Visual Tracking Via Multi-Layer Factorized Correlation Filter

Bin Kang, Gaowei Chen, Quan Zhou, Jun Yan, and Min Lin

***Abstract*—Pruning the parameters of basis filters can effectively eliminate the negative effect of redundant deep features in discriminative correlation filter based trackers. However, traditional methods often treat feature maps in Convolutional Neural Networks (CNN) as isolate observations, ignore the intrinsic correlation between partially attentional feature maps in multiple convolutional layers, when basis filter pruning is pursued. In this letter, we propose a multi-layer factorized discriminant correlation filter (MLF-DCF) for visual tracking. By integrating the multi-view discriminant learning and the discriminative correlation filter into a unified optimization problem, we can explore the correlation between different target sub-regions from multi-layer viewpoint, thus can effectively prune multi-layer basis filters. To enhance the efficiency of MLF-DCF in terms of speed and accuracy, we not only adopt alternating direction method of multipliers (ADMM) to solve the unified optimization, but also employ a mask estimation strategy to eliminate the background noise in deep features. A large number of experiments on challenging video sequences are given to illustrate the superiority of our tracking method.**

***Index Terms*—Visual tracking, factorized correlation filter, multi-view discriminant learning, ADMM.**

I. Introduction

VISUAL tracking, aiming at predicting the state of the target at each frame, plays a very important role in computer vision for its extensive applications in video analysis, vehicle navigation and human-computer interaction. A typical tracking algorithm mainly includes: 1) a motion model aiming to track the state of moving target; 2) an observation model evaluating the likelihood of each target candidate. According to different observation models, state-of-the-art visual tracking can be categorized into the support vector machine based trackers [1]–[3], the sparse representation based trackers [4]–[6], the deep learning based trackers [7]–[10] and the correlation filter based trackers [11]–[14] *et al.*

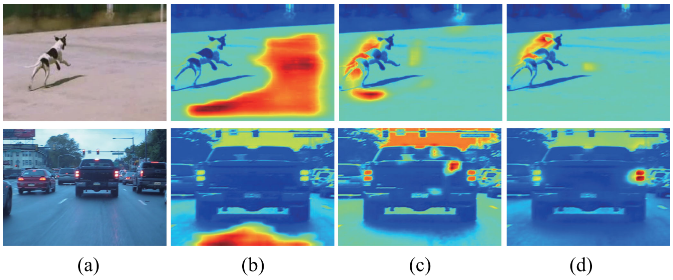

Fig. 1. Randomly selected heat maps of the last convolutional layers in pretrained VGG-m net. (a) are the examples of input images, which are obtained from the dog and blurcar4 video sequences. (b) and (c) are the examples of heat maps of the last convolutional layers in C-COT (the 70-th and 142-th channels of dog, the 202-th and 479-th channels of blurcar4). (d) are the examples of masked feature maps in MLF-DCF.

Recently, with the successful implementation of deep learning in object detection and image classification, it has aroused a great deal of attention in visual tracking. Unlike image classification, it is hard to build a large scale reliable dataset for pretraining the tracking model because labeled samples can only be collected in the first few video frames and the appearance of the tracked target often changes dramatically. Aforementioned limitations may prevent the end-to-end neural network based trackers unleashing their full potential. Due to this fact, a substantial effort has been put on the correlation filter based trackers due to their advantage of adopting large numbers of cyclically shifted samples for the training of appearance model. Since traditional correlation filter often uses handcraft target feature to achieve visual tracking, the tracking accuracy may be reduced in challenging scenarios. To overcome this limitation, C-COT [15], ECO [16] and STRDCF [17] are successively proposed to introduce deep features in discriminative correlation filter, which can give more semantic information for target regression.

Existing deep feature based discriminative correlation filters, such as [15], [18] and [17], often directly use pre-trained VGG-m net to yield partially attentional features maps. This may incur two challenges: 1) In some feature maps such as Fig. 1(b), the active regions (highlighted by warm color) can not locate the semantically meaningful parts of the targets. This kind of feature maps can not contribute to target localization, hence they are redundant. 2) In some feature maps such as Fig. 1(c), although the active regions can locate the semantically meaningful parts of the targets, those active regions may contain background noise. To overcome aforementioned challenges, the ECO [16] method

Manuscript received August 26, 2019; revised October 1, 2019; accepted October 4, 2019. Date of publication October 14, 2019; date of current version November 11, 2019. This work was supported in part by the NSFC under Grants 61801242, 61876093, 61871235, 61771258, and 61771256; in part by NSF of Jiangsu Province under Grants BK20170915 and BK20181393; in part by the Key International Cooperation Research Project under Grant 61720106003; in part by NUPT program under Grant NY218075; and in part by the Open Research Fund of Jiangsu Engineering Research Center of Communication and Network Technology , NUPT. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wei Li. *(Corresponding authors: Bin Kang; Quan Zhou.)*

B. Kang was with the Department of Internet of Things and the Jiangsu Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: kangb@njupt.edu.cn).

G. Chen was with the Department of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: s844975016@163.com).

Q. Zhou, J. Yan, and M. Lin were with the Department of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: quan.zhou@njupt.edu.cn; yanj@njupt.edu.cn; linmin@njupt.edu.cn).

Digital Object Identifier 10.1109/LSP.2019.2947185

introduced a factorized projection operator in discriminative correlation filter to prune the parameters of basis filters. This operator can reduce the negative effective of redundant deep features in Fig. 1(b). However, it treats each feature map as isolate observation, ignoring the intrinsic correlation between different convolutional layers and the background noise in Fig. 1(c). This may reduce the accuracy of target regression.

In this letter, we propose a multi-layer factorized discriminant correlation filter (MLF-DCF) for robust visual tracking. The main contributions of our method are listed as follows.

- To reduce the negative effect of redundant deep features, the basis filter selection in MLF-DCF is achieved by integrating the multi-view discriminant learning and the discriminative correlation filter into a unified optimization problem. In this way we can simultaneously make pruned multi-layer basis filters highlight their responses of target sub-regions through minimizing the variation between similar feature maps in different convolutional layers.
- An ADMM algorithm is developed for solving the joint optimization problem in MLF-DCF, where each sub-problem has the closed-form solution.
- To eliminate the background noise of feature maps, we employ a revised SCDA method to estimate a mask to effectively extract the actively semantic representation of target. This can enhance the generality of the pre-trained CNN model. The examples of masked feature maps are shown in Fig. 1(d).

II. Factorized Correlation Filter

A. Traditional Factorized Discriminant Correlation Filter

Suppose these exist a training set $B = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{M}$, where each training sample $\mathbf{x}_i = [\mathbf{x}_i^1, \ldots, \mathbf{x}_i^d, \ldots, \mathbf{x}_i^D]$ is composed of D feature maps, and $\mathbf{y}_i$ is the predefined Gaussian shaped labels. The ECO method can eliminate the negative effect of redundant feature maps through using a factorized projection to prune the parameters of basis filters. This method is formulated as [16]

$$\min_{\mathbf{f},\mathbf{P}} \sum_{i=1}^{M} \alpha_i \|S_{Pf}(\mathbf{x}_i) - \mathbf{y}_i\|_2^2 + \sum_{c=1}^{C} \|\mathbf{w} \cdot \mathbf{f}^c\|_2^2 + \lambda \|\mathbf{P}\|_F^2, \quad (1)$$

where $\mathbf{f} = \{\mathbf{f}^1, \ldots, \mathbf{f}^C\}$ denotes the multi-channel basis filter set, $\mathbf{P} = (\mathbf{P}_{d,c}) \in \mathbb{R}^{D \times C} (C < D)$ denotes a projection matrix and S_{Pf} denotes the factorized projection operator, which is calculated as that

$$S_{Pf}(\mathbf{x}_i) = \mathbf{f} * \mathbf{P}^T \mathbf{J}(\mathbf{x}_i) = \sum_{c,d} \mathbf{P}_{d,c} \mathbf{f}^c * J_d(\mathbf{x}_i^d, t), \quad (2)$$

with

$$J_d(\mathbf{x}_i^d, t) = \sum_{n=0}^{N_d - 1} \mathbf{x}_i^d[n] b_d \left(t - \frac{T}{N_d} n \right), \quad (3)$$

where $J_d(\mathbf{x}_i^d, t)$ is an interpolation operator aiming at transferring multi-channel feature maps $\mathbf{x}_i^d$ to the continuous spatial domain $t \in [0, T)$. In Eq. (2), the operator $S_{Pf}(\mathbf{x}_i)$ is aimed to prune the parameters of basis filters (reduce the size of basis filter set) through using projection $\mathbf{P}$ to reduce the dimension of feature map $\mathbf{J}(\mathbf{x}_i)$. The limitations of Eq. (1) are two-folds:

1) Feature maps in multi-channel often focus on different sub-regions of the target, this means that multi-channel features maps can be naturally clustered into groups. Eq. (1) treats each feature map as an isolate observation and could not explore the intrinsic group similarity in each training sample $\mathbf{x}_i$. This may lose some useful semantic information when reducing the dimension of feature maps.

2) $\{\mathbf{x}_i\}_{i=1}^{M}$ are obtained from pre-trained VGG-m net. The training dataset for VGG-m is ImagNet, which is not built for the specific visual tracing task. Hence the gap between the training and the online tracking datasets may make $\{\mathbf{x}_i\}_{i=1}^{M}$ involve background noise. The background noise may reduce the accuracy of target regression.

B. The Proposed Method

Eq. (1) is the factorized correlation filter model for a single convolutional layer. In this letter, we propose a MLF-DCF model to simultaneously optimize the factorized correlation filters in multiple convolutional layers. The MLF-DCF model is formulated as that

$$\min_{\mathbf{f},\mathbf{P}} \sum_{k=1}^{L} \left(\sum_{i=1}^{M} \alpha_i^k \|S_{Pf}(\bar{\mathbf{x}}_i^k) - \mathbf{y}_i^k\|_2^2 + \sum_{c=1}^{C} \|\mathbf{w}^k \cdot (\mathbf{f}^k)^c\|_2^2 \right) + \lambda Tr(\mathbf{P}^T(\mathbf{B} - \mathbf{S})\mathbf{P}), \quad (4)$$

where $\mathbf{P} = [(\mathbf{P}^1)^T, \ldots (\mathbf{P}^k)^T, \ldots, (\mathbf{P}^L)^T]^T$ and $\mathbf{P}^k$ denotes the projection matrix that can reduce the dimension of feature maps in the k-th convolutional layer, $(\mathbf{f}^k)^c$ denotes the c-th basis filter in the k-th layer and $\bar{\mathbf{x}}_i^k$ is the i-th training sample in the k-th layer after using group mask to highlight its useful information. Parameter matrices $\mathbf{B}$ and $\mathbf{S}$ are used to evaluate the within group variation and between group variation from both inter-view and intra-view [19].

It is observed from [20] and [21] that in a certain convolutional layer, a convolutional feature channel often corresponds to a certain type of visual pattern. Dividing feature channels into different groups can yield multi-attention target representation. On the other hand, in VGG net, feature maps of early layers contain more spatial information of target, while feature maps of latter layers capture the semantic information of target, which are robust to the target appearance change. Both feature maps of early and latter layers are useful for visual tracking. Considering aforementioned observations, feature maps of different layers can be considered as the observations of the same target from different viewpoints. And if the feature maps are clustered into groups, the same observation groups in different views contain inherent correlation. Based on above considerations, the motivation of Eq. (4) is to highlight active response and effectively reduce the redundant information of multi-layer multi-channel feature maps through simultaneously exploiting the group similarity that within a certain convolutional layer and between different convolutional layers. To achieve this purpose, we first divide $\{\bar{\mathbf{x}}_i^k\}_{i=1}^{M}$ into n groups, then refer to multi-view discriminant analysis [22] and [19] to calculate $\mathbf{B}$ and $\mathbf{S}$. In this way we can exploit the multi-layer group similarity to restrict projection $\mathbf{P}$.

The advantages of Eq. (4) over Eq. (1) include: i) Eq. (1) only can reduce the redundant information of deep features in a single convolutional layer. Different from Eq. (1), Eq. (4) uses the regularizer $Tr(\mathbf{P}^T(\mathbf{B} - \mathbf{S})\mathbf{P})$ to maximize the between-group variations and minimize the within-group variations in both inter-view and intra-view. This can simultaneously highlight the discriminative information of multi-layer feature maps, making the pruned multi-layer basis filters more class discriminative. ii) For a certain layer, we do not directly use $\mathbf{x}_i^k$ such as Eq. (1) to train the basis filters. Instead, we use the masked training sample $\bar{\mathbf{x}}_i^k$ to train the correlation filters. Specifically, since the feature maps in $\mathbf{x}_i^k$ can be naturally grouped, we firstly propose a revised SCDA method to estimate group masks for different groups in $\mathbf{x}_i^k$. Then, we use group masks to refine original feature map groups. After orderly refining each feature map group, we can obtain masked $\bar{\mathbf{x}}_i^k$. In aforementioned process, to estimate a group mask, our revised SCDA method firstly employs [23] to combine all the feature maps in a feature map group together to generate an aggregation map, then refers to [20] to use log-likelihood function to smooth the aggregation map, yielding the final group mask.

C. Reconstruction Method

To reduce the computational complexity of reconstruction method, we propose an ADMM method to solve Eq. (4), where we firstly use the Parsevals theorem to transform Eq. (4) into Fourier domain and obtain

$$\begin{aligned}\min_{\hat{\mathbf{f}},\mathbf{P}} \sum_{k=1}^{L} \left(\sum_{i=1}^{M} \alpha_i^k \|(\hat{\mathbf{z}}^k)^T \mathbf{P}^k \hat{\mathbf{f}}^k - \mathbf{y}_i^k\|_2^2 + \sum_{c=1}^{C} \|\hat{\mathbf{w}}^k * (\hat{\mathbf{f}}^k)^c\|_2^2 \right) \\ + \lambda Tr(\mathbf{P}^T(\mathbf{B}-\mathbf{S})\mathbf{P}),\end{aligned} \tag{5}$$

where $\hat{\mathbf{z}}^k$ denotes the discrete Fourier transform of $\mathbf{J}(\bar{\mathbf{x}}_i^k)$.

Then, we introduce a variable $\mathbf{H}$ into Eq. (5) and set $\mathbf{P} = \mathbf{H}$. After that, the Augmented Lagrangian function of Eq. (5) is that

$$\begin{aligned}L(\hat{\mathbf{f}},\mathbf{P},\mathbf{H}) = \sum_{k=1}^{L} \Bigg(& \sum_{i=1}^{M} \alpha_i^k \|(\hat{\mathbf{z}}^k)^T \mathbf{P}^k \hat{\mathbf{f}}^k - \mathbf{y}_i^k\|_2^2 \\ & + \sum_{c=1}^{C} \|\hat{\mathbf{w}}^k * (\hat{\mathbf{f}}^k)^c\|_2^2 + \frac{\mu}{2}\|\mathbf{P}^k - \mathbf{H}^k - \frac{1}{\mu}\mathbf{\Lambda}^k\|_F^2 \Bigg) \\ & + \lambda Tr(\mathbf{H}^T(\mathbf{B}-\mathbf{S})\mathbf{H}),\end{aligned} \tag{6}$$

where μ and $\mathbf{\Lambda}^k$ are the Lagrangian parameters.

Thirdly, we propose to use alternate strategy to alternately optimize three sub-problems: **P**-Step, **H**-Step and **F**-Step. For a single training sample $\hat{\mathbf{x}}^k$, the **P**-Step is formulated as

$$\min_{\mathbf{P}} \sum_{k=1}^{L} \alpha^k \|(\hat{\mathbf{z}}^k)^T \mathbf{P}^k \hat{\mathbf{f}}^k - \mathbf{y}^k\|_2^2 + \frac{\mu}{2} \left\| \mathbf{P}^k - \mathbf{H}^k - \frac{1}{\mu}\mathbf{\Lambda}^k \right\|_F^2 . \tag{7}$$

Setting the partial derivative of Eq. (7) with respect to $vec(\mathbf{P}^k)$ to zero, we can obtain

$$\begin{aligned}vec(\mathbf{P}^k) = \big(2\alpha^k \mathbf{M}^k (\mathbf{M}^k)^T + \mu \mathbf{I}\big)^{-1} \big(\mu \cdot vec(\mathbf{H}^k) \\ + vec(\mathbf{\Lambda}^k) + 2\alpha^k \mathbf{M}^k \mathbf{y}^k\big),\end{aligned} \tag{8}$$

where $\mathbf{M}^k = \hat{\mathbf{f}}^k \otimes \hat{\mathbf{z}}^k$ with $\otimes$ meaning Kronecker product, $vec(\cdot)$ denotes the vectorization operator. Eq. (8) is obtained due to $(\hat{\mathbf{z}}^k)^T \mathbf{P}^k \hat{\mathbf{f}}^k = (\hat{\mathbf{f}}^k \otimes \hat{\mathbf{z}}^k)^T vec(\mathbf{P}^k)$.

The **H**-Step is formulated as

$$\min_{\mathbf{H}} \sum_{k=1}^{L} \frac{\mu}{2} \left\| \mathbf{H}^k - \mathbf{P}^k + \frac{1}{\mu}\mathbf{\Lambda}^k \right\|_F^2 + \lambda Tr(\mathbf{H}^T(\mathbf{B}-\mathbf{S})\mathbf{H}), \tag{9}$$

$\|\mathbf{H}^k + \frac{1}{\mu}\mathbf{\Lambda}^k - \mathbf{P}^k\|_F^2$ is upper bounded by $\|\mathbf{H}^k\|_F^2$+$\|\frac{1}{\mu}\mathbf{\Lambda}^k - \mathbf{P}^k\|_F^2$. Based on this observation, we can simplify Eq. (9) as

$$\min_{\mathbf{H}} Tr\big(\mathbf{H}^T(\lambda\mathbf{B} - \lambda\mathbf{S} + \mathbf{I})\mathbf{H}\big), \tag{10}$$

Eq. (10) can be solved by eigenvalue decomposition.

Inspired by [15], the **F**-Step is formulated as

$$\min_{\mathbf{f}} \sum_{k=1}^{L} \alpha^k \|(\hat{\mathbf{z}}^k)^T \mathbf{P}^k \hat{\mathbf{f}}^k - \mathbf{y}^k\|_2^2 + \|\mathbf{W}\hat{\mathbf{f}}^k\|_2^2, \tag{11}$$

where Toeplitz matrix corresponding to convolution operator is denoted as $\mathbf{W}^d(\hat{\mathbf{f}}^k)^c = vec(\hat{\mathbf{w}}^k * (\hat{\mathbf{f}}^k)^c)$ and $\mathbf{W}$ is the block-diagonal matrix $\mathbf{W} = \mathbf{W}^1 \oplus \mathbf{W}^2 \ldots \oplus \mathbf{W}^C$. Setting the partial derivative of Eq. (11) to zero, we can obtain

$$\big(\alpha^k (\mathbf{A}^k)^H (\mathbf{A}^k) + \mathbf{W}^H \mathbf{W}\big)\hat{\mathbf{f}}^k = (\mathbf{A}^k)^H \mathbf{y}^k, \tag{12}$$

where $\mathbf{A}^k = (\hat{\mathbf{z}}^k)^T \mathbf{P}^k$

III. Experiments

In this section, we use two datasets OTB-50 [24] and OTB-100 [25] to test the tracking performance of our MLF-DCF method. Here, four objective measures (position error, overlap rate, precision plot and success plot) are used to evaluate the quantitative tracking performance. The position error is defined as the Euclidean distance between the central location of the tracked bounding box and the manually labeled ground truth. The overlap rate is defined as $\frac{area(B_T \cap B_G)}{area(B_T \cup B_G)}$, where B_T and B_G are the tracked bounding box for each video frame and the corresponding ground truth, respectively. The precision plot indicates accumulated position errors under different location error thresholds. The success plot reflects the accumulated success rates versus different overlap thresholds and the success rate counts the number of video frames whose the overlap rate is larger than 0.5. In this experiment, we use 11 state-of-the-art methods as comparison to test the tracking performance of our methods. The selected tracking methods include: C-COT [15], ECO [16], DSST [26], KCF [11], SiamFC [27], MCPF [13], VITAL [7], PTAV [28], HDT [29], CF2 [21], MCCT [18]. Specifically, C-COT, ECO, MCPF and MCCT are the top trackers that use the combination of deep feature and discriminant correlation filter to achieve visual tracking. SiamFC, CF2. HDT and VITAL are four well known deep learning based trackers. DSST, KCF and PTAV are the correlation filter based trackers with handcraft features.

Experiment setting: In our method, we do the same as ECO that uses the Conv-1 and Conv-5 convolutional layers in the VGG-m network to achieve visual tracking. The original number of basis filters for Conv-1 and Conv-5 layers are 96 and 512, respectively. Since $\mathbf{P}^1$ and $\mathbf{P}^2$ have the same size, we empirically reduce both of basis filters in Conv-1 and Conv-5 to 30. The parameters λ and μ in the reconstruction method are empirically set as $\lambda = 1$, $\mu = 0.1$. In online visual tracking, we update the parameters of basis filters every 6 frames.

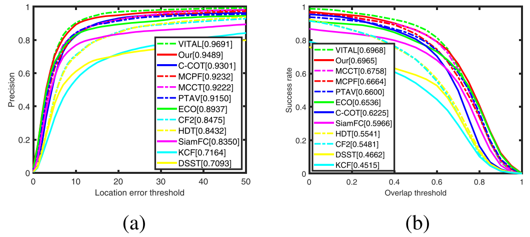

Fig. 2. The overall tracking performance in OTB-50 dataset: (a) Precision plot (b) Success plot. The distance precision and the AUC score are shown in the legend of precision and success plots, respectively.

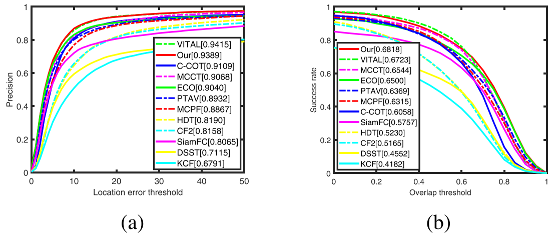

Fig. 3. The overall tracking performance in OTB-100 dataset: (a) Precision plot (b) Success plot. The distance precision score and the AUC score are shown in the legend of precision and success plots, respectively.

TABLE I
MEAN VALUE OF OVERLAP RATE OVER DIFFERENT VIDEO SUBSETS. THE BEST TWO RESULTS ARE DENOTED AS RED AND BLUE

Meth. Seq.	Our	ECO	C-COT	DSST	KCF	MCCT	SiamFC	MCPF	VITAL	PTAV	HDT	CF2
IV	0.71	0.66	0.59	0.42	0.39	0.67	0.58	0.66	0.69	0.65	0.52	0.51
SV	0.68	0.65	0.57	0.40	0.37	0.65	0.56	0.64	0.67	0.63	0.51	0.50
OCC	0.67	0.65	0.63	0.42	0.37	0.63	0.57	0.63	0.66	0.63	0.49	0.48
DEF	0.66	0.61	0.58	0.36	0.35	0.62	0.55	0.63	0.66	0.57	0.47	0.47
MB	0.68	0.67	0.61	0.38	0.39	0.66	0.56	0.62	0.66	0.60	0.51	0.52
FM	0.69	0.68	0.61	0.42	0.41	0.66	0.57	0.64	0.66	0.62	0.52	0.51
IPR	0.66	0.63	0.60	0.40	0.40	0.64	0.54	0.59	0.65	0.62	0.50	0.42
OPR	0.68	0.63	0.57	0.41	0.41	0.64	0.53	0.60	0.66	0.58	0.48	0.46
OV	0.63	0.64	0.57	0.38	0.30	0.64	0.54	0.60	0.66	0.60	0.41	0.45
BC	0.68	0.62	0.58	0.47	0.44	0.66	0.53	0.63	0.70	0.68	0.54	0.51
LR	0.64	0.63	0.57	0.35	0.30	0.68	0.63	0.60	0.66	0.51	0.40	0.39
Average	0.67	0.64	0.59	0.40	0.38	0.65	0.56	0.62	0.66	0.61	0.49	0.47

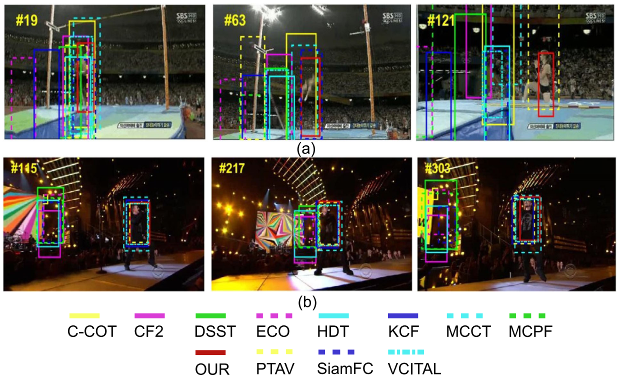

Fig. 4. The qualitative results on two video sequences in OTB-100 dataset. (a) Jump (b) Singer2.

TABLE II
FPS PERFORMANCE FOR DIFFERENT TRACKERS

Tracker	Our	ECO	VITAL	KCF	SiamFC	PTAV	MCPF	HDT	C-COT	DSST	MCCT
FPS	3.2	3.6	0.3	49.5	30.2	3.7	0.2	1.0	0.1	11.8	2.2

Evaluation on OTB-50 dataset: OTB-50 dataset contains 50 challenging video sequences. The tracking experiments on this dataset are shown in Fig. 2. From Fig. 2(a) and (b), we could see that, our method ranks second on both distance precision and AUC scores, outperforming ECO by 5% and 4%, respectively. Although VITAL gives slightly higher distance precision and AUC scores than our method, it requires expensive adversarial learning to enrich the training data. Different from VITAL, we only use the labeled candidates obtained from the first video frame to train the correlation filter. This can obviously reduce the computational complexity of tracking process.

Evaluation on OTB-100 dataset: Here, we use OTB-100 to carry out the experiments (see Fig. 3). OTB-100 dataset contains 100 video sequences, which is more challenging than OTB-50. From Fig. 3(b) we could see that our method gives the highest AUC score on success plots, outperforming ECO by 4%. Even compared with VITAL, we can outperform it by 1.4% in AUC performance.

In the OTB-100 dataset, the 100 video sequences are tagged by 11 attributes, which indicate the challenging aspects in visual tracking. Those 11 attributes include: Illumination Variation (IV), Scale Variation (SV), Occlusion (OCC), Deformation (DEF), Motion Blur (MB), Fast Motion (FM), In-Plane Rotation (IPR), Out-Plane Rotation (OPR), Out-of-View (OV), Background Clutters (BC) and Low Resolution (LR). We refer to [25] to divide 100 video sequences into 11 subsets, and give the overlap rate evaluation on different subsets (see Table I). From this test we can see that our method can give obviously higher overlap rate score than ECO, especially in IV, DEF, OPR and BC scenarios. This test can indicate the advantage of reducing the redundant deep features through exploring multi-layer group similarity.

Besides aforementioned quantitative evaluation, we also give qualitative evaluation on different trackers in Fig. 4. Jump video sequence is very challenging because it contains several adverse factors such as scale variation, fast motion and deformation. From Fig. 4(a) we could clearly see that our method can still track the athlete using an appropriate bounding box. This performance obviously outperforms other 11 methods. Besides jump sequence, we also give the tracking performance of singer 2 sequence in Fig. 4(b). From Fig. 4(b) we could see that our method can give similar tracking performance as MCCT and VITAL when facing severe illumination change and background clutter.

Finally, we evaluate the tracking speed of different methods. The average FPS is carry out on a desk with Inter(R) Core(TM) i3-2310 M CPU @ 2.10 Hz (2 GB RAM) (see Table II). From this test we can see that our MLF-DCF model does not involve higher computational complexity than ECO.

IV. CONCLUSION

In this letter, we have proposed a multi-layer factorized discriminant correlation filter for visual tracking. This method can effectively prune multi-layer basis filters through simultaneously exploring the group similarity that within a certain convolutional layer and between different convolutional layers. Compared with state-of-the-art trackers such as [15], [16] and [18], the extensive experiments show the superior performance of our method in challenging video sequences. Our future work is to extend MLF-DCF to temporal domain, exploring the multi-layer multi-frame partially attentional feature map correlation.

REFERENCES

[1] S. Zhang, X. Yu, Y. Sui, S. Zhao, and L. Zhang, "Object tracking with multi-view support vector machines," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 265–278, Mar. 2015.

[2] F. Liu, Z. Tao, G. Chen, K. Fu, B. Li, and Y. Jie, "Inverse nonnegative local coordinate factorization for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1752–1764, Aug. 2018.

[3] Y. Zheng, L. Sun, S. Wang, J. Zhang, and J. Ning, "Spatially regularized structural support vector machine for robust visual tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3024–3034, Oct. 2019.

[4] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2259–2272, Nov. 2011.

[5] Y. Qi, L. Qin, J. Zhang, S. Zhang, Q. Huang, and M. H. Yang, "Structure-aware local sparse coding for visual tracking," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3857–3869, Aug. 2018.

[6] T. Zhang, A. Bibi, and B. Ghanem, "In defense of sparse tracking: Circulant sparse tracker," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3880–3888.

[7] Y. Song *et al.*, "VITAL: Visual tracking via adversarial learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8990–8999.

[8] B. Li, W. Wu, Z. Zhu, and J. Yan, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8971–8980.

[9] Q. Wang, Z. Teng, J. Xing, J. Gao, and S. Maybank, "Learning attentions: Residual attentional siamese network for high performance online visual tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4854–4863.

[10] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019.

[11] 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.

[12] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1387–1395.

[13] T. Zhang, C. Xu, and M. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4819–4827.

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

[15] 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. Vision*, 2016, pp. 472–488.

[16] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6931–6939.

[17] L. Feng, T. Cheng, W. Zuo, Z. Lei, and M. H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4904–4913.

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

[19] M. Kan, S. Shan, H. Zhang, and S. Lao, "Multi-view discriminant analysis," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 188–194.

[20] H. Zheng, J. Fu, M. Tao, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5219–5227.

[21] C. Ma, J. Huang, X. Yang, and M. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3074–3082.

[22] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2016.

[23] X. S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2868–2881, Jun. 2017.

[24] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 2411–2418.

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

[26] M. Danelljan, G. Hger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vision Conf.*, 2015, pp. 1–5.

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

[28] H. Fan and H. Ling, "Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5487–5495.

[29] Y. Qi, S. Zhang, Q. Lei, H. Yao, and M. H. Yang, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4303–4311.