes *i.e.*, background clutter (BC), deformation (DEF), fast motion (FM), in-plane rotation (IPR), low resolution (LR),illumination variation (IV), out-of-plane rotation (OPR), motion blur (MB), occlusion (OCC), out-of-view (OV), and scale variation (SV), on OTB100 [8] in Table V. Our LSDCF outperforms all the other trackers in 7 attributes, *i.e.*, DEF, FM, IPR, LR, OPR, OV, and SV in terms of both DP and AUC. Our low-rank and sparse DCF formulation enables adaptive temporal-spatial-channel layout recognition, focusing on the relevant discriminative features. The results of LSDCF in the other 4 attributes are still among the top 3, demonstrating the effectiveness and robustness of our method. In particular, the performance of LSDCF exhibits significant performance boosting (2.3%/4.8%, 1.8%/1.8% and 1.1%/2.1% in terms of DP/AUC as compared with the second best one in the attributes of LR, OPR and SV, respectively.

3) Comparison in Qualitative Performance: Fig. 6 presents the qualitative results of the state-of-the-art methods, *i.e.*, BACF, STAPLE_CA, CFNet, C-COT, ECO, CREST, MCPF, VITAL, MetaT as well as our LSDCF, on some challenging video sequences. The difficulties are posed by rapid variations in the appearance of targets and surroundings. Our LSDCF performs well on these challenges, benefiting from learning in the framework of the low-rank and sparse DCF formulation and of the coarse-to-fine tracking strategy. Sequences with deformations (*Bolt*, *Dragonbaby*, and *Skiing*) and out of view (*Biker* and *Bird1*) can successfully be tracked by our methods without any failures. Videos with occlusions (*Dragonbaby*, *Girl2*, and *Bird1*) also benefit from our tracking strategy of unveiling the complementary characteristics of hand-crafted and deep features. Specifically, LSDCF is expert in solving in-plane and out-of-plane rotations (*MotorRolling*, *Dragonbaby*, and *Skiing*), because the proposed adaptive low-rank and sparse regularisation approach provide enhanced discrimination by highlighting specific appearance information from the central region and surroundings.

D. Sensitivity Analysis

In this part, we provide a sensitivity analysis of our proposed LSDCF method to regularisation parameters, λ_1 and λ_2, in order to assess the stability of the sparsity and low-rank components.

As shown in Fig. 7, the maximum performance gap between the filters designed with different λ_1 or λ_2 for both hand-crafted and deep features is less than 1.5% . The tracking results on OTB100 vary smoothly with respect to λ_1 or λ_2, demonstrating that our method achieves stable performance with the proposed low-rank and sparse DCF formulation, encouraging the learnt filter to be instantiated in a low-dimensional manifold space to to cope with diversity and to achieve considerable generalisation.

V. Conclusion

We proposed an effective appearance model based on discriminative correlation filters for visual object tracking in video sequences. By reformulating the appearance model learning so as to incorporate low-rank and sparse regularisation, we derived adaptive temporal-spatial-channel filters on a low dimensional manifold with enhanced interpretability. Hand-crafted and deep features are combined, in a complementary way, using an innovative coarse-to-fine tracking framework. The extensive experimental results on tracking benchmark datasets demonstrate the effectiveness and robustness of our method, compared to the state-of-the-art trackers.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (61672265, U1836218, 61876072, 61902153), the 111 Project of Ministry of Education of China (B12018), and the EPSRC Programme Grant (FACER2VM) EP/N007743/1, EPSRC/dstl/MURI project EP/R018456/1.

References

[1] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[2] S. Avidan, "Support vector tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 8, pp. 1064–1072, 2004.

[3] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 174–188, 2002.

[4] 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.

[5] J. F. Henriques, C. Rui, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[6] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, "Hedged deep tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4303–4311.

[7] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2411–2418.

[8] 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.

[9] 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.

[10] 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.

[11] M. Kristan, J. Matas, A. Leonardis, T. Vojíř, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2137–2155, 2016.

[12] M. Kristan, R. Pflugfelder, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, and G. Nebehay, "The visual object tracking vot2015 challenge results," in *IEEE International Conference on Computer Vision Workshop*, 2015, pp. 564–586.

[13] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin Zajc, T. Vojir, G. Häger, A. Lukežič, and G. Fernandez, "The visual object tracking vot2016 challenge results," Springer, Oct 2016. [Online]. Available: http://www.springer.com/gp/book/9783319488806

[14] R. M. Gray, "Toeplitz and circulant matrices : a review," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[15] J. Henriques, o F, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *European Conference on Computer Vision*, 2012, pp. 702–715.

[16] M. Danelljan, F. S. Khan, M. Felsberg, and J. V. D. Weijer, "Adaptive color attributes for real-time visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1090–1097.

[17] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *IEEE International Conference on Computer Vision*, 2015, pp. 4310–4318.

[18] D. Martin, R. Andreas, K. Fahad, and F. Michael, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *European Conference on Computer Vision*, 2016, pp. 472–488.

[19] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 38, 2016, pp. 1401–1409.

[20] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3623–3632.

[21] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, 2017.

[22] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 1, pp. 20–33, 2017.

[23] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2017.

[24] X. Lu, C. Ma, B. Ni, X. Yang, I. Reid, and M.-H. Yang, "Deep regression tracking with shrinkage loss," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 353–369.

[25] X. Dong, J. Shen, W. Wang, Y. Liu, L. Shao, and F. Porikli, "Hyperparameter optimization for tracking with continuous deep q-learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 518–527.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[29] 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.

[30] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, and R. Pflugfelder, "The visual object tracking vot2017 challenge results," 2017.

[31] M. Kristan, A. Leonardis, J. Matas, and M. Felsberg, "The sixth visual object tracking vot2018 challenge results," in *European Conference on Computer Vision workshops*, vol. 3, no. 5, 2018, p. 8.

[32] 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.

[33] A. Li, M. Lin, Y. Wu, M. H. Yang, and S. Yan, "Nus-pro: A new visual tracking challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 335–349, 2016.

[34] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[35] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 142–149.

[36] X. Mei and H. Ling, "Robust visual tracking using l1 minimization," in *IEEE International Conference on Computer Vision*, 2009, pp. 1436–1443.

[37] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2017, p. 3.

[38] R. G. Baraniuk, "Compressive sensing [lecture notes]," *IEEE signal processing magazine*, vol. 24, no. 4, pp. 118–121, 2007.

[39] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via structured multi-task sparse learning," *International Journal of Computer Vision*, vol. 101, no. 2, pp. 367–383, 2013.

[40] X. Jia, "Visual tracking via adaptive structural local sparse appearance model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1822–1829.

[41] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Low-rank sparse learning for robust visual tracking," in *European conference on computer vision*. Springer, 2012, pp. 470–484.

[42] 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.

[43] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *British Machine Vision Conference 2006, Edinburgh, Uk, September*, 2006, pp. 47–56.

[44] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *European conference on computer vision*. Springer, 2008, pp. 234–247.

[45] B. Babenko, M. H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.

[46] S. Hare, A. Saffari, and P. H. Torr, "Struck: Structured output tracking with kernels," in *Proceedings of the 2011 International Conference on Computer Vision*. IEEE Computer Society, 2011, pp. 263–270.

[47] Bertinetto, Luca and Valmadre, Jack and Henriques, Joao F and Vedaldi, Andrea and Torr, Philip HS, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*. Springer, 2016, pp. 850–865.

[48] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 5000–5008.

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

[50] 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.

[51] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1430–1438.

[52] P. Zhang, S. Yu, J. Xu, X. You, X. Jiang, X.-Y. Jing, and D. Tao, "Robust visual tracking using multi-frame multi-feature joint modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[53] S. Wang, D. Wang, and H. Lu, "Tracking with static and dynamic structured correlation filters," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2861–2869, 2018.

[54] T. Xu, X.-J. Wu, and J. Kittler, "Non-negative subspace representation learning scheme for correlation filter based tracking," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1888–1893.

[55] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6931–6939.

[56] K. Z, M. K, and M. J, "Tracking-learning-detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[57] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 798–805.

[58] Q. Liu, J. Fan, H. Song, W. Chen, and K. Zhang, "Visual tracking via nonlocal similarity learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2826–2835, 2018.

[59] T. Zhang, S. Liu, C. Xu, S. Yan, B. Ghanem, N. Ahuja, and M.-H. Yang, "Structural sparse tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 150–158.

[60] S. Liu, T. Zhang, X. Cao, and C. Xu, "Structural correlation filter for robust visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4312–4320.

[61] Y. Li, J. Zhu, and S. C. Hoi, "Reliable patch trackers: Robust visual tracking by exploiting reliable patches," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 353–361.

[62] R. Yao, Q. Shi, C. Shen, Y. Zhang, and A. van den Hengel, "Part-based robust tracking using online latent structured learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 6, pp. 1235–1248, 2017.

[63] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Joint group feature selection and discriminative filter learning for robust visual object tracking," *arXiv preprint arXiv:1907.13242*, 2019.

[64] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1396–1404.

[65] H. Kiani Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *IEEE International Conference on Computer Vision*, 2017.

[66] D. Wang, H. Lu, and M.-H. Yang, "Robust visual tracking via least soft-threshold squares," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1709–1721, 2016.

[67] R. J. Tibshirani, "Regression shrinkage and selection via the lasso. j r stat soc b," *Journal of the Royal Statistical Society*, vol. 58, pp. 267–288, 1996.

[68] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[69] J. V. D. Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1512–23, 2009.

[70] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[71] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Eprint Arxiv*, vol. 9, 2010.

[72] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[73] K. B. Petersen, M. S. Pedersen *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, no. 15, p. 510, 2008.

[74] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 689–692.

[75] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R. Lau, and M.-H. Yang, "Vital: Visual tracking via adversarial learning," *arXiv preprint arXiv:1804.04273*, 2018.

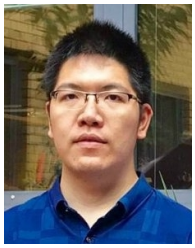
[76] E. Park and A. C. Berg, "Meta-tracker: Fast and robust online adaptation for visual object trackers," *arXiv preprint arXiv:1801.03049*, 2018.

[77] J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, J. Y. Choi *et al.*, "Attentional correlation filter network for adaptive visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2017.

[78] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4847–4856.

[79] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *European Conference on Computer Vision Workshops*. Springer, 2014, pp. 254–265.

[80] 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.



Tianyang Xu received the B.Sc. degree in electronic science and engineering from Nanjing University, Nanjing, China, in 2011. He is a PhD student at the School of Internet of Things Engineering, Jiangnan University, Wuxi, China. He is currently a visiting PhD student 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.

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

He has published more than 30 scientific papers in top-ranking conferences and journals, including IEEE Conference on Computer Vision and Pattern Recognition, IEEE International Conference on Computer Vision, IEEE Transactions on Image Processing, IEEE Transactions on Cybernetics, IEEE Transactions on Information Forensics and Security, Pattern Recognition, Information Sciences etc. He has received the 2017 European Biometrics Industry Award from the European Association for Biometrics (EAB).

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 Jiangnan University, Wuxi, China. His current research interests include pattern recognition, computer vision, fuzzy systems, neural networks and intelligent systems.

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 62,000 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.