Densely Connected Discriminative Correlation Filters for Visual Tracking

Cheng Peng, Fanghui Liu, Jie Yang, and Nikola Kasabov, *Fellow, IEEE*

***Abstract*—Discriminative Correlation Filters (DCFs)-based approaches have recently achieved competitive performance in visual tracking. However, such conventional DCF-based trackers often lack the discriminative ability due to the shallow architecture. As a result, they can hardly tackle drastic appearance variations and easily drift when the target suffers heavy occlusions. To address this issue, a novel densely connected DCFs framework is proposed for visual tracking. We incorporate multiple nested DCFs into the deep learning architecture, and then train the compact network with the data-specific target. Specifically, feature maps and interim response maps are shared and reused throughout the whole network. By doing so, the implicit information carried out by each DCF is fully exploited to enhance the model representation ability during the tracking process. Moreover, a multiscale estimation scheme is developed to account for scale variations. Experimental results on the benchmarks demonstrate that the proposed approach achieves outstanding performance compared to the existing state-of-the-art trackers.**

***Index Terms*—Correlation filters, deep learning, visual tracking.**

I. Introduction

VISUAL tracking is one of the most important research topics in computer vision with various applications [1]–[4]. It aims to track a target, only the initial location of which is given in a video sequence. The problem is quite challenging due to several factors including occlusions (OCC), deformation (DEF), and scale variations (SV) of the target.

In general, tracking algorithms can be either generative or discriminative. Generative methods typically search for the best image candidate with the minimal reconstruction error [5]–[9]. For example, Li *et al.* [6] construct object representation via three-dimensional discrete cosine transform. A spatial-color mixture of Gaussians appearance model is proposed by Wang *et al.* [5] to measure the similarity of patches. Comparably, the representative discriminative methods are Discriminative Correlation Filters (DCFs)-based trackers [10]–[18]. They aim to learn a correlation filter from training samples and distinguish the target from backgrounds. Bolme *et al.* [10] first introduce correlation filters into visual tracking. Numerous efforts have been dedicated to its improvement, e.g., by exploiting the circulant structure of image patches [12], utilizing multiple kinds of features [14], and incorporating the structured output support vector machine [16]. Although much progress has been made, these DCF-based trackers are often not robust to appearance variations and thus achieve unsatisfactory performance due to a simple or shallow architecture.

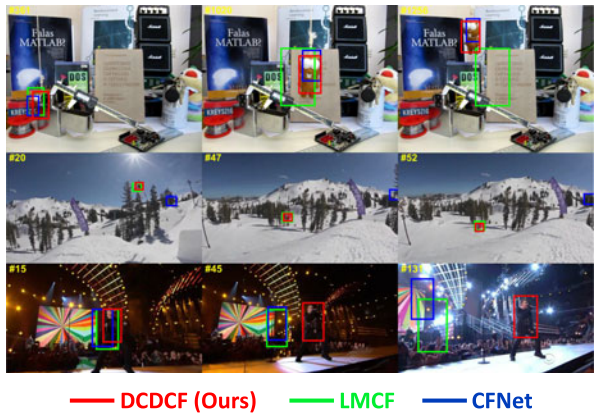

Fig. 1. Comparisons of the proposed DCDCF with the other state-of-the-art trackers. The representative frames are from the *lemming*, *skiing*, and *singer2* sequences. The results demonstrate that our DCDCF outperforms LMCF [16] (based on DCFs) and CFNet [19] (based on deep learning) in challenging situations of occlusions and deformation.

As demonstrated by the success use of deep learning [20]–[22], researchers develop several algorithms that incorporate the deep learning architecture into the tracking framework [19], [23]–[26]. Hong *et al.* [23] construct target-specific saliency maps based on a pretrained convolutional neural network. Valmadre *et al.* [19] apply an end-to-end network to learn tightly coupled deep features. Considering that there are extremely limited labeled data (the first frame in a video sequence) for training, these deep-learning-based trackers usually train their networks on large-scale datasets associated with object detection and classification [27], [28].[1] In this case, we have to be encountered with such task inconsistency. That is, the learned models in the above algorithms cannot adequately meet the requirement of a tracking task. And also, these algorithms equipped with such training strategy on large-scale datasets can hardly design an online updating scheme during the tracking process.

Motivated by the above observations, we propose a novel Densely Connected Discriminative Correlation Filters

Manuscript received March 13, 2018; revised May 3, 2018; accepted May 9, 2018. Date of publication May 15, 2018; date of current version June 7, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61572315 and Grant 6151101179; and in part by the 973 Plan of China under Grant 2015CB856004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mireille Boutin. *(Corresponding author: Jie Yang.)*

C. Peng, F. Liu, and J. Yang are with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: pc1899@outlook.com; lfhsgre@outlook.com; jieyang@sjtu.edu.cn).

N. Kasabov is with the Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Auckland 1010, New Zealand (e-mail: nkasabov@aut.ac.nz).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2018.2836360

[1]The visual object tracking committee has prohibited training and testing deep models for tracking using datasets from the same domain.

(DCDCF) tracking framework. To the best of our knowledge, this letter is the first to introduce the densely connected architecture [22] into the visual tracking community. It extends the dimension of the single DCF, and integrates multiple DCFs with dense connections into the deep learning architecture. Such structure of dense connections guarantees that the feature maps and interim response maps can be shared and reused throughout the whole network. As a result, the implicit information of the target are comprehensively exploited by various filters. And specifically, our deep network is trained by only single image, i.e., the first frame in a sequence. Accordingly, an online updating scheme is naturally developed to retrain the DCDCF network, which effectively learns the data-specific target appearance. Besides, we utilize a multiscale estimation scheme with the Peak-to-Sidelobe Ratio (PSR) as the confidence criterion, which helps to alleviate the drifting problem caused by SV. Fig. 1 provides a comparison of our tracker with other methods based on DCFs and deep learning. Experimental results on OTB-13 [29] and OTB-15 [30] benchmark datasets demonstrate that DCDCF outperforms the other state-of-the-art trackers.

II. Proposed Approach

In this section, we first review the DCF-based tracker which is closely related to this work, and then introduce the proposed DCDCF method in detail.

A. Review: DCF-Based Trackers

In the tth frame ($t \geqslant 1$), the DCF-based tracker aims to learn a correlation filter $\mathbf{h}_t$ based on a given training sample $\mathbf{x}_t$. The sample consists of D-channel feature maps $\{\mathbf{x}_t[d]\}_{d=1}^{D}$ extracted from a image patch. The learned correlation filter $\mathbf{h}_t$ minimizes the ℓ_2 loss between the output and the desired response $\bar{\mathbf{y}}$, which is defined as

$$\mathcal{L}(\mathbf{x}_t;\mathbf{h}_t) = \left\| \sum_{d=1}^{D} \mathbf{h}_t[d] \star \mathbf{x}_t[d] - \bar{\mathbf{y}} \right\|^2 + \lambda \sum_{d=1}^{D} \|\mathbf{h}_t[d]\|^2 \quad (1)$$

where λ is a regularization parameter and $\star$ refers to the convolution operator. The solution can be obtained in closed form by utilizing the Discrete Fourier Transform (DFT) as

$$\hat{\mathbf{h}}_t[d] = \frac{\hat{\bar{\mathbf{y}}}^* \odot \hat{\mathbf{x}}[d]}{\sum_{d=1}^{D} \hat{\mathbf{x}}[d]^* \odot \hat{\mathbf{x}}[d] + \lambda}, d \in \{1,\ldots,D\}. \quad (2)$$

Herein, $\hat{\mathbf{x}}$ is the DFT of $\mathbf{x}$, i.e., $\hat{\mathbf{x}} = \mathcal{F}(\mathbf{x})$, $\mathbf{x}^*$ represents the complex conjugation of $\mathbf{x}$, and the notation $\odot$ denotes the elementwise product.

In the $(t+1)$th frame, the target location can be estimated by searching for the maximal value in the response map $\mathbf{y}_{t+1}$, which arrives at

$$\mathbf{y}_{t+1} = \mathcal{F}^{-1}\left(\sum_{d=1}^{D} \hat{\mathbf{h}}_t[d] \odot \hat{\mathbf{x}}_{t+1}[d] \right) \quad (3)$$

where $\mathcal{F}^{-1}$ denotes the inverse DFT.

B. DCDCF Network

Motivated by [22], we introduce the densely connected network to visual tracking and propose DCDCF. Fig. 2 shows the structure of the proposed DCDCF network (the dashed box). The designed network consists of five DCFs. Structures with fewer DCFs are not deep or broad enough to capture the target appearance variations, while more DCFs lead to overfitting problems.

In the proposed network, the information of both input feature maps and interim response maps flows to all subsequent filters. Formally, assume that the original feature input is denoted as $\mathbf{x}^{(1)}$. The input $\mathbf{x}^{(i)}$ of the ith DCF on the main axis, which consists of D_i-channel feature maps $\{\mathbf{x}^{(i)}[d]\}_{d=1}^{D_i}$, can be concatenated as

$$\mathbf{x}^{(i)} = \Theta\left(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(i-1)}\right), i \in \{2,3\} \quad (4)$$

where the function Θ denotes the concatenation operation. Besides, we connect the input features and interim maps to the final response map, which ensures the maximum functionality of each DCF from the network. These direct connections also have a regularizing effect.

Note that, compared to the conventional DCF in (2), we extend its dimension in the proposed DCDCF network. To be specific, the ith DCF $\mathbf{h}^{(i)}$ has L_i layers $\{\{\mathbf{h}^{(i)}[d,l]\}_{d=1}^{D_i}\}_{l=1}^{L_i}$ over D_i channels. The L_i-channel response maps $\{\mathbf{y}^{(i)}[l]\}_{l=1}^{L_i}$ of the ith DCF are then yielded as

$$\mathbf{y}^{(i)}[l] = \sum_{d=1}^{D_i} \mathbf{h}^{(i)}[d,l] \star \mathbf{x}^{(i)}[d]$$
$$l \in \{1,\ldots,L_i\}, i \in \{1,2,3\}. \quad (5)$$

Specifically, we formulate the DCF as convolution operations by sliding the filter $\mathbf{h}^{(i)}$ in the spatial domain, i.e.,

$$\mathbf{y}^{(i)}[l]|_p = \sum_{d=1}^{D_i} \mathbf{h}^{(i)}[d,l] \odot \Upsilon\left(\mathbf{x}^{(i)}[d]; p\right)$$
$$l \in \{1,\ldots,L_i\}, i \in \{1,2,3\}, p \in \mathcal{P}. \quad (6)$$

Herein, $\mathbf{y}^{(i)}[l]|_p$ represents the score located at p in $\mathbf{y}^{(i)}[l]$ and $\Upsilon(\mathbf{x};p)$ refers to the circular shifts of $\mathbf{x}$ at each location p in the position space $\mathcal{P}$.

C. Tracking Framework

The overall tracking procedure is detailed in this section. We extract feature maps from a search window in the tth frame and feed them to the DCDCF network. The translation response map $\mathbf{y}_t^{\text{trans}}$ is obtained to derive the target position p_t. The current scale s_t is then adaptively estimated through the multiscale estimation to achieve the final result.

1) Training: At the initialization stage, we extract D_1-channel feature maps $\mathbf{x}_1^{(1)}$ from a search window based on the ground truth in the first frame. The forward propagation of the DCDCF network is denoted by the function f. We adopt the Adam optimizer [31] to backward propagate the ℓ_2-loss errors between the yielded output and the desired Gaussian-shaped response $\bar{\mathbf{y}}$, namely,

$$\mathcal{L}\left(\mathbf{x}_1^{(1)};\mathbf{h}_1\right) = \sum_{p\in\mathcal{P}} \left\| f\left(\Upsilon\left(\mathbf{x}_1^{(1)};p\right),\mathbf{h}_1\right) - \bar{\mathbf{y}}|_p \right\|^2. \quad (7)$$

With $\mathbf{x}_1^{(1)}$ as input, the network is trained by minimizing $\mathcal{L}$ for optimizing the DCFs $\mathbf{h}_1^{(i)}, i \in \{1,\ldots,5\}$.

2) Translation Estimation and Updating: In the tth frame ($t \geqslant 2$), we extract features $\mathbf{x}_t^{(1)}$ of the search window in the

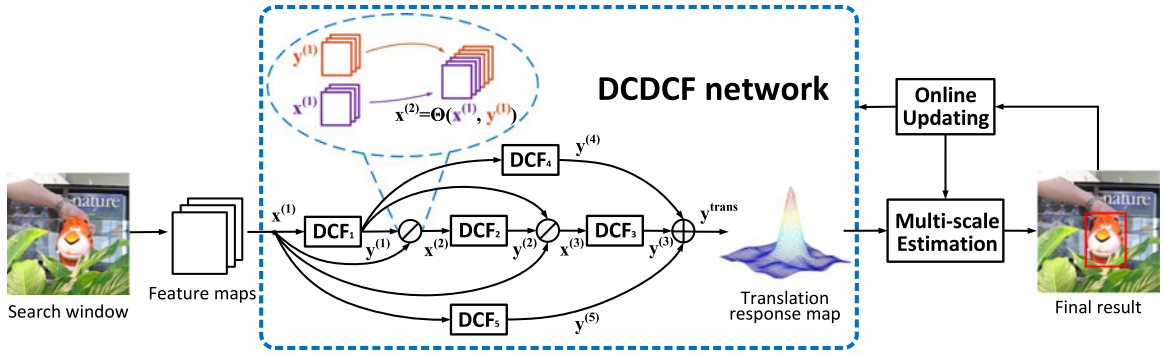

Fig. 2. Flowchart of the proposed DCDCF approach. The concatenation operation is denoted as $\oslash$. Based on the deep features extracted from the search window, the target is located by the proposed DCDCF network, and a multiscale scheme is developed for scale estimation.

same way based on the position p_{t-1} and scale s_{t-1} estimated in the $(t-1)$th frame. The feature maps are sent into the DCDCF net, and the translation response map $\mathbf{y}_t^{\text{trans}}$ can be obtained through the forward pass of the proposed architecture as

$$\mathbf{y}_t^{\text{trans}} = \sum_{i=3}^{5} \sum_{l=1}^{L_i} \mathbf{y}_t^{(i)}[l]. \tag{8}$$

Here, $\mathbf{y}_t^{(1)}, \mathbf{x}_t^{(2)}, \mathbf{y}_t^{(2)}$, and $\mathbf{x}_t^{(3)}$ can be successively derived based on $\mathbf{x}_t^{(1)}$ using (4) and (6) to achieve $\mathbf{y}_t^{(3)}$, while $\mathbf{y}_t^{(4)}$ and $\mathbf{y}_t^{(5)}$ can be obtained as

$$\mathbf{y}_t^{(4)}[l] = \sum_{d=1}^{L_1} \mathbf{h}_{t-1}^{(4)}[d,l] \star \mathbf{y}_t^{(1)}[d] \tag{9a}$$

$$\mathbf{y}_t^{(5)}[l] = \sum_{d=1}^{D_1} \mathbf{h}_{t-1}^{(5)}[d,l] \star \mathbf{x}_t^{(1)}[d]. \tag{9b}$$

The new position p_t of the target is estimated to locate at the maximum value in $\mathbf{y}_t^{\text{trans}}$, i.e.,

$$p_t = \arg\max_{p \in \mathcal{P}} f\left(\Upsilon\left(\mathbf{x}_t^{(1)}; p\right), \mathbf{h}_{t-1}\right). \tag{10}$$

For each frame, the DCDCF model is updated after detection in the similar way of training with a lower learning rate and fewer iterations.

3) Multiscale Estimation: To tackle SV, similar to [32], we use DCDCF at different scales to search the scale space $\mathcal{S}$ in parallel. Note that, PSR [33] is introduced as the confidence criterion of each scale result. The PSR value $\Gamma(\mathbf{y})$ of a response map $\mathbf{y}$ is defined as

$$\Gamma(\mathbf{y}) = \frac{\max(\mathbf{y}) - \mu_\Phi(\mathbf{y})}{\sigma_\Phi(\mathbf{y})}. \tag{11}$$

Herein, Φ refers to the part of $\mathbf{y}$ except the sidelobe area around the peak, the mean value, and standard deviation of which are μ_Φ and σ_Φ, respectively. In the tth frame, the target scale s_t is adaptively estimated from tracking results of different scales with the maximum PSR score.

III. Experimental Results

A. Implementation Details

We implemented the proposed tracker via MATLAB on a PC with an Intel Xeon E5-2695 CPU (2.30 GHz) and a GeForce GTX TITAN X GPU, and the tracker runs about 5.2 frames per second on average. All the parameters were fixed to all experimental sequences as follows.

Supposing that the initial target size was $M \times N$ pixels, the search window size was then set to $\max(4M, 4N) \times \max(4M, 4N)$. We adopted the trained VGG-Net [21] and utilized the outputs of the *conv3-4*, *conv4-4*, and *conv5-4* convolutional layers as feature maps, which was same to [13]. Principal Component Analysis was also applied to reduce the feature dimension D_1 to 64, which contributed to computational efficiency. The numbers of map channels and filter layers were $D_2 = 128$, $D_3 = 192$, $L_1 = L_2 = 64$, $L_3 = L_4 = L_5 = 1$, respectively. The spatial size of each filter was $(M - M \bmod 2 + 1) \times (N - N \bmod 2 + 1)$. The scale space $\mathcal{S}$ was set as $\{0.995, 1, 1.005\}$, which was similar to [32]. The sidelobe area was 10% of the response map around the peak.

B. Benchmarks and Evaluation Metrics

We implement experiments on the OTB-13 [29] and OTB-15 [30] benchmark datasets. All the video sequences are annotated with different attributes, namely, fast motion (FM), background clutter (BC), motion blur (MB), DEF, illumination variation (IV), in-plane rotation (IPR), low resolution (LR), OCC, out-of-plane rotation (OPR), out-of-view (OV), and SV.

The precision and accuracy of a tracker are quantitatively validated by center location error (CLE) and the success rate. CLE represents the average Euclidean distance between the center locations of B_T and B_G, where B_T and B_G refer to the tracked bounding box and the ground truth bounding box, respectively. The representative precision score for each tracker is chosen with the CLE threshold $= 20$ pixels in precision plots.

On the other side, the success rate is relevant to the overlap score, which is defined as $\frac{|B_T \cap B_G|}{|B_T \cup B_G|}$. Here, $|\cdot|$ denotes the number of pixels in the bounding box. The successful frames refer to those the overlap scores of which are larger than a threshold ε (e.g., $\varepsilon = 0.5$), and the success rate is the ratio of the number of successful frames to the total number. Trackers are ranked by the area under curve (AUC) of each success plot.

C. Comparisons With State-of-the-Art Trackers

We conduct quantitative and qualitative evaluations compared with the other state-of-the-art methods on benchmarks. In the following section, results and analysis in different aspects are discussed in detail.

TABLE I
RANKED AUC SCORES (%) OF THE SUCCESS RATES UNDER ELEVEN ATTRIBUTES

Tracker \ Attribute	FM (17)	BC (21)	MB (12)	DEF (19)	IV (25)	IPR (31)	LR (4)	OCC (29)	OPR (39)	OV (6)	SV (28)
DCDCF (Ours)	**64.4**	**67.5**	**65.9**	**69.3**	**65.8**	**64.2**	61.3	**66.2**	**66.1**	**70.2**	**63.2**
CFNet (2017) [16]	55.3	58.4	53.4	50.9	55.1	58.7	**63.2**	56.5	59.1	56.1	60.5
LMCF (2017) [13]	55.5	62.5	54.2	65.0	58.9	57.9	43.5	63.4	61.1	59.8	57.6
fDSST (2017) [14]	55.6	61.7	59.3	56.4	59.7	58.4	38.9	55.8	57.2	55.5	57.1
ACFN (2017) [22]	52.7	54.6	52.1	63.2	55.7	56.5	35.2	60.4	60.0	62.4	59.2
CSR-DCF (2017) [15]	53.3	55.9	55.2	62.5	57.0	55.4	34.2	57.1	58.1	56.4	52.8
SiamFC (2016) [21]	56.1	55.4	54.3	53.7	54.2	58.2	56.6	59.4	58.8	63.5	60.3
DLSSVM (2016) [31]	53.3	58.3	54.5	65.5	56.4	58.0	41.8	59.2	59.1	51.2	53.1
Staple (2016) [11]	50.1	55.7	52.6	60.7	56.1	57.6	39.5	58.5	56.9	51.8	54.5
HDT (2016) [32]	57.4	61.0	61.4	62.7	55.7	58.0	55.1	60.3	58.4	56.9	52.3
SRDCF (2015) [33]	56.9	58.7	60.1	63.5	57.6	56.6	42.6	62.7	59.9	55.5	58.7
HCFT (2015) [10]	57.8	62.3	61.6	62.6	56.0	58.2	55.7	60.6	58.7	57.5	53.1
RPT (2015) [30]	53.2	59.8	53.6	52.7	53.2	54.6	35.7	51.9	54.1	53.7	51.5

The number of sequences associated with the corresponding attribute is shown in parenthesis. The best performance in each column is indicated by **bold** fonts.

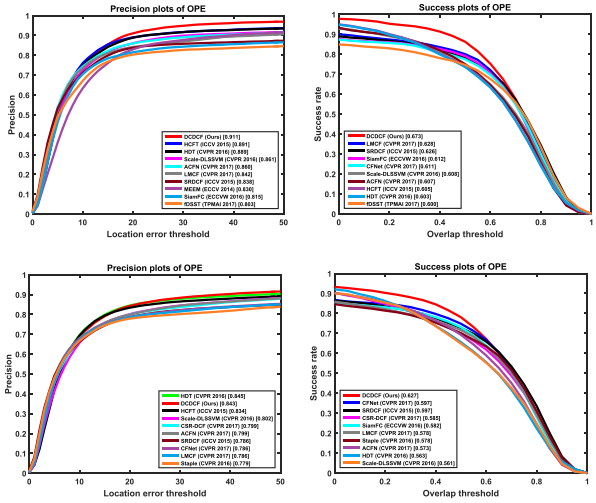

Fig. 3. Precision and success plots of OPE compared with other state-of-the-art approaches on OTB-13 [29] (first row) and OTB-15 [30] (second row).

1) Overall Performance: We compare the proposed DCDCF method with other 18 state-of-the-art trackers including CFNet [19], LMCF [16], ACFN [25], CSR–DCF [18], staple [14], fDSST [17], DLSSVM [34], HDT [35], SiamFC [24], struck [37], HCFT [13], KCF [12], RPT [33], SRDCF [36], TGPR [38], CN [11], SAMF [32], MEEM [39], and TLD [40]. Fig. 3 demonstrates the overall results under One Pass Evaluation (OPE) on OTB-13 [29] and OTB-15 [30]. Only the top 10 trackers are listed for presentation clarity. The proposed DCDCF obtains favorable results among most state-of-the-art trackers. It significantly outperforms trackers based on multiple DCFs (including SRDCF, HCFT, fDSST, and HDT) and methods based on deep learning (ACFN, CFNet, and SiamFC) in success rate on benchmarks.

2) Attribute-Based Evaluation: The attribute-based comparison on OTB-13 [29] between different methods in recent years (2015, 2016, and 2017) is shown in Table I. We show the AUC scores under 11 video attributes. The results demonstrate that the proposed DCDCF tracker outperforms others in most challenging situations, especially on OV, IV, and DEF attributes. It is mainly because that the densely connected structure naturally integrates the information carried out by all DCFs. Consequently, the learned discriminative model is compact and robust against drastic appearance variations.

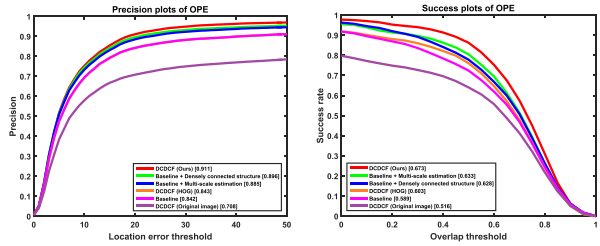

Fig. 4. Precision and success plots of OPE for the ablation analysis of the proposed approach on OTB-13 [29]. We take a cascade of three DCFs without scale estimation scheme as baseline.

3) Ablation Analysis: We provide an ablation analysis on OTB-13 [29] to evaluate the effectiveness of the key components from the proposed approach. The baseline tracker has a cascade of three DCFs without scale estimation scheme. Fig. 4 shows the precision and success plots of these six variants: the baseline tracker, the baseline tracker with multiscale estimation, the baseline tracker using the densely connected structure, the proposed DCDCF tracker with both components, DCDCF with original image as input, and DCDCF using histogram of oriented gradients [41] features. The result demonstrates that the components of the proposed DCDCF are significantly conducive to the improvement of final tracking performance.

IV. CONCLUSION

This letter introduces a DCDCF tracker that extends the dimension of the single DCF, and exploits the densely connected structure of multiple DCFs to visual tracking. Our framework has the capability of reusing all the feature maps and interim response maps from various filters, which significantly enhances the discriminative ability of the learned model. Besides, DCDCF is able to effectively capture the appearance variations thanks to our online updating scheme. Quantitative and qualitative evaluations on OTB-13 and OTB-15 demonstrate the effectiveness and robustness of the proposed DCDCF tracker when compared to the existing state-of-the-art trackers.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surveys*, vol. 38, no. 4, 2006, Art no. 13.

[2] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, Jul. 2014.

[3] Q. Jin, I. Grama, C. Kervrann, and Q. Liu, "Nonlocal means and optimal weights for noise removal," *SIAM J. Imag. Sci.*, vol. 10, no. 4, pp. 1878–1920, 2017.

[4] P. Liang, Y. Pang, C. Liao, X. Mei, and H. Ling, "Adaptive objectness for object tracking," *IEEE Signal Process. Lett.*, vol. 23, no. 7, pp. 949–953, Jul. 2016.

[5] H. Wang, D Suter, K Schindler, and C. Shen, "Adaptive object tracking based on an effective appearance filter," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1661–1667, Sep. 2007.

[6] X. Li, A. Dick, C. Shen, A. Van den Hengel, and H. Wang, "Incremental learning of 3D-DCT compact representations for robust visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 863–881, Apr. 2013.

[7] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparse collaborative appearance model," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2356–2368, May 2014.

[8] F. Liu, C. Gong, T. Zhou, K. Fu, X. He, and J. Yang, "Visual tracking via nonnegative multiple coding," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2680–2691, Dec. 2017.

[9] F. Liu, C. Gong, X. Huang, T. Zhou, J. Yang, and D. Tao, "Robust visual tracking revisited: From correlation filter to template matching," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2777–2790, Jun. 2018.

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

[11] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van De Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1090–1097.

[12] 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.

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

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

[15] C. Ma, Y. Xu, B. Ni, and X. Yang, "When correlation filters meet convolutional neural networks for visual tracking," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1454–1458, Oct. 2016.

[16] M. Wang, Y. Liu, and Z. Huang, "Large margin object tracking with circulant feature maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4800–4808.

[17] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2017.

[18] A. Lukežič, T. Vojář, L. Čehovin, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4847–4856.

[19] 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 Recog.*, 2017, pp. 5000–5008.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

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

[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4700–4708.

[23] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 597–606.

[24] 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. Vis. Workshops*, 2016, pp. 850–865.

[25] J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, and J. Y. Choi, "Attentional correlation filter network for adaptive visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4828–4837.

[26] J. Guo and T. Xu, "Deep ensemble tracking," *IEEE Signal Process. Lett.*, vol. 24, no. 10, pp. 1562–1566, Oct. 2017.

[27] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[28] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

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

[30] 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.

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[32] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2014, pp. 254–265.

[33] Y. Li, J. Zhu, and Steven C. H. Hoi, "Reliable patch trackers: Robust visual tracking by exploiting reliable patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 353–361.

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

[35] Y. Qi *et al.*, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 4303–4311.

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

[37] S. Hare *et al.*, "Struck: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, Oct. 2016.

[38] J. Gao, H. Ling, W. Hu, and J. Xing, "Transfer learning based visual tracking with gaussian processes regression," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 188–203.

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

[40] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[41] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, wno. 9, pp. 1627–1645, Sep. 2010.