IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 12, 2022, accepted February 25, 2022, date of publication February 28, 2022, date of current version March 9, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3155660

A Novel Algorithm Based on a Common Subspace Fusion for Visual Object Tracking

SAJID JAVED[1,2], **ARIF MAHMOOD**[3], **IHSAN ULLAH**[4], **THIERRY BOUWMANS**[5], **(Member, IEEE)**, **MAJID KHONJI**[1,2], **(Member, IEEE)**, **JORGE MANUEL MIRANDA DIAS**[1,2], **AND NAOUFEL WERGHI**[1,2], **(Senior Member, IEEE)**

[1]Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[2]Khalifa University Centre for Autonomous Robotics Systems (KUCARS), Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[3]Department of Computer Science, Information Technology University, Lahore 25000, Pakistan
[4]School of Computer Science, National University of Ireland Galway, Galway, H91 TK33 Ireland
[5]Laboratoire MIA, La Rochelle Université, 17031 La Rochelle, France

Corresponding author: Sajid Javed (sajid.javed@ku.ac.ae)

This work was supported by the Khalifa University of Science and Technology under Award RC1-2018-KUCARS. The work of Sajid Javed was supported by the Faculty Start Up (FSU-2022-003) Grants under Project 000628-00001.

ABSTRACT Recent methods for visual tracking exploit a multitude of information obtained from combinations of handcrafted and/or deep features. However, the response maps derived from these feature combinations are often fused using simple strategies such as winner-takes-all or weighted sum approaches. Although some efficient fusion methods have also been proposed, these methods still do not leverage the individual strengths of the different features being fused. In the current work, we propose a novel information fusion strategy comprising a common low-rank subspace for the fusion of different types of features and tracker responses. Firstly, we interpret the response maps as smoothly varying functions which can be efficiently represented using individual low-rank matrices, thus removing high frequency noise and sparse artifacts. Secondly, we estimate a common low-rank subspace which is constrained to remain close to each individual low-rank subspace resulting in an efficient fusion strategy. The proposed algorithm achieves good performance by integrating the information contained in heterogeneous feature types. We demonstrate the efficiency of our algorithm using several combinations of features as well as correlation filter and end-to-end deep trackers. The proposed common subspace fusion algorithm is generic and can be used to efficiently fuse the response maps of varying types of feature representations as well as trackers. Extensive experiments on several tracking benchmarks including OTB100, TC128, VOT-ST 2018, VOT-LT 2018, UAV123, GOT-10K and LaSoT have demonstrated significant performance improvements compared to many SOTA tracking methods.

INDEX TERMS Visual object tracking, features fusion, correlation filters, deep features.

I. INTRODUCTION

Visual Object Tracking (VOT) is one of the most fundamental tasks in computer vision having a wide range of applications across several domains [1], [2] for example autonomous driving [3], anomaly detection [4], augmented reality [5], action recognition [6], surveillance, and security [7]. Numerous research directions have been investigated in recent years for VOT [8]–[24]. Despite a lot of research focus, VOT in challenging environments is still an open problem which needs to be further investigated [25]–[30]. Among the most investigated tracking approaches, Correlation Filters (CFs) have attained significant attention because of their impressive performance in terms of speed and accuracy [9]–[12], [31]–[40]. In most of these methods, a correlation filter is trained over a region of interest in the current frame which is then employed to track the target object in the subsequent frames by maximizing the filter response [18], [41], [42]. More recently end-to-end deep learning-based trackers have also been proposed which have achieved excellent performance [43]–[45]. In many cases, classical object detectors such as Faster R-CNN have also been adapted for

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

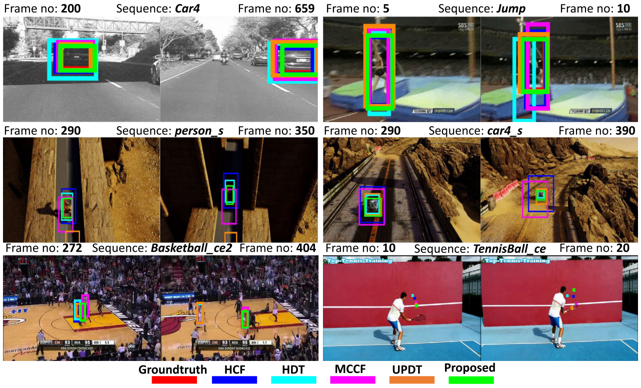
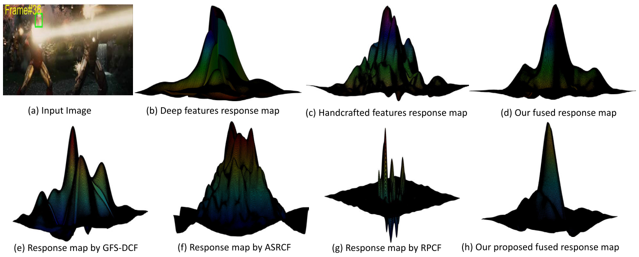


FIGURE 1. Many existing tracking methods including HCF [9], HDT [10], MCCF [12], and UPDT [11] are not able to effectively handle VOT in the presence of challenging scenarios. These sequences suffer from illumination variation, fast motion, motion blur, and occlusion challenges and selected from OTB100 [26] and UAV123 [28] datasets. In contrast to the compared methods, the proposed CSF tracker has better handled these sequences.

tracking-by-detection tasks [13], [14]. The performance of CF-based trackers is further enhanced through scale invariance [42], target re-detection [46], deep end-to-end training [43], local and global filter ensembles [12], and the combination of deep CNN and handcrafted features [11], [47].

Most existing CNN-based methods only use features from later layers to represent target objects, because these features capture rich category-level semantic information. However, spatial details captured by earlier layers are also important for accurately localizing a target [9]. Although the features from these earlier layers are relatively less discriminative than those of later layers, often leading to failure in more challenging tracking scenarios. Consequently, many trackers complement deep representations with shallow activations or handcrafted features for improved localization [41], [42], [48]. This raises the question of how to optimally fuse the fundamentally different properties of shallow and deep features in order to achieve both accuracy and robustness [11]. For optimal tracking performance, it is imperative that handcrafted features be combined with deep features from different CNN layers to best discriminate between target object and background clutter. Fusion of different feature representations has been shown to improve VOT performance [9]–[12]. For instance, Ma *et al.*, aggregate response maps extracted from earlier and later CNN layers by manually assigning a relative weight to each [9]. Qi *et al.* proposed fusion of response maps from six CNN features using a hedging method [10]. Wang *et al.* proposed feature-level and decision-level strategies to fuse multi-expert response maps [12]. Bhat *et al.* recently proposed a Unveiling the Power of Deep Tracking (UPDT) tracker in which the relative weights of features are learnt from training samples [11]. Although feature-level strategies have demonstrated competitive performance for VOT, the initial weights of deeper layers tend to be higher than those of shallow layers, due to their ability to encode more semantic information. However, it is observed that shallow layers can improve localization performance [9], [11], suggesting that in some tracking scenarios shallow layers should be given significant weightage. This has been addressed by decision-level tracker, however the early feature-level fusion strategy is an important factor to consider [12]. Figure 1 presents a challenging situation in which the aforementioned trackers have faced many difficulties to track the target objects.



(a) Input Image (b) Deep features response map (c) Handcrafted features response map (d) Our fused response map (e) Response map by GFS-DCF (f) Response map by ASRCF (g) Response map by RPCF (h) Our proposed fused response map

FIGURE 2. (a) Shows the search regions of sequence *Ironman*. (b) The response map of deep features using KCF as a baseline tracker [31]. (c) Fused response map of handcrafted features including HOG, IC, and CN. (d) Fusion of the response maps shown in (b) and (c) using our proposed CSF algorithm. (e)-(g) Show the response maps of SOTA CFs-trackers, and (h) shows the fusion of response maps shown in (e)-(g) using our proposed CSF algorithm.

In the current work, we propose to learn a common subspace-based response map which compliments the information captured by handcrafted as well as deep features. For each response map, we estimate its low-rank representation using non-negative matrix factorization [49]. Then we compute a common low-rank representation across all these response maps which is constrained to remain close to each individual low-rank representation [50]. Thus a consensus is achieved by those response maps which correctly estimate the target position while the incorrect ones do not get accumulated resulting in a more robust VOT. We observe the effectiveness of proposed algorithm by comparing its response map with various State-Of-The-Art (SOTA) trackers as shown in Figure 2. In the first case, Figure 2 (d) shows the fused response map of the proposed CSF tracker using KCF as a baseline tracker on deep features (Figure 2 (c)) and on handcrafted features (Figure 2 (d)). In the second case, Figure 2 (h) shows the fusion by the proposed CSF tracker over three existing SOTA correlation filters-based trackers including GFS-DCF [41], ASRCF [42], and RPCF [48]. In both cases, the fused response map shows a higher signal peak and suppressed noisy peaks.

The proposed tracker, which we name Common Subspace Fusion (CSF), is evaluated on seven tracking benchmark datasets and compared with the many SOTA trackers. Our experiments demonstrate a significant performance improvement in terms of both speed and accuracy. Specifically, our tracker demonstrates a 7.0% improvement in terms of Expected Average Overlap (EAO) as compared to baseline GFS-DCF tracker [41] and an 3.0% improvement as compared to PrDiMP tracker [16] on VOT2018 dataset [25]. Further experiments on GOT-10k [51], OTB100 [26], UAV123 [28] and LaSoT [30] datasets have demonstrated significant improvement over existing SOTA trackers. The main contributions of the current work are as follows:

- A novel common subspace fusion algorithm is proposed based on low-rank response map representation of various types of features and trackers. Using the individual low-rank representation response map, a common subspace-based representation is estimated which is constrained to be close to each individual representation on the Grassmannian manifold.
- The proposed fusion scheme is employed on correlation filter-based trackers using different features resulting in significant performance improvement in all cases. It is also employed to fuse the predicted response maps of deep trackers which again results in significant performance improvement. Rigorous evaluations are performed on two long-term and six short-term tracking datasets. The proposed CSF tracker consistently demonstrated improved performance.

The rest of this paper is organized as follows. Section II summarizes related work. Section III presents the proposed methodology in detail. Section IV describes our experimental evaluation and results and Section V presents our conclusions and future directions.

II. LITERATURE REVIEW

In the past decade, a number of studies have demonstrated improved performance for the task of VOT [9]–[20], [52]. Since the current work is focused on the fusion of various types of features and trackers, we particularly review those studies which present some type of fusion scheme.

Many researchers have aimed to tap the complementary information contained in various types of handcrafted and deep features by using different fusion strategies. These schemes may be categorized into two groups: feature-level and decision-level fusion. Feature-level fusion is an intermediate level fusion in which each feature representation is used to obtain a probability map of the target location. These probability maps, also known as response maps, are then fused using different strategies such as pre-defined or learned weights. This fusion strategy assigns relative weights to different types of features based on semantic information, therefore semantically rich high-level features get higher weights compared to their shallow counterparts. It has been observed that in many tracking scenarios, shallow features are more effective than deep features, resulting in performance degradation of feature fusion strategies that prioritize deep features. For instance, Ma *et al.* trained correlation filters on each feature layer of VGG-19 [9]. The fused responses were estimated by aggregating all feature maps using a manually hard-coded weighting scheme. Qi *et al.* proposed a fusion method for hedging correlation filter responses based on relative hard-coded weights into a single response map for target detection [10]. These manually assigned hard-coded weighting schemes may not be optimal in all tracking scenarios. To address this problem, Bhat *et al.* proposed to learn the weights of the individual feature representation and demonstrated improved VOT performance as compared to the aforementioned fusion techniques [11]. UPDT learned optimal fusion hyper-parameters on the OTB100 dataset [26], which were subsequently applied to other tracking datasets, albeit with no guarantee of the effectiveness of these learned parameters across different tracking challenges. It is observed that when weights are learned, higher priority is still given to deep features over shallow or handcrafted features.

Decision-level fusion is exploited by the MCCF tracker in which a result is selected from multiple proposals based on the agreement of multiple feature combinations as well as temporal consistency [12]. While it has been shown to improve performance in some scenarios, the significance of decision-level fusion strongly depends on the design of the baseline feature combinations. Decision-level fusion is again strongly dependent on a feature-level fusion in which semantic information is given high significance. Decision-level fusion is also prone to errors in scenarios where multiple feature combinations contain similar errors. In many tracking scenarios, semantic information may cause errors that could be overcome by prioritizing low-level information. Some studies are also reported on other imaging modalities such as thermal infrared for robust object tracking [53]–[55].

In the current work, we address this shortcoming by using a feature-level fusion strategy based on the common subspace spanned by individual response maps. In contrast to the aforementioned fusion strategies, we consider each feature representation to be equally significant so that a broader range of tracking scenarios can be effectively handled compare to the prior unequal weighting schemes. We learn a common subspace across all feature representations which is constrained to be close to each low-rank representation on the Grassmannian manifold. Our proposed fusion scheme is generic, allowing us to demonstrate its efficacy by plugging it into many recent SOTA trackers resulting in significant performance improvement.

III. PROPOSED COMMON SUBSPACE FUSION ALGORITHM

In our proposed Common Subspace Fusion (CSF) algorithm, each response map is considered equally important, therefore we do not compute any weights for shallow or deep feature maps. Thus we address the problem of tracking errors caused by incorrect semantic information being given too much importance. Compared to the aforementioned fusion schemes, which improve performance by using weak classifiers, our proposed fusion strategy is more generic and improves performance beyond current SOTA trackers. The system diagram of our proposed CSF tracker is shown in Figure 3. For each response map estimated by a set of SOTA trackers, a low-rank representation is computed. Then using multiple low-rank representations, our aim is to compute a common subspace representation resulting in fusion over multiple response maps.

A. MATHEMATICAL FORMULATION

An ideal tracking response map $\mathbf{R}_k \in \mathbb{R}^{m\times m}$ should be a smoothly varying continuous function, however when working with real-world data, it may contain high frequency artifacts, where $m \times m$ is the size of the response map and k denotes feature representation. We therefore propose to compute a low-rank representation $\mathbf{L}_k \in \mathbb{R}^{m\times c}$ of $\mathbf{R}_k$ where $c < m$ and maximum rank of $\mathbf{L}_k \leq c$. For this purpose, we convert the response map $\mathbf{R}_k$ into an affinity matrix $\mathbf{S}_k = \mathbf{R}_k\mathbf{R}_k^\top \in \mathbb{R}^{m\times m}$ which is symmetric and positive semi-definite and may be factorized into a low-rank sparse matrix. $\mathbf{S}_k$ contains the structure of the corresponding response map $\mathbf{R}_k$ such that one cluster in $\mathbf{S}_k$ belongs to the target region while the remaining clusters correspond to non-target region in the search space.

Non-negative Matrix Factorization (NMF) has been widely employed for the estimation of low-rank approximation [56], [57]. NMF factorizes an input data matrix $\mathbf{S}_k$ into two non-negative matrices $\mathbf{L}_k$ and $\mathbf{G}_k$, i.e., $\mathbf{S}_k \approx \mathbf{L}_k\mathbf{G}_k^\top$. The rank of both non-negative matrices $\mathbf{L}_k$ and $\mathbf{G}_k$ is significantly lower than $\mathbf{S}_k$. For the purpose of uniqueness and clustering interpretation, $\mathbf{G}_k$ is enforced to be orthogonal $\mathbf{G}_k^\top\mathbf{G}_k = \mathbf{I}$. We consider to enforce orthogonality constraints on both non-negative matrices $\mathbf{L}_k$ and $\mathbf{G}_k$, so that $\mathbf{L}_k$ can be considered as cluster indicator matrix for rows clustering and $\mathbf{G}_k$ as the cluster indicator matrix for columns clustering. Such a configuration assists us to identify the target region as an intersection of rows and columns corresponding to the target clusters. The objective function for such decomposition is formulated as follows:

$$\min_{\mathbf{L}_k\leq 0,\mathbf{G}_k\leq 0} ||\mathbf{S}_k - \mathbf{L}_k\mathbf{G}_k^\top||^2, \text{ s.t. } \mathbf{L}_k^\top\mathbf{L}_k = \mathbf{I}, \mathbf{G}_k^\top\mathbf{G}_k = \mathbf{I} \quad (1)$$

However, this double orthogonality is very restrictive and it gives a rather poor matrix low-rank approximation. One needs an extra factor $\mathbf{B}_k$ to absorb the different scales of $\mathbf{S}_k$, $\mathbf{L}_k$, and $\mathbf{G}_k$, i.e., $\mathbf{S}_k \approx \mathbf{L}_k\mathbf{B}_k\mathbf{G}_k^\top$. In case of symmetric input matrix $\mathbf{S}_k$, $\mathbf{S}_k = \mathbf{S}_k^\top$, the non-negative matrices become same i.e., $\mathbf{L}_k = \mathbf{G}_k$. Using Symmetric Non-negative Matrix Tri-Factorization (SNMTF), we factorize each $\mathbf{S}_k$ as $\mathbf{S}_k \approx \mathbf{L}_k\mathbf{B}_k\mathbf{L}_k^\top$ by solving the following objective function [49]:

$$d_k(\mathbf{S}_k; \{\mathbf{L}_k, \mathbf{B}_k\}) = \min_{\{\mathbf{L}_k,\mathbf{B}_k\}\geq 0} ||\mathbf{L}_k\mathbf{B}_k\mathbf{L}_k^\top - \mathbf{S}_k||_F^2, \quad (2)$$

where $||\cdot||_F$ is the Frobenius norm [58] and $\mathbf{B}_k$ is a non-negative auxiliary matrix. The matrix $\mathbf{L}_k$ contains the feature specific response map structure such that one particular cluster corresponds to the target region while the remaining clusters belong to non-target regions.

In order to obtain a common subspace-based tracking response maps structure across all feature representations, we compute a common low-rank representation $\mathbf{M} \in \mathbb{R}^{m\times c}$ which should be close to each individual low-rank response representation $\mathbf{L}_k$. The common representation contains a unified target region cluster over all feature response maps such that the individual target region gets superimposed and resulting in an amplified target response. Matrix $\mathbf{M}$ can be computed using Eq. (2), in which $\mathbf{M}$ replaces $\mathbf{L}_k$, and minimizing the objective function across all k feature representations:

$$d_k(\mathbf{S}_k; \{\mathbf{M}, \mathbf{B}_k\}) = \min_{\{\mathbf{M},\mathbf{B}_k\}\geq 0} ||\mathbf{M}\mathbf{B}_k\mathbf{M}^\top - \mathbf{S}_k||_F^2, \quad (3)$$

If each minimization problem is solved independently, matrix $\mathbf{M}$ can be further from some low-rank representations $\mathbf{L}_k$ than the others. A set of c dimensional linear subspaces of $\mathbb{R}^m$ can be considered a Grassmann manifold $G(c, m)$, such that each point in this manifold corresponds to a unique subspace. Each subspace can be represented using its basis vectors as an orthonormal matrix whose columns span the corresponding c dimensional subspace in $\mathbb{R}^m$. In order to ensure that $\mathbf{M}$ is close to the majority of $\mathbf{L}_k$, it is enforced that the subspace spanned by $\mathbf{M}$ is close to the subspace spanned by each $\mathbf{L}_k$ on a Grassmann manifold [50], [59]. Each $\mathbf{L}_k$ spans a corresponding c' dimensional subspace where $c' \leq c \leq m$ and is mapped to a unique point on the Grassmann manifold $\mathbf{G}(c', m)$ defined as a set of c' dimensional linear subspaces in $\mathbb{R}^m$.

The geodesic distance between two subspaces $\mathbf{M}$ and $\mathbf{L}_k$ on the Grassmann manifold can be computed by projection distance as follows [50], [59]:

$$\begin{aligned} d_k(\mathbf{M}; \mathbf{L}_k) &= \sum_{j=1}^{c'} sin^2\theta_j^k = c' - \sum_{j=1}^{c'} cos^2\theta_j^k \\ &= c' - \text{tr}(\mathbf{M}\mathbf{M}^\top\mathbf{L}_k\mathbf{L}_k^\top), \end{aligned} \quad (4)$$

where $\{\theta_j^k\}_{j=1}^{c'}$ are principal angles between c'-dimensional subspaces spanned by $\mathbf{L}_k$ and $\mathbf{M}$ and tr$(\cdot)$ denotes the trace of a matrix. In order to ensure $\mathbf{M}$ to be close to each $\mathbf{L}_k$, the overall objective function is given by:

$$\min_{\mathbf{M}\geq 0} \Psi = \sum_k (d_k(\mathbf{S}_k; \{\mathbf{M}, \mathbf{B}_k\}) + \gamma d_k(\mathbf{M}; \mathbf{L}_k)) \quad (5)$$

where $\gamma > 0$ is a weighting parameter. The second term is the sum of the projection distances between $\mathbf{M}$ and each $\mathbf{L}_k$. Minimizing this term will ensure that the matrix $\mathbf{M}$ will be close to each individual matrix $\mathbf{L}_k$ on the Grassmann manifold in terms of geodesic distance. In order to minimize Eq. (5), we first formulate the multiplicative update rules to compute $\mathbf{L}_k$ using the SNMTF method [49] and then we jointly optimize our objective function (Eq. (5)) to derive the multiplicative update rules for our common low-rank matrix $\mathbf{M}$.

1) INDIVIDUAL LOW-RANK REPRESENTATION COMPUTATION

Following multiplicative update rules are derived for Eq. 2 using the SNMTF method [49] for estimating $\mathbf{L}_k$ as follows:

$$\begin{aligned} \mathbf{L}_k(i,j) &\leftarrow \frac{[\mathbf{S}_k\mathbf{L}_k\mathbf{B}_k]_{(i,j)}}{[\mathbf{L}_k\mathbf{L}_k^\top\mathbf{S}_k\mathbf{L}_k\mathbf{B}_k]_{(i,j)}}\mathbf{L}_k(i,j), \\ \mathbf{B}_k(i,j) &\leftarrow \frac{[\mathbf{L}_k^\top\mathbf{S}_k\mathbf{L}_k]_{(i,j)}}{[\mathbf{L}_k^\top\mathbf{L}_k\mathbf{B}_k\mathbf{L}_k^\top\mathbf{L}_k]_{(i,j)}}\mathbf{B}_k(i,j), \end{aligned} \quad (6)$$

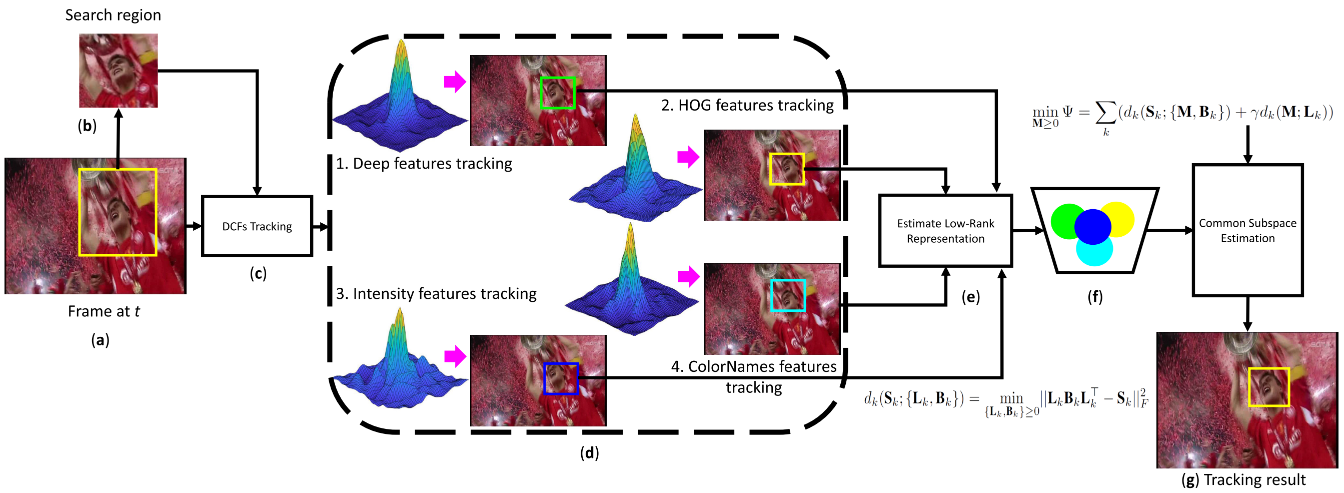

Search region
(b)
Frame at *t*
(a)
DCFs Tracking
(c)
1. Deep features tracking
2. HOG features tracking
3. Intensity features tracking
4. ColorNames features tracking
(d)
Estimate Low-Rank Representation
(e)
(f)
Common Subspace Estimation
$\min_{\mathbf{M}\geq 0} \Psi = \sum_k (d_k(\mathbf{S}_k; \{\mathbf{M}, \mathbf{B}_k\}) + \gamma d_k(\mathbf{M}; \mathbf{L}_k))$
$d_k(\mathbf{S}_k; \{\mathbf{L}_k, \mathbf{B}_k\}) = \min_{\{\mathbf{L}_k, \mathbf{B}_k\}\geq 0} ||\mathbf{L}_k\mathbf{B}_k\mathbf{L}_k^\top - \mathbf{S}_k||_F^2$
(g) Tracking result

FIGURE 3. System diagram of the proposed Common Subspace Fusion (CSF) algorithm. (a) Shows input frame at time *t*. (b) Shows search regions of the input frame. (c) DCFs-based tracker is used to compute the response maps of varying feature representations. (d) Shows the response maps of different types of features using a baseline KCF tracker [31]. (e) Low-rank representation is estimated for each response map. (f) Each color show low-rank representation computed in step (e). (g) Fusion of low-rank representations using our proposed common subspace estimation algorithm and the tracking result.

where $\mathbf{L}_k(i,j)$ is the (i,j)-th element of the low-rank representation $\mathbf{L}_k$. Eq. (6) converges to the optimal solution if $\frac{||\mathbf{L}_k^t - \mathbf{L}_k^{(t-1)}||_F}{||\mathbf{L}_k^{(t-1)}||_F} \leq \zeta$, where ζ is a tolerance factor.

2) COMMON LOW-RANK REPRESENTATION MATRIX COMPUTATION

Following the constrained optimization theory [60] and non-negative matrix factorization [61], we take the derivative of (5) with respect to $\mathbf{M}$ as follows:

$$\nabla_M \Psi = -\sum_k 4\mathbf{S}_k\mathbf{M}\mathbf{B}_k + \sum_k 4\mathbf{M}\mathbf{B}_k\mathbf{M}^\top\mathbf{M}\mathbf{B}_k - 2\gamma \sum_k \mathbf{M}_k\mathbf{M}_k^\top\mathbf{M} \quad (7)$$

The ordinary gradient of the optimization problem Eq. (7) does not represent its steepest direction because the matrix $\mathbf{M}$ spans the Grassmann manifold [62]. However, the steepest direction can be obtained by using the notion of natural gradient [59], [62], [63]. The natural gradient of Ψ on the Grassmann manifold at $\mathbf{M}$ can be written in terms of the ordinary gradient as follows [62], [63]:

$$\widetilde{\nabla}_{\mathbf{M}}\Psi = \nabla_{\mathbf{M}}\Psi - \mathbf{M}\mathbf{M}^\top\nabla_{\mathbf{M}}\Psi, \quad (8)$$

where $\nabla_{\mathbf{M}}\Psi$ is the ordinary gradient given by Eq. (7). Combining Eq. (7) and Eq. (8), we get

$$\widetilde{\nabla}_{\mathbf{M}}\Psi = -\underbrace{(\sum_k \mathbf{S}_k\mathbf{M}\mathbf{B}_k + \frac{\gamma}{2}\sum_k \mathbf{L}_k\mathbf{L}_k^\top\mathbf{M})}_{[\widetilde{\nabla}_{\mathbf{M}}\Psi]^-} + \underbrace{\mathbf{M}\mathbf{M}^\top(\sum_k \mathbf{S}_k\mathbf{M}\mathbf{B}_k + \frac{\gamma}{2}\sum_k \mathbf{L}_k\mathbf{L}_k^\top\mathbf{M})}_{[\widetilde{\nabla}_{\mathbf{M}}\Psi]^+} \quad (9)$$

In order to ensure the positivity constraints on $\mathbf{M}$, the natural gradient is decomposed into two non-negative terms [63], [64] such that: $\widetilde{\nabla}_{\mathbf{M}}\Psi = \widetilde{\nabla}_{\mathbf{M}}\Psi^+ - \widetilde{\nabla}_{\mathbf{M}}\Psi^-$. The two terms are enforced to be positive as follows:

$$(\widetilde{\nabla}_{\mathbf{M}}\Psi^+)_{(i,j)} = \begin{cases} (\widetilde{\nabla}_{\mathbf{M}}\Psi^+)_{(i,j)} & \text{if } (\widetilde{\nabla}_{\mathbf{M}}\Psi^+)_{(i,j)} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

and

$$(\widetilde{\nabla}_{\mathbf{M}}\Psi^-)_{(i,j)} = \begin{cases} -(\widetilde{\nabla}_{\mathbf{M}}\Psi^-)_{(i,j)} & \text{if } (\widetilde{\nabla}_{\mathbf{M}}\Psi^-)_{(i,j)} < 0 \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

Following the KKT condition [60] and preserving the non-negativity of $\mathbf{M}$, we derive the multiplicative update rules for matrix $\mathbf{M}$ using the natural gradient as follows:

$$\mathbf{M}(i,j) \leftarrow \mathbf{M}(i,j)\frac{[\widetilde{\nabla}_{\mathbf{M}}\Psi]^+_{(i,j)}}{[\widetilde{\nabla}_{\mathbf{M}}\Psi]^-_{(i,j)}}, \quad (12)$$

where the non-negative parts of the normal gradient are given in Eq. (8). Substituting Eq. (9) in to Eq. (12), we obtain the multiplicative update rules for the common subspace matrix $\mathbf{M}$ as follows:

$$\mathbf{M}(i,j) \leftarrow \mathbf{M}(i,j)\frac{\sum_k[\mathbf{S}_k\mathbf{M}\mathbf{B}_k + \frac{\gamma}{2}\mathbf{L}_k\mathbf{L}_k^\top\mathbf{M}]_{(i,j)}}{[\mathbf{M}\mathbf{M}^\top\sum_k(\mathbf{S}_k\mathbf{M}\mathbf{B}_k + \frac{\gamma}{2}\mathbf{L}_k\mathbf{L}_k^\top\mathbf{M})]_{(i,j)}} \quad (13)$$

The target detection is then estimated by seeking the maximum value in the common low-rank representation matrix $\mathbf{M}$. Algorithm 1 summarizes the steps of the proposed CSF algorithm.

Algorithm 1: Pseudocode of the Proposed CSF Algorithm

```
Input: Response maps of the target object $\mathbf{R}_k \in \mathbb{R}^{m\times m}$
    using any DCFs-based tracker.
Initialization: $\gamma = 0.8$, $\zeta = 0.3$, $t = 0$.
Compute $\mathbf{S}_k = \mathbf{R}_k\mathbf{R}_k^\top \in \mathbb{R}^{m\times m}$.
While not converged & $t$ <max_iterations=6
1. Compute $\mathbf{L}_k$ using (6).
2. Compute $\mathbf{B}_k$ using (6).
3. Compute $\mathbf{M}$ using (13).
4. Convergence: $\frac{||\mathbf{M}^t - \mathbf{M}^{(t-1)}||_F}{||\mathbf{M}^{(t-1)}||_F} \leq \zeta$
5. t = t+1
Output: M
Find maximum value in common low-rank
representation matrix M for target localization.
```

IV. EXPERIMENTAL EVALUATIONS

The performance of the proposed CSF algorithm is evaluated on seven tracking datasets including OTB100 [26], UAV123 [28], VOT2018 Short Term challenge (VOT2018-ST) [25], TC128 [27], GOT-10K [51], VOT2018 Long Term challenge [25], and LaSoT [30]. These datasets comprise a variety of tracking challenges including occlusion, background clutter, and scale variations to name a few [28]. The description of each dataset is shown in Table 1. Our proposed algorithm is implemented on a PC with an Intel core i7 4GHz, Titan Xp GPU, and 64 GB RAM.

The performance of the proposed CSF tracker is compared with 29 existing SOTA trackers including ASRCF [42], GFS-DCF [41], RPCF [48], ATOM [65], PrDiMP [16], UPDT [11], HDT [10], HCF [9], SRDCF [66], HCFTs [67], STRCF [68], CCOT [69], ECO [70], DeepMCCT [12], DGL [71], TADT [72], DeepSRDCF [73], GradNet [74], MCCT [12], BACF [75], TRACA [76], DeepRSST [77], MUSTER [78], LCT [46], UDT [79], CREST [43], DSLT [80], CFNET [45] and GCT [81].

The tracking performance is evaluated using two popular measures known as precision and success rates [26] for OTB100, TC128, UAV123, and LaSoT datasets. The precision rate is defined as the percentage of frames with Euclidean distance between the predicted and ground truth target location less than 20 pixels threshold [26]. The success rate is defined as the percentage of frames with overlap ratio $\frac{b_1 \cap b_2}{b_1 \cup b_2} > 0.5$ [26], where b_1 and b_2 are the predicted and the ground truth bounding boxes, respectively. By varying the threshold from 0 to 1, the success plots are generated and the area under the curve is estimated. Moreover, following the protocols defined in VOT2017/VOT2018 [25], we used three primary measures including Expected Average Overlap (EAO), Robustness (R), and Accuracy (A) to compare the performance of different trackers on VOT2018-ST dataset. The EAO estimates the average overlap a tracker is expected to obtain on a large set of short-length sequences with the same visual properties as a given dataset. The robustness measures the number of times a tracker fails (loss the target) during tracking while accuracy is the average overlap between the ground truth and estimated bounding box during the successful tracking periods.

The proposed CSF algorithm is tested in two different configurations: fusion using varying types of features for the same tracker (`Config-1`) and integrating multiple deep trackers (`Config-2`). In `Config-1`, we fuse the feature responses obtained by varying types of features while using the same tracker as a baseline.

In `Config-2` we fuse the responses of multiple trackers. Figures 2 (e)-(g) show the response maps of three SOTA CF based trackers. The fusion map is smoothed and has a high signal to noise ratio. The objective is to complement the information captured by different trackers and to analyse the capability of our proposed CSF algorithm to fuse multiple deep trackers into a unified framework.

The performance of the two configurations is evaluated using the VOT2018-ST challenge dataset in terms of Expected Average Overlap (EAO), Robustness (R), and Accuracy (A) using the protocols provided by the authors [25]. Each of these experiments is discussed in detail in the following subsections.

A. FEATURE FUSION (`CONFIG-1`)

Feature fusion is performed using the proposed CSF algorithm with three recent SOTA trackers as baselines: GFS-DCF, ASRCF, and RPCF. We use the same feature set for each tracker including HOG, Intensity Channel (IC), Color Names (CN), and deep features extracted from the 4*th* block of ResNet50. The performance comparison on the VOT2018 dataset is shown in Table 2, demonstrating that the proposed CSF fusion algorithm improves the performance of each baseline tracker by a significant margin. In this experiment, the baseline trackers are used as the same as proposed by the original authors while the features set is proposed in GFS-DCF. The EAO measure has improved by up to 7.0% for CSF-GFS-DCF while accuracy (A) is improved by 4.0% and robustness by 3.0%. In terms of robustness measure, RPCF is improved by 4.0% (CSF-RPCF). This experiment demonstrates the effectiveness of the proposed algorithm for fusing information across varying types of features.

B. TRACKER FUSION (`CONFIG-2`)

In this experiment, the proposed CSF algorithm is used to fuse the output response maps of three existing SOTA deep trackers including ATOM, PrDiMP, and DSLT. Performance is evaluated on the VOT2018-ST dataset, with results shown in Table 3. The features and parameters suggested by the original authors are used in each case. We observe a significant EAO improvement of 3.0% beyond PrDiMP tracker. In terms of accuracy, we observe an improvement of up to 3.0% while in terms of robustness an improvement of up to 2.0% is observed. This simple experiment demonstrates the effectiveness of the proposed CSF algorithm in fusing

TABLE 1. Details of the datasets used in experimental evaluations.

Datasets Name	VOT2018-LT [25]	OTB100 [26]	TC128 [27]	UAV123 [28]	VOT2018-ST [25]	GOT-10K [51]	LaSoT [30]
Total Sequences	35	100	129	123	60	10000	1400
Minimum Frames	1389	71	71	109	41	51	1000
Maximum Frames	29700	3872	3872	3085	1500	920	11397
Total Frames	146847	59040	55346	112578	21973	56000 (Testing)	3.52M
Average Resolution	468 × 785	356 × 530	458 × 731	1280 × 720	465 × 758	929 × 1638	632 × 1089

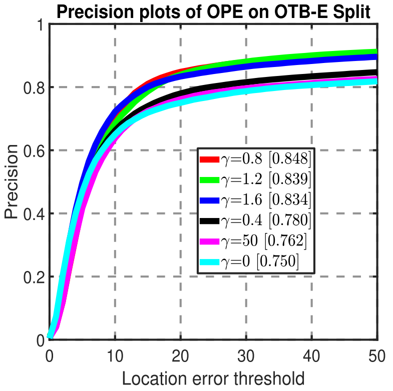


FIGURE 4. Variations of precision rate with varying values of hyper-parameter γ on OTB-E split. $\gamma = 0.8$ has produced the best performance.

TABLE 2. `Config-1`: performance evaluation of the proposed CSF algorithm on the short term VOT2018-ST dataset in configuration 1. The trackers CSF-ASRCF, CSF-GFS-DCF, and CSF-RPCF use the CSF algorithm while ASRCF, GFS-DCF and RPCF are baselines. The best and second best results are shown in red and blue respectively.

Measures	ASRCF	GFS-DCF	RPCF	CSF-ASRCF	CSF-GFS-DCF	CSF-RPCF
EAO↑	0.32	**0.39**	0.31	0.35	**0.46**	0.34
A↑	0.49	0.51	0.50	0.52	**0.55**	**0.54**
R↓	0.23	**0.14**	0.23	0.20	**0.11**	0.19

TABLE 3. `Config-2`: Performance evaluation of the proposed CSF algorithm on the short term VOT2018-ST dataset in configurations 2. The best and second best results are shown in red and blue respectively.

Measures	ATOM	PrDiMP	DSLT	`Config-2`
EAO↑	0.40	**0.44**	0.27	**0.47**
A↑	0.59	**0.61**	0.50	**0.63**
R↓	0.20	**0.16**	0.27	**0.14**

complementary information from different deep trackers, resulting in a significant performance improvement.

C. COMPARISON OF CSF ALGORITHM WITH EXISTING FUSION SCHEMES

The proposed CSF algorithm is also compared with four existing SOTA fusion-based trackers: UPDT, HCF, HDT, and MCCF. For a fair comparison among the compared methods, the classical KCF tracker [31] is used as a baseline and the same set of features including HOG, IC, CN, and deep features extracted from the 4*th* block of ResNet50 are used. Thus, this experiment only compares the strengths and weaknesses of different fusion schemes while keeping all other variables fixed. The experiments are repeated with and without scale estimation on three datasets: OTB100, TC128, and UAV123. The scale of the target object is estimated using the same coarse-to-fine search strategy on HOG features for all compared trackers as proposed in ASRCF tracker [42].

Figure 5 (a)-(d) show the precision and success plots of all trackers on the three compared datasets. The proposed CSF algorithm has consistently demonstrated the best performance in all experiments. On average UPDT gives the second best results on OTB100 while MCCF gives the second best results on the TC128 and UAV123 datasets. In most experiments, HCF shows the lowest performance among the compared fusion-based trackers. The scale estimation step assisted all compared trackers in obtaining better performance, however our proposed CSF algorithm remains the best performer.

D. ABLATION STUDY

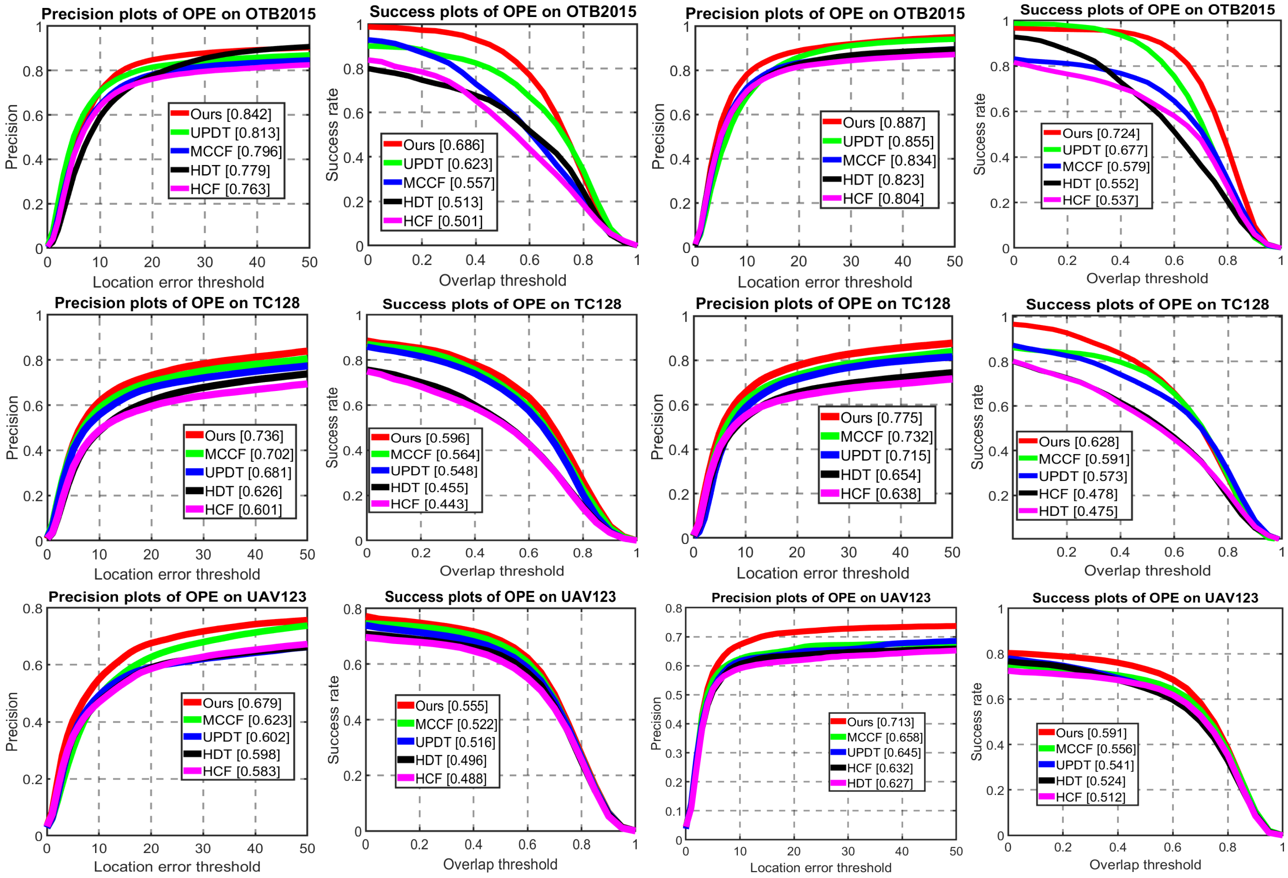
The proposed CSF algorithm has only one hyper-parameter to tune which is γ in Eq. (5). To find a good value of γ, experiments are performed on OTB100 dataset using the protocols defined in [11][1] The authors divided OTB100 sequences into hard videos (OTB-H) as a validation set and easy videos (OTB-E) as a test set considering the performance of different trackers. The value of γ is varied from 0 to 1.6 in intervals of 0.4. In addition, a high value of $\gamma = 50$ is also tested. Figure 4 shows the performance comparison as a precision plot of results from OTB-E with varying values of hyper-parameter γ.

For $\gamma = 0$ in our objective function given by Eq. (5), only the performance using the common low-rank representation component is evaluated while for $\gamma = 50$ the second term, which is the sum of projection distances between common subspace and individual subspaces, becomes more important. In both cases, we observe a graceful degradation of the proposed fusion algorithm while for $\gamma = 50$ the performance was better than for $\gamma = 0$ suggesting that the second term plays a more important role than the first. The best performance is observed when $\gamma = 0.8$, which is the value used in all other experiments.

E. QUALITATIVE RESULTS

To evaluate the performance of the proposed CSF tracker, we present rigorous visual results on key frames of 13 challenging sequences selected from OTB100 dataset and five

[1] Please see supplementary material of UPDT tracker (https://arxiv.org/pdf/1804.06833.pdf) for the list of videos included in OTB-H and OTB-E.



(a) Precision Plots (without scale estimation)

(b) Success Plots (without scale estimation)

(c) Precision Plots (with scale estimation)

(d) Success Plots (with scale estimation)

FIGURE 5. Precision and success plots using OPE for the proposed CSF algorithm against other fusion schemes. Top row shows the results on OTB100, middle row shows the results on TC128, and bottom row shows results on UAV123 datasets. The legends of the precision plots contain threshold scores at 20 pixels [26] while the legends of success plots contain area-under-the-curve scores for each compared tracker. Our proposed CSF algorithm has performed consistently better against these SOTA fusion schemes in all experiments.

sequences from UAV123 dataset. Figure 6 presents the visual results of the proposed CSF tracker. The bounding boxes of the tracked objects are overlaid on the input images and the comparisons are shown with six existing trackers including ATOM, PrDiMP, DSLT, GFS-DCF, ASRCF, and RPCF. The sequences presented in this figure undergo a variety of tracking challenges including occlusion, background clutter, scale variation, deformation, in-plane rotation, out-of-plane rotation, out-of-view, illumination variation, fast motion, motion blur, and low resolution. Overall, the proposed CSF tracker has performed much better than the compared trackers in all these sequences which can be attributed to the fusion of multiple trackers and variety of features within the proposed objective function.

F. EVALUATIONS ON SHORT-TERM TRACKING DATASETS

In addition to the VOT2018 short term tracking dataset evaluated in the previous Section, we have also performed experiments on OTB100, UAV123 and GOT-10K datasets.

1) OTB100 DATASET

The proposed CSF algorithm is evaluated on OTB100 dataset (average video length is 590 frames with 100 videos). The performance is evaluated using the success and precision over varying overlap thresholds. Some visual results of OTB100 dataset are presented in Figure 6. Figure 7 shows the precision and success plots of the proposed CSF tracker with other SOTA trackers including GFS-DCF, RPCF, ASRCF, MCCT, ECO, CCOT, HCFTs, STRCF, and SRDCF. It should be noted that the proposed CSF (ASRCF) tracker used the response maps estimated by ASRCF tracker on different features.

In terms of precision plot, the proposed tracker has obtained 95.1% precision score while the second best GFS-DCF tracker obtained 93.2%. Compared to the baseline ASRCF tracker, the performance of the proposed tracker has improved by 2.8%. In terms of success plot, the proposed tracker CSF tracker has obtained 72.2% success rate while the second best is ECO tracker obtaining 70.0% success rate. Compared to baseline ASRCF tracker, the performance

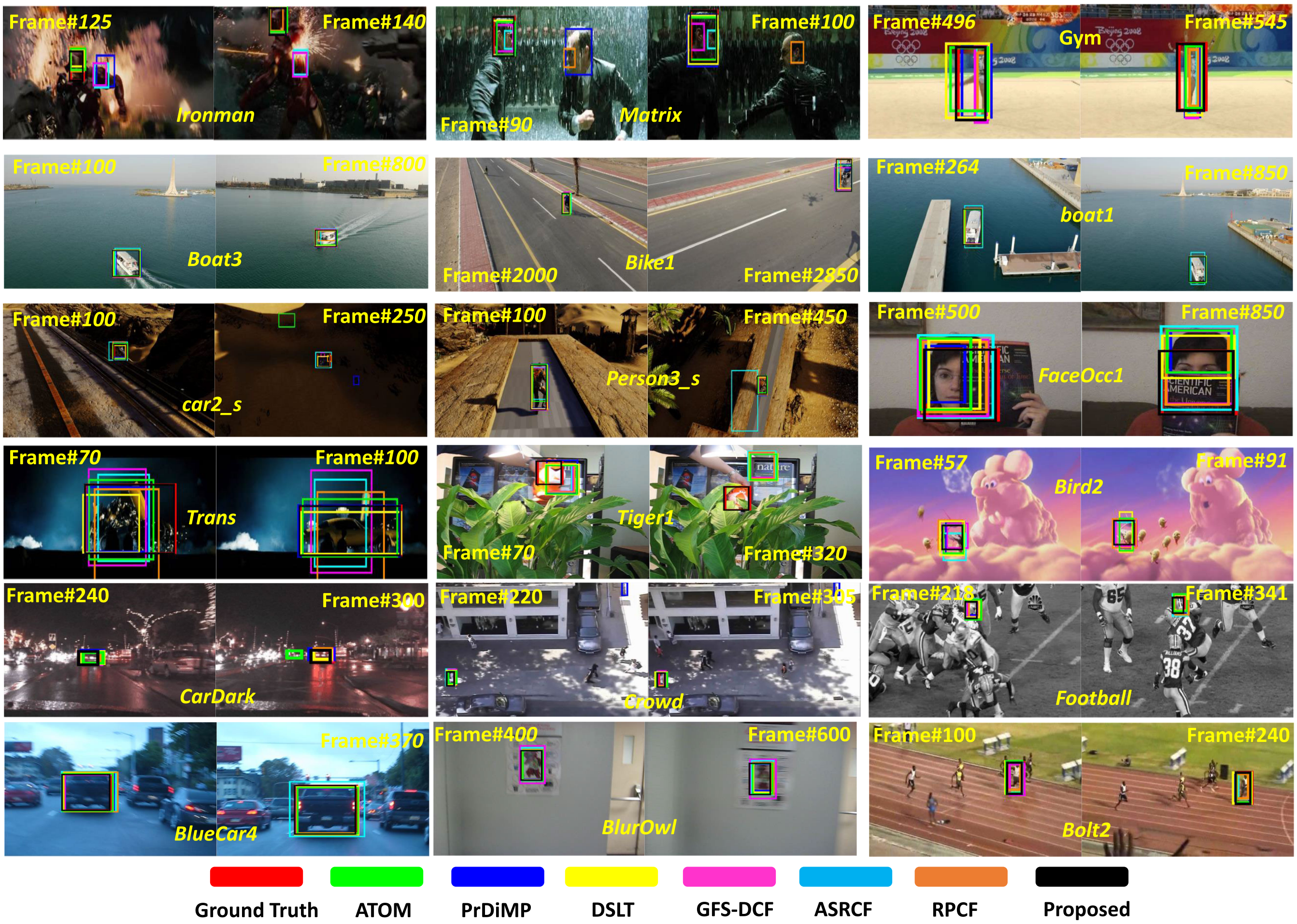
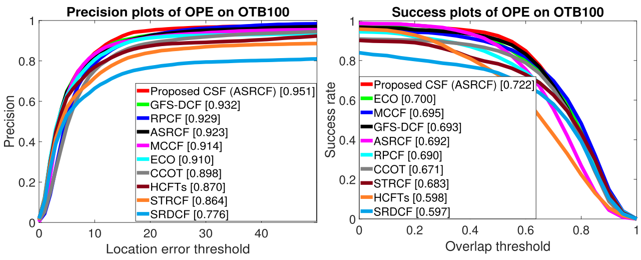


FIGURE 6. Visual results of the proposed CSF algorithm and its comparison with existing SOTA trackers including ATOM [65], PrDiMP [16], DSLT [80], GFS-DCF [41], ASRCF [42], and RPCF [48] on 12 challenging sequences selected from the OTB100 [26] and 6 sequences from UAV123 [28] datasets. Frame indexes and sequence names are shown for each video. Our proposed CSF algorithm has consistently performed better than the compared trackers.

improvement is 3.0%. It demonstrates the effectiveness of our proposed fusion algorithm.

We have also evaluated the attribute-based performance on the OTB100 dataset. The 11 different attributes including Illumination Variation (IV), Occlusion (Occ), Out-of-Plane Rotation (OPR), In-Plane Rotation (IPR), Deformation (DEF), Out of View (OV), Background Clutter (BC), Motion Blur (MB), Low Resolution (LR) and Fast Motion (FM) are evaluated in terms of Precision Rate (PR) and Success Rate (SR) and compared with many SOTA trackers. Table 4 shows the attribute-based performance comparison of the proposed CSF tracker with SOTA trackers.

In terms of Precision Rate (PR), the proposed CSF tracker (baseline ASRCF) achieves the best results under 5 out of 11 challenging tracking attributes including OCC (91.6%), BC (95.1%), DEF (93.1%), OPR (93.2%) and OV (93.9%). For sequences with IV, SV, MB, FM, and IPR tracking challenges, the proposed tracker achieves the second best performance compared to other competing trackers. In terms of Success Rate (SR), the proposed CSF tracker (baseline ASRCF) achieves the best results under 7 out of 11 challenging tracking attributes including IV (72.4%), SV (68.1%), OCC (69.3%), BC (72.2%), DEF (67.6%), OPR (69.3%) and IPR (68.2%). For sequences with MB, FM, OV and LR tracking challenges, the proposed tracker achieves the second best performance compared to other competing trackers. The improved performance of the



(a) Precision Plot

(b) Success Plot

FIGURE 7. Precision and success plots using OPE of the proposed CSF tracker against other SOTA trackers on OTB100 dataset [26]. The legend of precision plot contains threshold scores at 20 pixels, while the legend of success rate contains area-under-the-curve score for each tracker.

TABLE 4. Attribute-based performance comparison of the proposed and existing SOTA trackers in terms of Precision Rate (PR)|Success Rate (SR) on OTB100 dataset. The PR is reported at a threshold of 20 pixels while AUC is shown for SR.

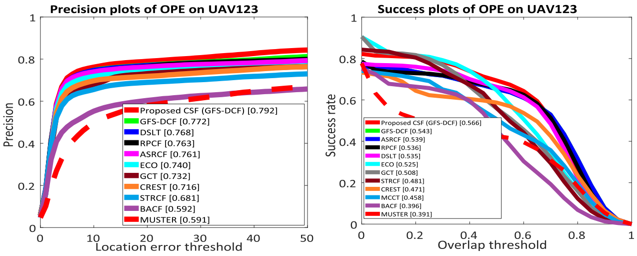
Trackers	IV(38)	SV(64)	Occ(49)	BC(31)	MB(29)	DEF(44)	OPR(63)	FM(39)	IPR(51)	OV(14)	LR(9)
GFS-DCF	**95.1**|**72.2**	**91.7**|**67.6**	88.0|66.6	92.1|68.3	**92.1**|69.9	89.6|65.4	**92.9**|**68.0**	**93.7**|**69.7**	**94.7**|**67.3**	**93.2**|**69.0**	**95.0**|**65.0**
ASRCF	92.8|70.9	89.3|66.1	89.1|67.9	93.2|70.0	88.6|68.0	**90.2**|**66.0**	91.6|67.5	88.4|67.1	89.8|65.9	91.8|66.6	85.8|61.2
DeepMCCT	88.6|68.5	88.7|64.8	86.0|64.7	92.5|**70.1**	85.7|66.9	88.1|63.5	89.6|65.8	87.4|65.0	**91.3**|65.6	86.3|64.3	88.6|61.1
ECO	91.4|71.3	88.1|66.7	**90.8**|**68.1**	**94.2**|**70.1**	90.4|**71.0**	85.9|63.3	90.7|67.3	86.5|68.4	89.2|65.6	91.2|66.3	88.1|60.3
CCOT	87.5|67.4	87.6|65.4	90.2|67.4	88.2|65.2	90.3|**70.5**	86.0|61.4	89.2|65.2	86.8|67.6	86.8|62.7	89.5|64.8	**97.5**|62.9
DGL	86.0|62.0	85.8|57.9	84.8|60.9	88.5|65.2	79.2|60.6	86.2|59.5	87.1|60.1	80.5|59.2	86.3|59.7	80.2|58.6	79.4|51.8
HCFT	88.8|60.3	82.6|52.5	81.3|55.8	88.6|62.1	82.2|60.6	82.6|56.0	84.9|57.3	82.2|58.1	89.5|60.3	74.5|52.0	82.2|48.8
TADT	86.4|67.6	86.5|65.4	84.4|64.2	80.9|62.5	83.7|67.3	82.2|60.4	87.2|64.6	83.8|65.9	83.5|62.2	82.5|63.0	88.1|63.4
DeepSTRCF	84.1|65.3	84.2|63.2	81.4|61.6	87.2|64.8	82.6|65.3	84.4|60.7	83.5|62.6	80.2|62.9	81.1|60.2	76.6|58.5	73.7|53.8
GradNet	76.3|59.7	78.9|59.0	75.8|57.3	81.3|61.2	79.1|62.6	76.5|56.3	77.7|56.9	73.7|57.4	76.0|55.5	77.9|59.4	78.8|56.4
DeepSRDCF	79.1|62.1	81.9|60.6	82.5|60.2	84.1|62.8	82.2|64.3	78.3|56.6	83.5|60.7	81.4|62.9	81.8|58.9	77.9|46.3	70.8|47.5
MCCT	77.9|60.7	80.2|59.2	77.2|59.6	85.4|64.8	75.2|59.2	80.8|60.2	81.5|60.4	76.3|59.0	78.2|57.8	76.9|57.8	68.6|47.9
HDT	73.9|49.2	70.0|44.1	68.0|47.9	70.0|48.5	69.2|51.4	71.0|49.0	70.5|48.4	72.6|52.4	72.8|49.4	59.6|44.6	57.1|28.9
BACF	80.8|62.4	77.1|57.4	73.6|56.8	80.1|60.6	74.1|57.5	77.0|57.4	77.9|57.8	78.7|59.9	79.2|58.3	74.7|54.8	74.1|53.2
TRACA	84.1|62.2	76.9|55.7	77.9|57.3	80.6|59.7	76.3|59.4	76.9|56.1	82.3|59.3	75.2|57.8	81.6|58.2	71.5|55.7	70.7|46.7
DeepRSST	83.6|55.2	82.6|50.0	76.6|51.9	86.9|60.3	77.5|57.0	79.4|53.1	81.3|53.7	81.5|57.0	86.3|56.8	70.3|50.7	82.6|45.4
SRDCF	76.7|60.0	73.3|55.5	71.3|55.2	73.2|56.6	73.6|57.9	72.6|54.1	72.7|54.1	75.3|58.7	71.5|52.8	58.7|46.3	64.0|49.1
MUSTer	78.2|60.0	71.5|51.8	73.4|55.4	78.6|57.9	69.9|55.7	68.9|52.4	74.8|54.1	69.1|53.9	77.3|55.1	59.1|46.9	67.7|47.7
LCT	74.6|56.6	68.0|48.8	68.2|50.7	73.3|55.0	66.8|53.4	68.9|49.9	74.6|53.8	68.0|53.5	78.1|55.7	59.0|45.3	53.7|35.4
UDT	76.3|59.7	78.9|59.0	75.8|57.3	81.3|61.2	79.1|62.6	76.5|56.3	77.7|56.9	73.7|57.4	76.0|55.5	77.9|59.4	78.8|56.4
CFNET	69.4|54.4	71.5|53.6	67.9|51.9	73.3|55.4	67.4|55.4	64.3|47.3	73.4|54.2	70.3|55.5	77.2|57.1	55.3|42.6	82.9|61.9
Proposed CSF	**94.3**|**72.4**	**91.5**|**68.1**	**91.6**|**69.3**	**95.1**|**72.2**	**90.5**|**70.5**	**93.1**|**67.6**	**93.2**|**69.3**	**90.4**|**69.0**	**91.3**|**68.2**	**93.9**|**68.5**	88.1|**63.5**

proposed tracker demonstrates the effectiveness of the common subspace fusion mechanism.

2) UAV123 DATASET

This dataset contains 123 video sequences with an average length of 915 frames [28]. The results are compared with 15 SOTA trackers: ECO, GCT, CREST, SRDCF, STRCF, MEEM, BACF, MUSTER, DSST, MCCT, STAPLE, ASRCF, GFS-DCF, RPCF and DSLT. Some visual results from the UAV123 dataset are shown in Figure 6. Figure 8 shows the performance comparison of the proposed CSF tracker with other SOTA trackers in terms Precision Rate (PR) and Success Rate (SR).

Overall, the proposed CSF tracker achieves the best precision rate of 79.2% which is 2.0% better than the baseline tracker, GFS-DCF (77.2%), and 2.4% better than the deep tracker, DSLT. The proposed CSF algorithm also achieves the best success rate (AUC) of 56.6% which is 2.3% better than GFS-DCF and 3.1% better than DSLT. The best performance achieved by the proposed tracker is because of the subspace fusion mechanism across varying types of feature representations incorporated within the baseline tracker GFS-DCF.



(a) Precision Plot

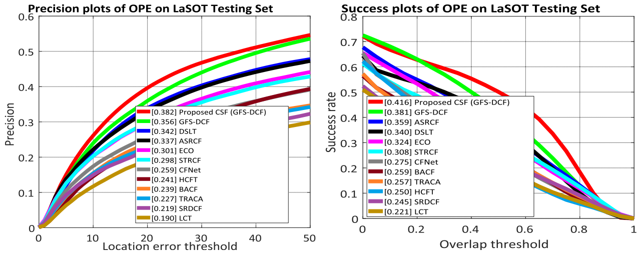
(b) Success Plot

FIGURE 8. Precision and success plots using OPE of the proposed CSF tracker against other SOTA trackers on UAV123 dataset [28]. The legend of precision plot contains threshold scores at 20 pixels, while the legend of success rate contains area-under-the-curve score for each tracker.

3) GOT-10K DATASET

Its test tracking split consists of 180 videos with an average length of 127 frames [51]. In this dataset, our proposed tracker, CSF (GFS-DCF), using GFS-DCF as a baseline is compared with 11 SOTA trackers including HCF, STAPLE, DSST, ECO, STRCF, CCOT, CFNET, BACF, MEEM, SRDCF, ASRCF, and GFS-DCF as shown in Table 5. The performance is evaluated in terms of mean Average Overlap (AO), mSR$_{0.5}$, and mSR$_{0.75}$ using official online protocols.

In terms of mAO, the proposed CSF tracker achieves 46.4% performance which is 4.2% better than the baseline tracker, GFS-DCF (42.20%), and 5.80% better performance than the CCOT tracker (40.6%). In terms of mSR$_{0.50}$ measure, CSF tracker obtains best score of 48.20% which is again 2.20% better than the second best performing tracker GFS-DCF (46.0%) and 6.70% than the CCOT tracker (41.50%). Similarly, the proposed tracker achieves mSR$_{0.75}$ sore of 16.20% which is 0.80% less than the best performing tracker, ECO (17.0%), and 0.10% less than the second best performing tracker, CCOT (16.1 %). The improved performance in terms of both measures, mAO and mSR$_{0.50}$, is because of the fusion component introduced within the baseline tracker GFS-DCF.



(a) Precision Plot

(b) Success Plot

FIGURE 9. Precision and success plots using OPE of the proposed CSF tracker against other SOTA trackers on LASoT dataset [30]. The legend of precision plot contains threshold scores at 20 pixels, while the legend of success rate contains area-under-the-curve score for each tracker.

G. EVALUATION ON LONG-TERM TRACKING DATASET

We have also evaluated the performance of the proposed tracker on two long term visual object tracking datasets including LaSoT [30] and VOT2018-LT [25]. In the below subsections, we describe the performance comparison of the proposed tracker on these datasets.

1) LaSoT DATASET [30]

The test split of this dataset consists of 280 videos with an average length of 2448 frames [30]. The results of the proposed CSF tracker are compared with 15 SOTA trackers including ECO, DSLT, BACF, HCFTs, CFNET, LCT, SRDCF, TRACA, STAPLE, STRCF, ASRCF, and GFS-DCF. The performance is reported in terms of PR and SR by using the protocols provided by the original authors [30].

In terms of PR, CSF tracker has obtained 38.20% accuracy which is 2.6% better than the existing baseline tracker, GFS-DCF (35.60%), and 4.50% better than the ASRCF tracker as shown in Figure 9. In terms of SR, CSF tracker has achieved 3.50% performance improvements compared to GFS-DCF and up to 5.70% better accuracy than ASRCF. This experiment demonstrates the effectiveness of our proposed fusion algorithm on long term tracking challenges.

2) VOT2018-LT DATASET [25]

The long term challenge of VOT2018 dataset consists of 35 video sequences with an average resolution of 468 × 785 as shown in Table 1. The proposed CSF tracker has also been evaluated on this dataset in terms of F-score, Recall (Re), and Precision (Pr) measures as defined in the VOT2018-LT evaluation kit [25]. In this dataset, ranking is achieved using maximum F-score attained by each tracker.

Table 6 shows the performance comparison of the proposed CSF tracker with five SOTA trackers including CCOT, DeepSRDCF, DeepSTRCF, UDT, and GFS-DCF. Overall, it can be observed that our proposed CSF tracker has attained best F-score of 67.80% and outperforms GFS-DCF, UDT, and CCOT trackers by 2.60% and 5.80% margin, respectively. The corresponding Re and Pr scores also demonstrated the improvements of the proposed tracker compared to second best tracking method.

TABLE 5. Performance comparison of the proposed CSF tracker with current SOTA trackers on GOT-10K dataset [51]. The performance is reported in terms of mAO (mean Average Overlap), mSR$_{0.50}$ and mSR$_{0.75}$ (Success Rate that measures the percentage of successfully tracked frames, where the overlap precision exceeds a threshold of 0.50 and 0.75). The best and second best results are shown in red and blue respectively. Our proposed tracker has consistently performed better than the SOTA trackers.

Trackers	mAO	mSR$_{0.50}$	mSR$_{0.75}$
HCF	0.379	0.380	0.134
ECO	0.395	0.407	**0.170**
CCOT	0.406	0.415	0.161
CFNET	0.364	0.365	0.150
BACF	0.346	0.365	0.150
SRDCF	0.312	0.310	0.134
ASRCF	0.313	0.317	0.113
GFS-DCF	**0.422**	**0.460**	0.120
Proposed CSF (GFS-DCF)	**0.464**	**0.482**	**0.162**

TABLE 6. Performance comparison of the proposed CSF tracker with current SOTA trackers on VOT2018-LT dataset [25]. The performance is reported in terms of F-socre, Precision (Pr) and Recall (Re). The best and second best results are shown in red and blue respectively. Our proposed tracker has consistently performed better than the SOTA Trackers.

Trackers	F-score	Pr	Re
CCOT	0.620	0.590	0.660
DeepSRDCF	0.570	0.410	0.930
DeepSTRCF	0.510	0.520	0.510
UDT	0.620	0.820	0.50
GFS-DCF	**0.652**	**0.850**	**0.532**
Proposed CSF (GFS-DCF)	**0.678**	**0.861**	**0.563**

H. EXECUTION TIME COMPARISON

The execution time of the proposed CSF tracker is measured on a PC with an Intel core i7 4.0 GHz, Titan Xp GPU, and 64-GB RAM. Our complete tracking pipeline including feature extractions, using KCF filter and CSF algorithm (IV-C) runs at 15.71 frames per second (fps), while the existing fusion-based trackers UPDT, MCCF, HDT, and HCF process frames at the rate of 8.11fps, 4.98fps, 7.41fps and 13.52fps respectively. For the case of `config-1` and `config-2`, the additional time taken by the CSF algorithm depends on the number of features or trackers to be fused. For the case of `config-1`, fusion is performed on four different types of features (as discussed in IV-A). The time taken by ASRCF is 28.32fps while the time taken by the proposed CSF-ASRCF is 34.19fps. Thus, the additional time taken in this configuration is 5.87fps. Thus, the proposed CSF fusion algorithm is computationally efficient and does not incur significant computational overhead beyond the underline baseline trackers.

V. CONCLUSION

In this work, an information fusion algorithm is proposed to encode the complementary information contained by various types of features and trackers to improve VOT performance. For this purpose, low-rank representations of response maps are computed which remove unwanted boundary effects and suppress noise. A common low-rank subspace representation is estimated such that it is close to each individual subspace on the Grassmann manifold in terms of projection distance. The common subspace representation acts as the fusion scheme which integrates information encoded by individual low-rank response maps. The CSF algorithm is generic and works well with various types of features, correlation filter-based trackers, and deep trackers. Evaluations are performed for feature fusion and tracker fusion on seven challenging tracking benchmark datasets and compared with several SOTA trackers. Our algorithm has consistently demonstrated significant performance improvements over various baseline methods. The CSF algorithm has also outperformed existing fusion schemes using the same features and baseline tracker. We observe that the fusion of deep correlation filters-based trackers has resulted in the highest performance gain. The SOTA fusion-based tracking methods assign weights to different feature or response map. An advantage of the proposed fusion algorithm is that it does not require weight assignment to different feature representations or response maps. The proposed fusion algorithm finds it challenging to handle significant target scale and orientation variations. In future, the proposed fusion algorithm will be implemented as a deep layer in an end-to-end network for VOT.

REFERENCES

[1] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 28, 2021, doi: 10.1109/TITS.2020.3046478.

[2] S. Javed, M. Danelljan, F. S. Khan, M. H. Khan, M. Felsberg, and J. Matas, "Visual object tracking with discriminative filters and Siamese networks: A survey and outlook," 2021, *arXiv:2112.02838*.

[3] A. Ess, K. Schindler, B. Leibe, and L. Van Gool, "Object detection and tracking for autonomous navigation in dynamic environments," *Int. J. Robot. Res.*, vol. 29, no. 14, pp. 1707–1725, 2010.

[4] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[5] D. Koller, G. Klinker, E. Rose, D. Breen, R. Whitaker, and M. Tuceryan, "Real-time vision-based camera tracking for augmented reality applications," in *Proc. ACM Symp. Virtual Reality Softw. Technol. (VRST)*, 1997, pp. 87–94.

[6] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.

[7] O. Javed and M. Shah, "Tracking and object classification for automated surveillance," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2002, pp. 343–357.

[8] S. Javed, M. Danelljan, F. S. Khan, M. H. Khan, M. Felsberg, and J. Matas, "Visual object tracking with discriminative filters and Siamese networks: A survey and outlook," 2021, *arXiv:2112.02838*.

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

[10] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4303–4311.

[11] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg, "Unveiling the power of deep tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 483–498.

[12] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4844–4853.

[13] G. Wang, C. Luo, X. Sun, Z. Xiong, and W. Zeng, "Tracking by instance detection: A meta-learning approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6288–6297.

[14] P. Voigtlaender, J. Luiten, P. H. S. Torr, and B. Leibe, "Siam R-CNN: Visual tracking by re-detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6578–6588.

[15] A. Lukezic, J. Matas, and M. Kristan, "D3S—A discriminative single shot segmentation tracker," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7133–7142.

[16] M. Danelljan, L. Van Gool, and R. Timofte, "Probabilistic regression for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7183–7192.

[17] T. Yang, P. Xu, R. Hu, H. Chai, and A. B. Chan, "ROAM: Recurrently optimizing tracking model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6718–6727.

[18] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "AutoTrack: Towards high-performance visual tracking for UAV with automatic spatio-temporal regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11923–11932.

[19] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, "Ocean: Object-aware anchor-free tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 771–787.

[20] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Know your surroundings: Exploiting scene information for object tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 205–221.

[21] S. M. Marvasti-Zadeh, H. Ghanei-Yakhdan, and S. Kasaei, "Adaptive exploitation of pre-trained deep convolutional neural networks for robust visual tracking," *Multimedia Tools Appl.*, vol. 80, no. 14, pp. 22027–22076, Jun. 2021.

[22] M. Jiang, Y. Zhao, and J. Kong, "Mutual learning and feature fusion Siamese networks for visual object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3154–3167, Aug. 2021.

[23] W. Zhang, R. Song, and Y. Li, "Online decision based visual tracking via reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 11778–11788.

[24] G. S. Walia, H. Ahuja, A. Kumar, N. Bansal, and K. Sharma, "Unified graph-based multicue feature fusion for robust visual tracking," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2357–2368, Jun. 2020.

[25] M. Kristan *et al.*, "The sixth visual object tracking VOT2018 challenge results," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 3–53.

[26] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

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

[28] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 445–461.

[29] H. K. Galoogahi, A. Fagg, C. Huang, D. Ramanan, and S. Lucey, "Need for speed: A benchmark for higher frame rate object tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1125–1134.

[30] H. Fan, H. Ling, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, and C. Liao, "LaSOT: A high-quality benchmark for large-scale single object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5374–5383.

[31] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[32] S. Javed, X. Zhang, J. Dias, L. Seneviratne, and N. Werghi, "Spatial graph regularized correlation filters for visual object tracking," in *Proc. Int. Conf. Soft Comput. Pattern Recognit.* Cham, Switzerland: Springer, 2020, pp. 186–195.

[33] S. Javed, X. Zhang, L. Seneviratne, J. Dias, and N. Werghi, "Deep bidirectional correlation filters for visual object tracking," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Jul. 2020, pp. 1–8.

[34] S. Javed, A. Mahmood, J. Dias, and N. Werghi, "Multi-level feature fusion for nucleus detection in histology images using correlation filters," *Comput. Biol. Med.*, vol. 143, Apr. 2022, Art. no. 105281.

[35] H. Zhu, H. Peng, G. Xu, L. Deng, Y. Cheng, and A. Song, "Bilateral weighted regression ranking model with spatial-temporal correlation filter for visual tracking," *IEEE Trans. Multimedia*, early access, Apr. 28, 2021, doi: 10.1109/TMM.2021.3075876.

[36] Q. Wang, C. Yuan, J. Wang, and W. Zeng, "Learning attentional recurrent neural network for visual tracking," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 930–942, Apr. 2019.

[37] X.-F. Zhu, X.-J. Wu, T. Xu, Z.-H. Feng, and J. Kittler, "Robust visual object tracking via adaptive attribute-aware discriminative correlation filters," *IEEE Trans. Multimedia*, vol. 24, pp. 301–312, 2022.

[38] K. Yang, Z. He, W. Pei, Z. Zhou, X. Li, D. Yuan, and H. Zhang, "SiamCorners: Siamese corner networks for visual tracking," *IEEE Trans. Multimedia*, early access, Apr. 21, 2021, doi: 10.1109/TMM.2021.3074239.

[39] Y. Zhang, B. Ma, J. Wu, L. Huang, and J. Shen, "Capturing relevant context for visual tracking," *IEEE Trans. Multimedia*, vol. 23, pp. 4232–4244, 2021.

[40] S. Javed, A. Mahmood, J. Dias, N. Werghi, and N. Rajpoot, "Spatially constrained context-aware hierarchical deep correlation filters for nucleus detection in histology images," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102104.

[41] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7950–7960.

[42] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4670–4679.

[43] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M.-H. Yang, "CREST: Convolutional residual learning for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2555–2564.

[44] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[45] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2805–2813.

[46] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Adaptive correlation filters with long-term and short-term memory for object tracking," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 771–796, Aug. 2018.

[47] S. Javed, A. Mahmood, J. Dias, L. Seneviratne, and N. Werghi, "Hierarchical spatiotemporal graph regularized discriminative correlation filter for visual object tracking," *IEEE Trans. Cybern.*, early access, Jul. 7, 2021, doi: 10.1109/TCYB.2021.3086194.

[48] Y. Sun, C. Sun, D. Wang, Y. He, and H. Lu, "ROI pooled correlation filters for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5783–5791.

[49] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2006, pp. 126–135.

[50] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov, "Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds," *IEEE Trans. Signal Process.*, vol. 62, no. 4, pp. 905–918, Feb. 2014.

[51] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1562–1577, May 2021.

[52] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6668–6677.

[53] Q. Liu, X. Lu, Z. He, C. Zhang, and W.-S. Chen, "Deep convolutional neural networks for thermal infrared object tracking," *Knowl.-Based Syst.*, vol. 134, pp. 189–198, Oct. 2017.

[54] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, and H. Wang, "Learning deep multi-level similarity for thermal infrared object tracking," *IEEE Trans. Multimedia*, vol. 23, pp. 2114–2126, Jul. 2020.

[55] Q. Liu, D. Yuan, N. Fan, P. Gao, X. Li, and Z. He, "Learning dual-level deep representation for thermal infrared tracking," *IEEE Trans. Multimedia*, early access, Jan. 6, 2022, doi: 10.1109/TMM.2022.3140929.

[56] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.

[57] N. Gillis, *Nonnegative Matrix Factorization*. Philadelphia, PA, USA: SIAM, 2020.

[58] X. Peng, C. Lu, Z. Yi, and H. Tang, "Connections between nuclear-norm and Frobenius-norm-based representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 218–224, Jan. 2018.

[59] V. Gligorijević, Y. Panagakis, and S. P. Zafeiriou, "Non-negative matrix factorizations for multiplex network analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 928–940, Apr. 2019.

[60] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[61] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2001, pp. 556–562.

[62] S. I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.

[63] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 576–588, Mar. 2010.

[64] Z. Yang and J. Laaksonen, "Multiplicative updates for non-negative projections," *Neurocomputing*, vol. 71, nos. 1–3, pp. 363–373, Dec. 2007.

[65] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4660–4669.

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

[67] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Robust visual tracking via hierarchical convolutional features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2709–2723, Nov. 2019.

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

[69] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 472–488.

[70] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6638–6646.

[71] C. Li, L. Lin, W. Zuo, J. Tang, and M.-H. Yang, "Visual tracking via dynamic graph learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2770–2782, Nov. 2019.

[72] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1369–1378.

[73] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 58–66.

[74] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang, and H. Lu, "GradNet: Gradient-guided network for visual object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6162–6171.

[75] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1135–1143.

[76] J. Choi, H. J. Chang, T. Fischer, S. Yun, K. Lee, J. Jeong, Y. Demiris, and J. Y. Choi, "Context-aware deep feature compression for high-speed visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 479–488.

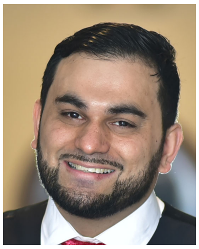
[77] T. Zhang, C. Xu, and M.-H. Yang, "Robust structural sparse tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 473–486, Feb. 2019.

[78] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "MUlti-store tracker (MUSTer): A cognitive psychology inspired approach to object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 749–758.

[79] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, and H. Li, "Unsupervised deep tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1308–1317.

[80] X. Lu, C. Ma, J. Shen, X. Yang, I. Reid, and M.-H. Yang, "Deep object tracking with shrinkage loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 30, 2020, doi: 10.1109/TPAMI.2020.3041332.

[81] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4649–4659.



SAJID JAVED received the B.Sc. degree in computer science from the University of Hertfordshire, U.K., in 2010, and the combined master's and Ph.D. degrees in computer science from Kyungpook National University, Republic of Korea, in 2017. He is currently an Assistant Professor with the Computer Vision in Electrical and Computer Engineering (ECE) Department, Khalifa University of Science and Technology, United Arab Emirates. Prior to that, he was a Research Scientist with the Khalifa University Center for Autonomous Robotics System (KUCARS), United Arab Emirates, from 2019 to 2021. Before joining Khalifa University, he was a Research Fellow with the University of Warwick, U.K., from 2017 to 2018, where he worked on histopathological landscapes for better cancer grading and prognostication. His research interests include visual object tracking in the wild, multi-object tracking, background-foreground modeling from video sequences, moving object detection from complex scenes, cancer image analytics, including tissue phenotyping, nucleus detection, and nucleus classification problems. His research themes involve developing deep neural networks, subspace learning models, and graph neural networks.

ARIF MAHMOOD has worked as a Research Assistant Professor with the School of Mathematics and Statistics (SMS), University of the Western Australia (UWA). In SMS, he worked on community detection in social and scientific networks. Before that, he was a Research Assistant Professor with the School of Computer Science and Software Engineering, UWA, and performed research on face recognition, object classification, and action recognition. He is currently a Professor with the Punjab and Director of Computer Vision Laboratory, Department of Computer Science, Information Technology University (ITU). He is actively working on cancer grading and prognostication using histology images, predictive auto-scaling of services hosted on the cloud and the fog infrastructures, and ocean color monitoring using remote sensing. His current research interests include person pose detection and segmentation, crowd counting and flow detection, background-foreground modeling in complex scenes, object detection, human-object interaction detection, and abnormal event detection.

IHSAN ULLAH received the Ph.D. degree from the University of Milan, specializing in designing lightweight deep neural network architectures with the pyramidal approach. He has more than nine year's of research and development experience in applying deep learning to a variety of images, video, text, and time-series recognition problems while working with renowned labs in the USA (Computational Vision and Geometry Laboratory, Stanford University), Europe (CVPR Laboratory, University of Naples Parthenope, Italy), and the Middle East (Visual Computing Laboratory, King Saud University, Saudi Arabia). Before joining the School of Computer Science, NUI Galway, he was a Senior Research Data Scientist with the CeADAR Ireland's Centre for Applied AI, University College Dublin, where he was the Head of the Special Projects Group and was actively involved in applying for various national and international funding's e.g., Horizon Europe, SFI, and EI. Prior to that, he worked with the Data Mining and Machine Learning Group, School of Computer Science, NUI Galway, as a Senior Postdoctoral Researcher, an Adjunct Lecturer, and the Project Manager of the H2020 Project ROCSAFE. He also worked as a Postdoctoral Researcher with the INSIGHT Research Centre, NUI Galway, and a Research Engineer at Prosa S.r.l., Italy. His current research interests include designing lightweight deep learning models, computer vision and pattern recognition, explainable AI, federated learning, and differential privacy. He is currently an Invited Member of the National Standards Authority of Ireland prestigious "Top Team" on setting the national Standards in AI. He is also a Steering Committee Member of Oblivious.ai.

THIERRY BOUWMANS (Member, IEEE) is currently an Associate Professor with the University of La Rochelle, France. He is also the Creator and the Administrator of the Background Subtraction Website. He has recently authored more than 30 papers in the field of background modeling and foreground detection. These papers investigated particularly the use of fuzzy concepts, discriminative subspace learning models, and robust PCA. They also develop surveys on mathematical tools used in the field and particularly on robust PCA via principal component analysis. He has supervised Ph.D. students in this field. His research interest includes the detection of moving objects in challenging environments. He served as the lead guest editors in two editorial works, such as the special issue in MVA on "Background Modeling for Foreground Detection in Real-World Dynamic Scenes" and the Handbook on "Background Modeling and Foreground Detection for Video Surveillance" in CRC Press. He has served as a reviewer for numerous international conferences and journals.

MAJID KHONJI (Member, IEEE) received the M.Sc. degree in security, cryptology, and coding of information systems from the Ensimag, Grenoble Institute of Technology, France, and the Ph.D. degree in interdisciplinary engineering from the Masdar Institute, United Arab Emirates, in collaboration with MIT, in 2016. He is currently an Assistant Professor with the EECS Department, Khalifa University, United Arab Emirates, and a Research Affiliate with the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), USA. Previously, he was a Visiting Assistant Professor with MIT CSAIL, a Senior Research and Development Technologist in Dubai electricity and water authority (DEWA), and an Information Security Researcher with the Emirates Advanced Investment Group (EAIG).

JORGE MANUEL MIRANDA DIAS received the Habilitation and Ph.D. degrees in electrical engineering from the University of Coimbra, Portugal, with a focus on specialization in control and instrumentation. He is currently a Professor in ECE/robotics with Khalifa University, Abu Dhabi. His expertise has been in the area of computer vision and robotics and has contributions on the field, since 1984. He has several publications in international journals, books, and conferences. He was a principal investigator from several international research projects. He has published several articles in the area of computer vision and robotics that include more than 80 publications in international journals, one published book, 15 books chapters, and more than 280 articles in international conferences with referee.

NAOUFEL WERGHI (Senior Member, IEEE) received the Habilitation and Ph.D. degrees in computer vision from the University of Strasbourg. He has been a Research Fellow with the Division of Informatics, University of Edinburgh, and a Lecturer with the Department of Computer Sciences, University of Glasgow. He is currently an Associate Professor with the Electrical Engineering and Computer Science Department, Khalifa University for Science and Technology. He has been a Visiting Professor with the University of Louisville, University of Florence, University of Lille, and the Korean Advanced and Institute of Sciences and Technology, South Korea. His research interests include 2D/3D image analysis and interpretation, where he has been leading several funded projects related to biometrics, medical imaging, remote sensing, and intelligent systems. He is an Associate Editor in the *EURASIP Journal for Image and Video Processing*. He is a member of the IEEE Signal Processing Society and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.

• • •