IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 21, 2019, accepted December 11, 2019, date of publication December 16, 2019,
date of current version December 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2960105

Robust Object Tracking Using Affine Transformation and Convolutional Features

YINGHONG XIE[1], JIE SHEN[2], AND CHENGDONG WU[3]
[1]College of Information Engineering, Shenyang University, Shenyang 110044, China
[2]College of Engineering and Computer Science, University of Michigan—Dearborn, Dearborn, MI 48128, USA
[3]College of Faculty of Robot Science and Engineer, Northeastern University, Shenyang 110819, China

Corresponding author: Yinghong Xie (yinghxie@umich.edu)

This work was supported in part by the National Science Foundation of China under Grant 61503274, Grant 61603415, and Grant 61603080, in part by the Science and Technology Project of Shenyang city under Grant 18-013-0-15, and in part by the Double Hundred Project of Shenyang City under Grant Z18-5-013.

ABSTRACT The state-of-the-art trackers using deep learning technology have no special strategy to capture the geometric deformation of the target. Based on that the affine manifold can better capture the target shape change and that the higher level of Convolutional Neural Network (CNN) can better describe semantic information of objects, we propose a new tracking algorithm combining affine transformation with convolutional features to track targets with dramatic deformation. First, the affine transformation is applied to predict possible locations of a target, then a correlative filter is designed to compute the appearance confidence score for determining the final target location. Furthermore, a standard discriminative correlation filter is used to develop the effect of convolutional features, which is more efficient than other methods used for CNN Networks. Comprehensive experiments demonstrate the outstanding performance of our tracking algorithm compared to the state-of-the-art techniques in the public benchmarks.

INDEX TERMS Object tracking, CNN networks, affine manifold, geometric transformation, convolutional features.

I. INTRODUCTION

Visual object tracking is one of the fundamental tasks in computer vision with various applications from missile guidance and computer vision to autonomous driving.

The deformation modeling of the target is the key to obtain stable tracking result. Considering that affine manifold can better describe the geometric deformation of the target, the deformation models of visual tracking algorithms are largely built on the affine group. Reference [1] uses affine transformation to depict the deformation process of the target, merges the particle filter framework, and can better remove the background interference. Reference [2], [3] uses affine transformation to depict the deformation process of the target, and proposes an target tracking algorithm by using Riemannian Manifold geometry structure. In reference [4], the target trapezoid region is extracted by the model and then refined through the affine transformation. Based on the particle filtering-based tracking framework, the fusion of color and shape is used as the main feature for target tracking. Based on the affine transformation (GAT) correlation matching proposed by Wakahara et al [5], the paper developed an acceleration method for GPT correlation matching [6].

Although conventional tracking methods [7]–[16] using handcrafted feature have achieved computational efficiency and comparable tracking performance. Some milestones include IVT [17], MIL [18], TLD [19], SCM [20], STRUCK [21], ASLAS [22], APGL1 [23] and so on. However the performance of these conventional methods is far from the requirement of realistic application [23]–[25].

In the past few years, convolutional neural networks have significantly outperformed other state-of-art algorithms in many video processing problems, such as video surveillance [26] and object recognition [27], [28]. Convolutional filter has been widely used for visual tracking due to its high computational efficiency in Fourier domain. These kinds of tracking methods [18], [29] don't need hard-threshold samples of target appearance because they regress all the circular-shifted versions of input features to a Gaussian function. Bolme [30] modelled the target appearance by a minimum output sum of squared error filter for difficult tracking scenarios. Some efforts have been made to considerably

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

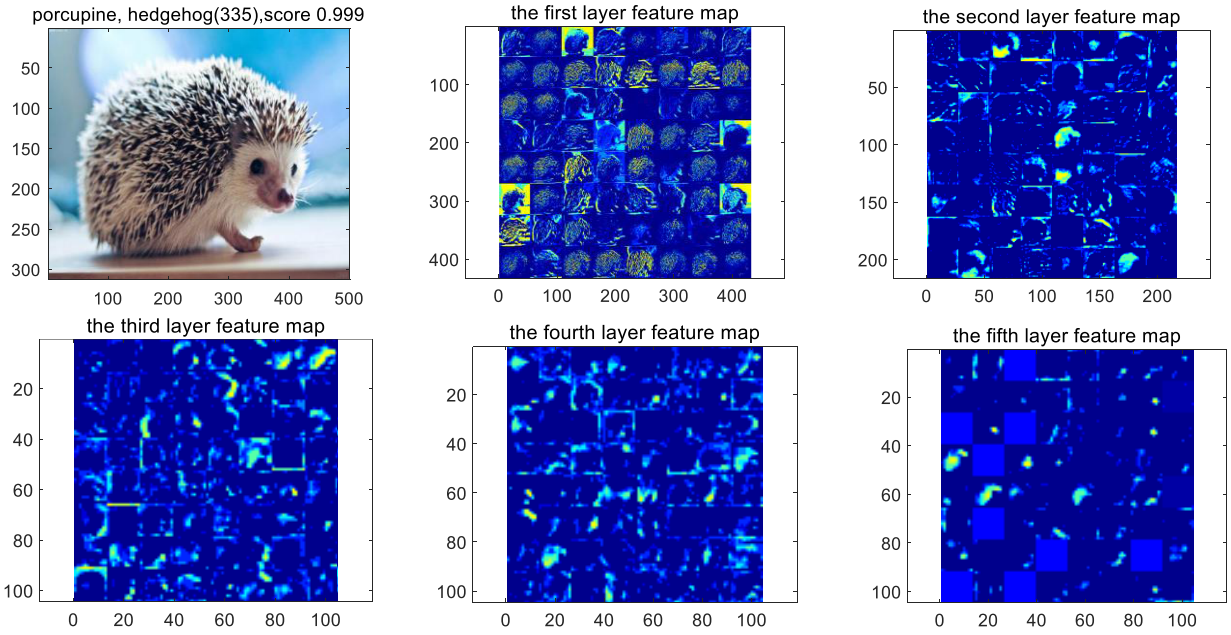


FIGURE 1. Visualization of the CNN features of an image. The first 64 principle components of features from the five layers are visualized.

improve tracking accuracy. The methods include kernelized correlation filters [31], [32], context learning [33], scale estimation [33], multiple dimensional features [34], [35], re-detection [12], short-term and long-term memory [36], and spatial regularization [37].

The current tracking methods based on convolutional neural network were mainly to train the tracking network in advance and design on discriminative or regression models [12], [31], [34], [38], [39]. The development of discriminative correlation filter is from single feature channel to multi feature channels. A high-speed tracking with kernelized correlation filters was proposed in [38] on the fact that the samples used to train the classifier are riddled with redundancies. Danelljan [12] presented a novel approach for robust scale estimation in a tracking-by-detection framework through learning discriminative correlation filters based on a scale pyramid representation. Bolme [34] presented a robust correlation filter which is minimum output sum of squared error filter. And a single frame is used to initialize the filter. Danelljan [31] proposed spatially regularized discriminative correlation filters to design a more discriminative feature model. And Kaihua Zhang [38] designed a light weight structure convolutional networks and achieved high speed performance for online tracking.

Because most existing CNN trackers learn online feature classifiers by training positive and negative samples according to the estimated location, two weaknesses lie in these methods. Firstly, only the output of the top layer is used to determine the tracking target, which is effective for target recognition, because the highest layer has the closest relationship to category-level semantics information while spatial resolution is gradually reduced with the increase of the depth of convolutional layers, and Figure 1 shows 64 feature maps for the first five layers, note that the special resolution is less useful when the layers get higher. But for video tracking problems, target location information is more important than semantics information for tracking target precisely. Furthermore, most of the existing convolutional methods have no special strategy to capture the geometric deformation of the target, and these methods can't change the size and shape of the tracking box to fit for the geometric change.

For dealing with the above issues, we apply affine manifold to capture target geometric transformation and the output of the highest layer of CNN network to describe the semantic information of target appearance in building a new tracker. The main advantages of the proposed tracker include

(1) The last convolutional layer is used to encode the semantic information of the objects, which is robust to significant appearance variations.
(2) The affine transformation is applied to predict possible locations of a target, which can be more accurately for the dynamical geometric deformation.
(3) By the combination of affine transformation and the last convolutional layer of correlative filters, both semantics and geometric deformation are simultaneously applied to handle large appearance and geometrical variations without drifting.
(4) Comprehensive experiments are conducted on a large benchmark dataset to demonstrate that the proposed method outperforms the existing tracking algorithms.

The rest of this paper is organized as follows: In section 2 we build affine manifold and the geometric transformation model. In section 3 we model a discriminative correlation filter for our tracker. And the steps of the tracking algorithm

are designed in section 4. Then, the next section describes the implementation details and evaluates the experimental effectiveness by comparing with other state-of-the-art trackers. Finally, conclusions are drawn in section 6.

II. AFFINE MANIFOLD AND THE GEOMETRIC TRANSFORMATION MODEL

A. AFFINE MANIFOLD AND ITS METRIC

In this paper, affine transformation is applied to represent the target geometric deformation.

Let $I(X)$ represent the gray value of the template image position $X = (x, y)'$. The Cartesian coordinate system is established with the center of the target as the coordinate origin. The gray value of the target in the input image after affine transformation is $I(W(x : r))$, where $W(x : r)$ represents the affine transformation of the object in the input image with respect to the template, $r = (r_1, r_2, r_3, r_4, r_5, r_6)'$ is a parameter vector.

$$W(x : r) = \begin{bmatrix} r_1 x + r_2 y + r_5 \\ r_3 x + r_4 y + r_6 \end{bmatrix} \tag{1}$$

The transformation matrix is represented with homogeneous coordinates as:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & r_5 \\ r_3 & r_4 & r_6 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

The affine transformation matrix

$$T(r) = \begin{bmatrix} r_1 & r_2 & r_5 \\ r_3 & r_4 & r_6 \\ 0 & 0 & 1 \end{bmatrix}$$

has the structure of Lie group $GA(2)$, and $ga(2)$ is Lie algebra corresponding to affine Lie group $GA(2)$. And matrix $G_i(\forall i \in \{1, 2, \cdots 6\})$ is the generators of $GA(2)$ and the basis of matrix $ga(2)$. For matrix $GA(2)$, the generators are

$$G_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad G_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad G_3 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$
$$G_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad G_5 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad G_6 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \tag{3}$$

For Lie group matrix, Riemann distance is defined by matrix logarithmic operation:

$$d'(X, Y) = \left\| \log(YX^{-1}) \right\| \tag{4}$$

where X and Y are the elements of Lie group matrix. Given N symmetric positive definite matrices $\{X_i\}_{i=1}^{N}$,the intrinsic mean is defined as

$$\mu^* = \exp(\frac{1}{N} \sum_{i=1}^{N} \log(X_i)) \tag{5}$$

For more knowledge of the Lie group, refer to the reference [40], [41].

B. DESIGN THE GEOMETRIC TRANSFORMATION MODEL

In the proposed method, affine transformation is applied to represent the process of object geometrical deformation during tracking. And the geometrical changes between two adjacent frames can be viewed as the movement of corresponding points of affine matrices on Riemann manifold, because affine transformation matrix is a positive definite symmetric manifold, which is a Lie group and no longer obeys Euclidean space. The basic idea for establishing the model of the target deformation is to find the transformation relationship between two adjacent points on the manifold. In this algorithm, the tangent vector of the point on the manifold is used to describe this kind of relationship. The objective deformation model is built in Riemannian manifold and tangent space, respectively:

$$S_t = S_{t-1} \exp(v_{t-1}) \tag{6}$$
$$v_t = a(v_{t-1} - v_{t-2}) + \mu_{t-1} \tag{7}$$

where the vector $S_t = [x_1, x_2, x_3, x_4, x_5, x_6]^{\mathrm{T}}$ is the affine transformation parameter of the target geometric deformation. v_t denotes as the velocity vector from point S_{t-1} to point S_t on the tangent space, and it describes the movement of the target, which is the tangent vector from point S_t on manifold. Suppose v_t is obeying the Gauss distribution, $\mu_{1:t}$ is Gauss white noise. and a is autoregressive coefficient.

The algorithm makes full use of the Lie group structure of affine transformation parameters space, with the status spaces being described on the manifold. The geometrical transformation is achieved on low dimensional manifold. Thus, it can decrease the dimension of tracking coefficient, and improve the tracking performance.

III. DISCRIMINATIVE CORRELATION FILTER

A correlation filter is typically trained by searching for the maximum value of correlation response maps to predict the object translation. In the proposed method, a standard discriminative correlation filters is used to develop the effect of convolutional features for object tracking. The feature maps of the final convolutional layer are applied as the input of the discriminative classifier. Compared with other costly methods used for CNNs training, the discriminative classifier filter is trained by computing a linear least-square and using Fast Fourier Transform (FFT), which is more efficient. Finally, the target appearance feature is determined by using the discriminative classifier.

Let x_k represents the input sample at frame k, where $k = 1, 2, \ldots, t$. t denotes the current frame number. And y_k which is set to Gaussian function label represents the desired correlation output at frame k.The aim is to gain a minimum loss by learning a correlation filter w, and the formula is

$$w^* = \arg\min_{w} \sum_{k=1}^{t} \|w_t \cdot x_k - y_k\|^2 + \lambda \|w_t\|^2 \tag{8}$$

where λ is a regularization parameter ($\lambda >= 0$), x_k^i denotes the ith feature channel of x_k.and the inner

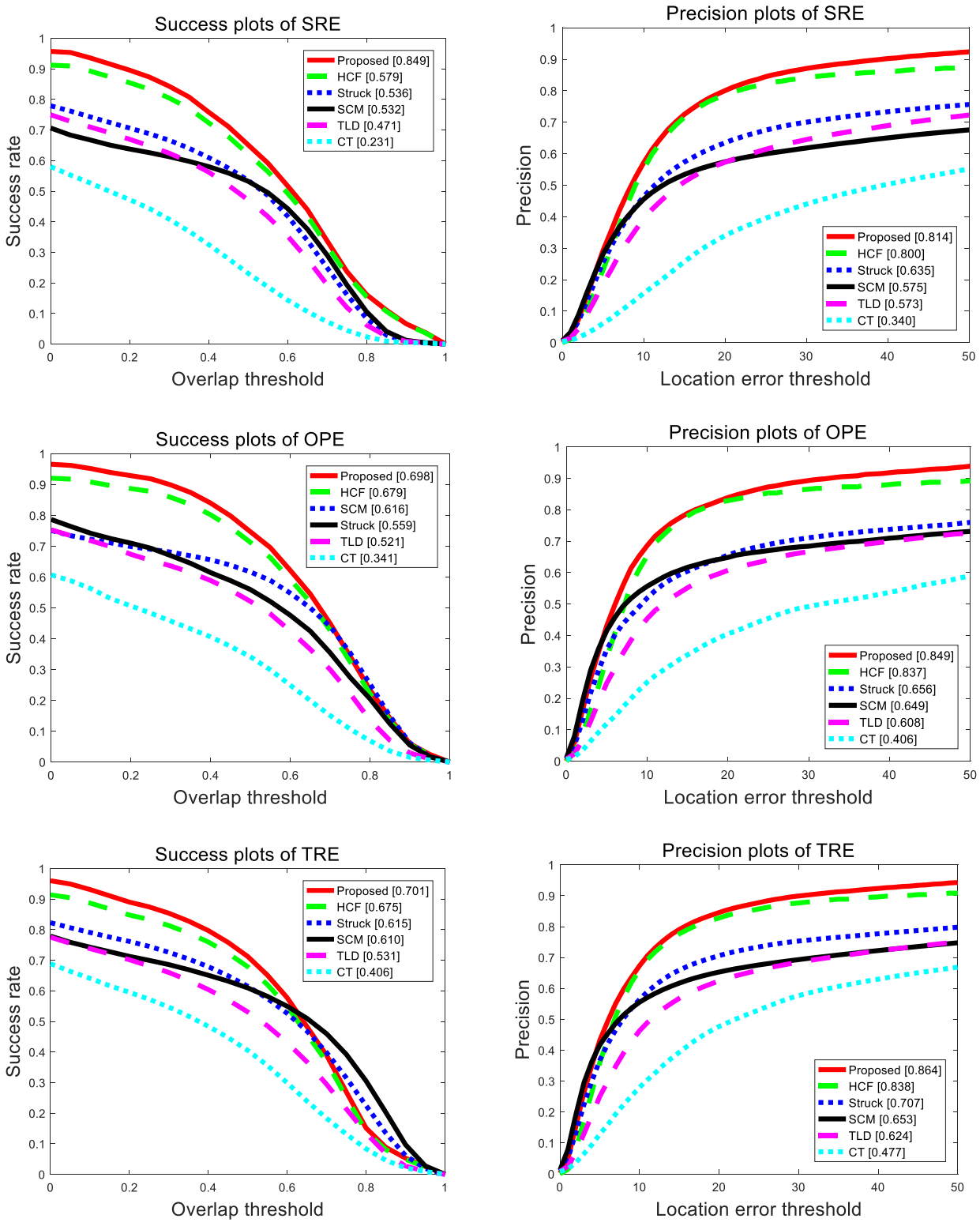
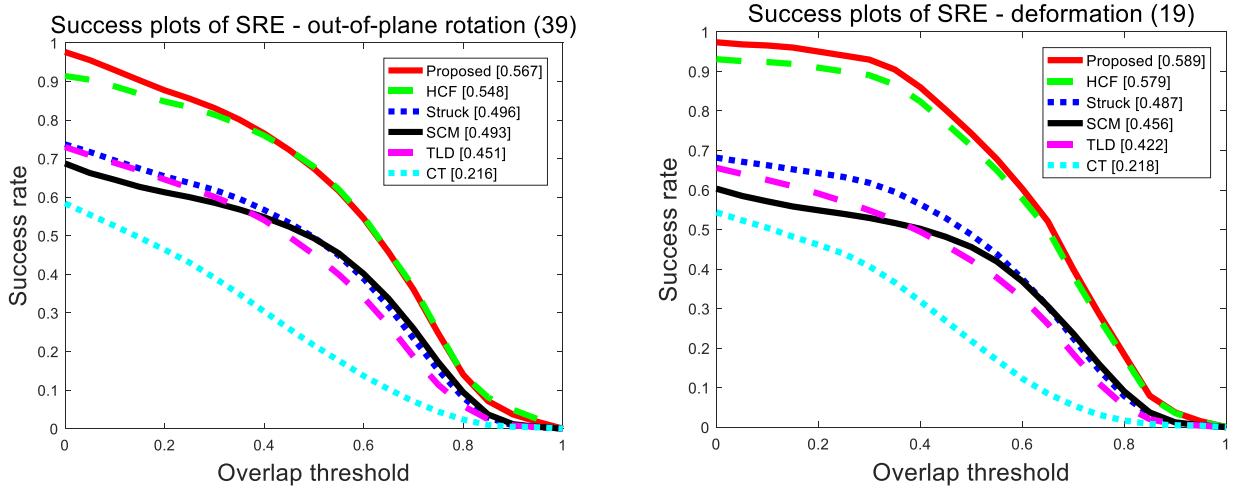
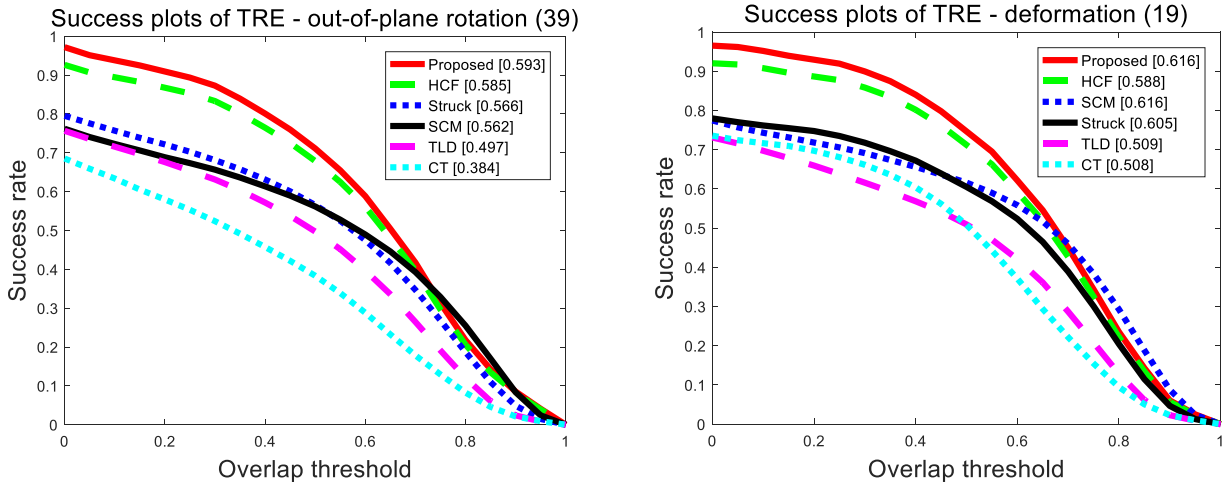


FIGURE 2. Bounding box overlap success rate plots and the center location error precision plots under SRE, OPE and TRE over benchmark sequences. The overlap success contains AUC score for each tracker, and the distance precision contains threshold scores at 20 pixels.

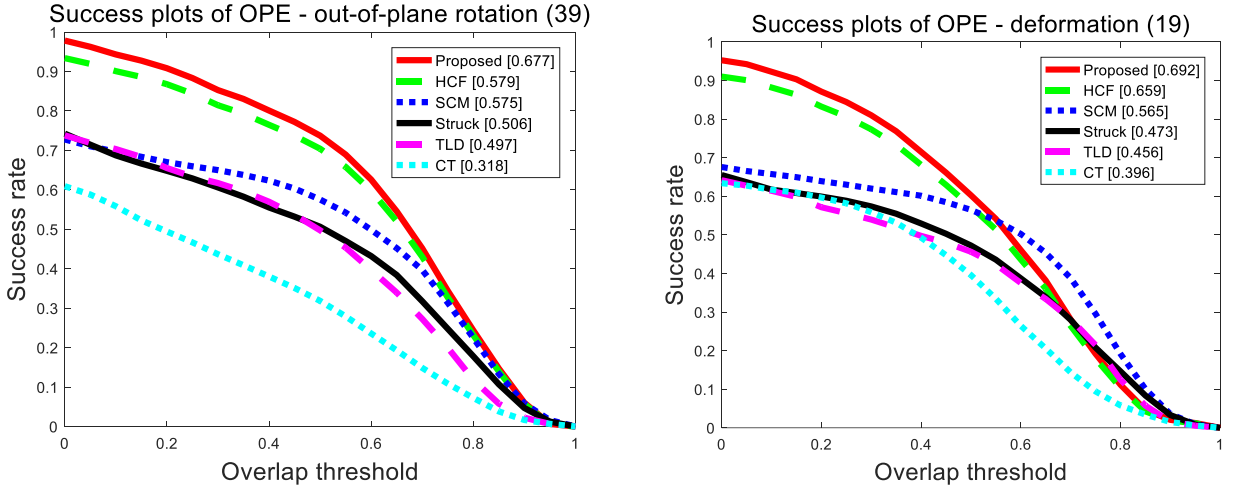
product is induced by a linear kernel in the Hilbert space. And the label y_k is soft, so hard-threshold samples is not required. So the minimization problem in (8) is converted to training the vector correlation filters, and can be solved in each feature channel using fast Fourier transformation (FFT).



(a) **Bounding box overlap success rate plots under SRE over sequences with out-of-plane rotation, deformation respectively.**



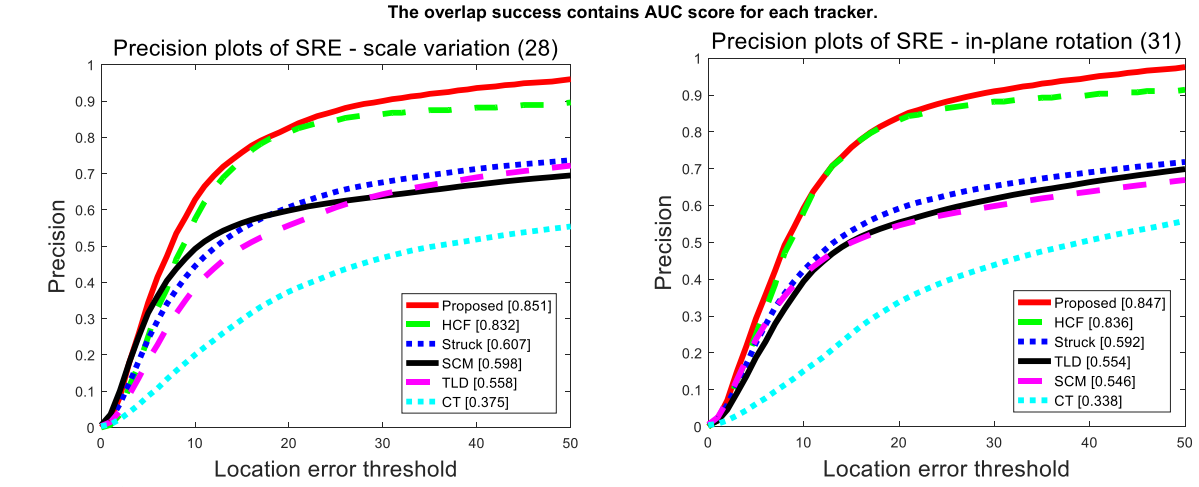
(b) **Bounding box overlap success rate plots under TRE over sequences with out-of-plane rotation, deformation respectively.**



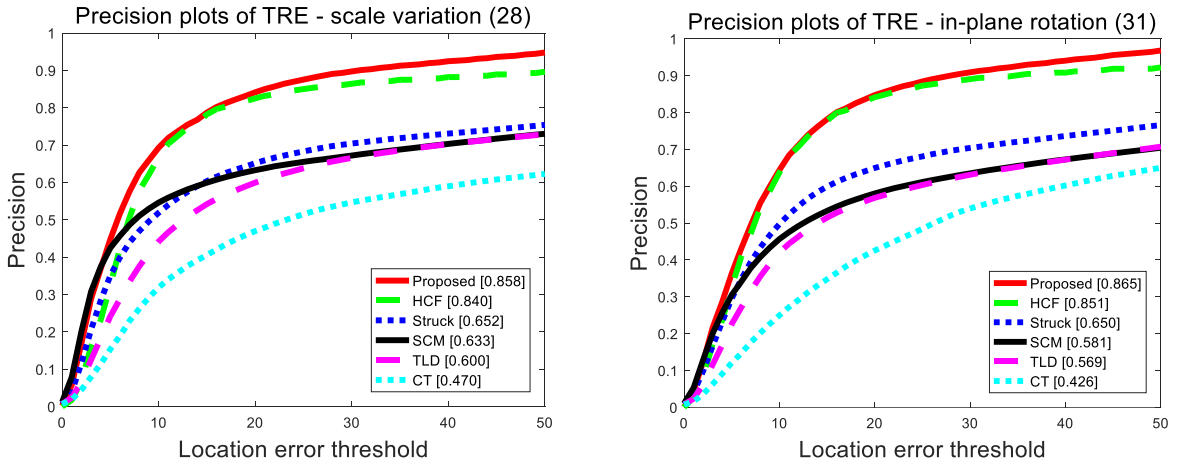
(c) **Bounding box overlap success rate plots under OPE over sequences with out-of-plane rotation, deformation respectively.**

FIGURE 3. **Bounding box overlap success rate plots under SRE, TRE and OPE over sequences with out-of-plane rotation, deformation respectively, The overlap success contains AUC score for each tracker.**

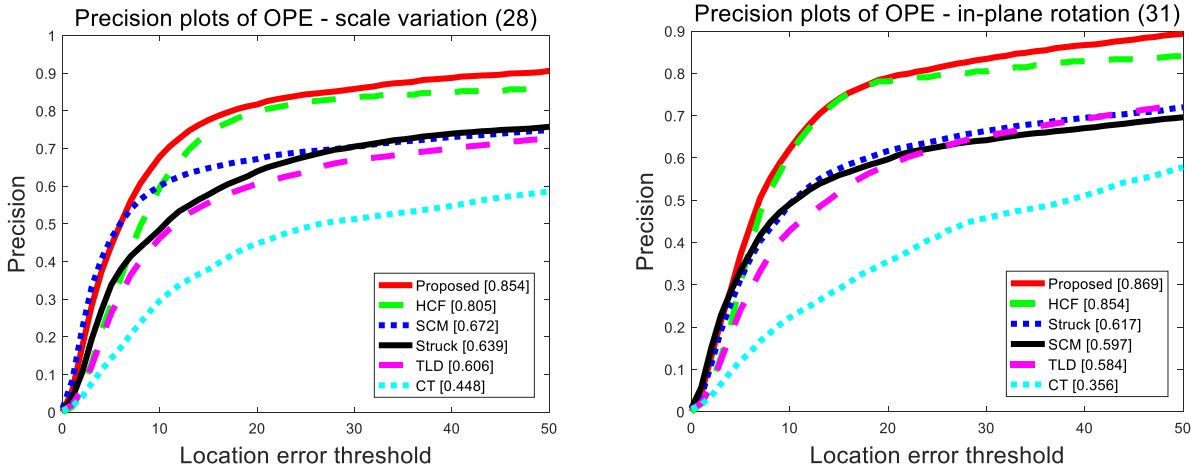
We update the optimal filter on l-th layer by minimizing the output error over all tracked results using the method proposed in [42]. However, a D $\times$ D linear system of equations must be solved for each location at frame t, which is computationally expensive as the channel number is usually large with the CNN features. To obtain a robust approximation,



(a) The center location error precision plots under SRE over sequences with scale variation and in-plane rotation respectively.



(b) The center location error precision plots under TRE over sequences with scale variation and in-plane rotation respectively.



(c) The center location error precision plots under OPE over sequences with scale variation and in-plane rotation respectively.

FIGURE 4. The center location error precision plots under SRE, TRE and OPE over sequences with scale variation and in-plane rotation respectively, and the distance precision contains threshold scores at 20 pixels.

we update the numerator M_t^i and denominator N_t of the FFT filter w^i at frame t as follows,

$$M_t^i = (1-\delta)M_{t-1}^i + \delta\, \bar{Y}_t \cdot X_t^i \tag{9a}$$

$$N_t = (1-\delta)N_{t-1} + \delta \sum_{i=1}^{d} \bar{X}_{t^i} \cdot X_t^i \tag{9b}$$

where the capital letter denotes the two-dimensional Fourier transformation from the corresponding letter, the operator $\cdot$ is element-wise multiplication, and δ is the learning rate.

The learned filter can be designed as follows,

$$W_t^i = \frac{M_t^i}{N_t} \tag{9c}$$

FIGURE 5. Qualitative evaluation for the proposed algorithm and the other four algorithms (Struck, CT, TLD, SCM).

$$r_t = \left\{ \sum_{i=1}^{d} \bar{W}_{t-1}^{i} \cdot Z_t^d \right\} \tag{10}$$

The filter is designed as formula (10), and the correlation scores are computed in the Fourier domain. To determine the location and the transformation for the target of frame t, firstly for a series of input appearance features, we compute the correlation scores for each of them. Then, the tracked target is determined by the appearance features that have the maximum value of all the correlation scores. Secondly, the target location and transformation of frame t are obtained by the affine manifold S_t, which is corresponding to the appearance features with maximum correlation score.

IV. TRACKING WITH CNN

Our method is designed on the observation that the last layer of CNNs encodes the semantic abstraction of objects and their outputs are robust to appearance variations. We apply affine transformation to predict possible locations of a target. For each possible location, the corresponding appearance feature is abstracted and input into the correlative filters to compute the confidence scores. The affine transformation sample corresponding to the maximum correlation score is the output of tracking result. The detail is shown in algorithm 1.

V. DETAILS AND EXPERIMENTAL EVALUATION

In section IV, we have designed the detailed steps of the proposed method shown in Algorithm 1. In this section, firstly, the implementation details of the proposed method are introduced. Then, experiments have been conducted for comparing the efficiency of the proposed method with the state-of-the-art methods on the benchmark datasets.

The implementation details are as follows. The proposed tracker is implemented in MATLAB using toolbox MatConvNet on the computer with Intel I7.40 GHz CPU and 64GB memory and use a standard discriminative correlation filer as framework to investigate the impact of convolutional features. and VGG-NET-16 [43] is applied to be trained on ImageNet [44] layers. Each layer of the correlation filters is designed with the same training parameters. The regularization parameter λ of equation (8) equals 10^{-4}. And the Gaussian function labels are generated by the kernel width 0.1. We also set the learning rate in equations (9a) and (9b) to $\delta = 0.01$. Moreover, the parameter setting of each

TABLE 1. Comparisons with the other four trackers on benchmark sequences. The proposed method performs better in distance precision (DP) rate at the threshold of 20 pixels, overlap success (OS) rate at an overlap threshold of 0.5 and center location error (CLE). And the tracking speed.

	The proposed	HCF	Struck	SCM	TLD	CT
DP rate (%)	85.7	83.7	63.5	57.5	57.3	34.0
OS rate (%)	68.8	65.5	51.6	51.1	49.5	27.6
CLE rate (%)	20.7	22.8	47.1	61.4	60.0	79.8
Speed (FPS)	10.1	10.4	10.0	0.37	21.7	38.8

Algorithm 1 The Proposed Algorithm

Input: the affine transformation of the first frame S_0, and the video sequence to be tracked.

Output:the tracked affine transformation S_t, and the learned correlation filters W_t^i.

Step1: Compute the candidate affine transformation sample $\{S_t^i, i = 1, 2, \cdots, M\}$ by using equations (6) and (7), where M is the sample number and t is the current frame number.

Step 2: For each affine transformation sample S_t^i, calculate the appearance feature $\{x_t^i, i = 1, 2, \cdots, M\}$ for the corresponding sample patch.

Step 3: Compute confidence scores $r_t^i = \left\{ \sum_{j=1}^{d} \bar{W}_{t-1}^i \cdot Z_t^d \right\}$ using equations (9a) through (10) for each $\{x_t^i, i = 1, 2, \cdots, M\}$.

Step4: Gain the maximum correlation score by using $r_t = \max_{i=1}^{m}(r_t^i)$.

Step5: Suppose the maximum correlation score corresponding to the affine transformation sample p, then $S_t = S_t^p$ is the output learned affine transformation parameter at frame t.

Step6: update the correlation filter with W_t^p.

Step7: $t = t + 1$, go to step 2.

compared-to methods is given in accordance with the original of the respective method.

Our tracker is evaluated with state-of-the-art trackers on the benchmark datasets OTB-2013 [45], OTB2015 [46] and OTB2016 [47]. The results are as follows.

A. QUANTITATIVE EVALUATION

We compare the proposed tracker with 5 representative trackers using online classifiers: HCF [38], Struck [21], CT [48], TLD [19], SCM [20]. The comparison is processed on the 100 sequences (Benchmark 2). We evaluate the efficiency in the quantitative and qualitative aspects via the following metrics: success of spatial robustness evaluation (SRE), success of temporal robustness evaluation (TRE) and success of one-pass evaluation (ORE). First, experiments were carried out on all the testing sequences. Then, our method was executed using the types of sequences classified according to the Benchmark database, separately. The investigated geometric transformations include out-of-plane rotation, scale variation, deformation, in-plane rotation.

1) OVERALL EVALUATION

We evaluate the overall performance of the proposed tracker and the compared trackers by success plots and precision plots, respectively.

The area under curve (AUC) scores is used to rank the trackers for success plots, while for precision plots, we use the tracking results on different sequences at the error threshold equaling 20 for trackers ranking. Figure 1 shows the results under evaluation metrics: SRE, OPE and TRE by using the above two different evaluation methods.

On the comparison in Figure 2, we can conclude that the same tracker gains different scores under the two metrics, because the two metrics measure different features of trackers. The results also tell us that the proposed method outperforms other four state-of-the-art methods in all three evaluation metrics: SRE, OPE and TRE. Since the precision plots only measure the efficiency of one tracker at one error threshold while the success plots measure the overall efficiency of the tracker, we mainly analyze the results of success plots while the results of precision plots are used as supplementary.

Furthermore, Table 1 shows quantitative results in distance precision (DP) rate at the threshold of 20 pixels, overlap success (OS) rate at an overlap threshold of 0.5 and center location error (CLE). From the table, our method performs better than the other four methods in distance precision rate, overlap success rate and center location error.

2) GEOMETRIC DEFORMATION EVALUATION

In the process of tracking, the geometric deformation of the target often occurs, and therefore it is of great significance for a tracking method to tackle the problem of tracking target with geometric deformation. In this section, we analyze the performance of trackers on different videos with out-of-plane rotation, scale variation, deformation, in-plane rotation. We report the tracking results of our method and the other four methods on benchmark sequence at the metrics of success plots and precision plots in Figures 3 and 4, respectively. Struck only predicts the location of the target and doesn't tackle scale variation or geometric deformation, and SCM uses sparse representation for appearance change, TLD applies dense sampling based trackers, and CT designs trackers which extract features from the multi-scale image feature space. While our method applies the affine group to describe the target transformation, and uses CNN network to gain a maximum confidence scores to locate the target. And

HCF can not describe the geometric deformation accurately for it only uses rectangular frames to determine tracking targets area. The figures show that our method is more efficient in all of the different target deformation.

B. QUALITATIVE EVALUATION

We test all the sequences provided by the benchmark I [38]. Because of page limit, just four sequence results are presented. And Figure 5 shows some tracking results of the tracking methods: Struck, CT, TLD, SCM, HCF and the proposed method on some challenging sequences.

In the first video sequence, the geometrical transformation occurs during the tracking. From the tracking results we can see that when the tracked object experiences obviously geometrical deformation, the proposed method can recognize the change, and the tracking boxes are not a rectangle, but a parallelogram, which can locate the object more accurately, while the tracking boxes drawn by HCF are still rectangles.

In the second video sequence, both geometrical deformation and occlusion occur when tracking. Both CT algorithm and TLD algorithm fail when the object is temporarily occluded as they don't have an effective strategy to process occlusion. While the proposed method, Struck method and SCM method can track the object. The proposed method can track the geometric deformation, while the Struck method can not. Because the Struck method only predict the position of the object, it doesn't deal with the scale varaition.

The third video is a long sequences with object apperance and illumination change during the tracking. SCM method uses sparse representation which is much effective for appearance change including occlusion. And TLD method fails to track in some sequences, but performs also well in long sequences with a re-detection module. Struck method is less as effective as others because it is more sensitive to illumination change. And the proposed method outperforms others on processing theappearance and illumination variations.

In the fourth video sequence, the tracked object undergoes dramaticly illumination change. All trackers except CT can track the object during the whole testing sequence. But Struck can not capture the scale changes of the target. And SCM and TLD methods can not capture the geometric deformation either. And CT method is less effective than others.

VI. CONCLUDING REMARKS

Because geometric deformation of targets occurs frequently during the tracking process, it is important for a robust tracker to capture the geometric deformation. But the state of art trackers using deep learning technology have no special strategy to handle this problem. In this paper, based on the observation that the higher level of CNN Network can better describe semantic information of objects and the outputs are robust to appearance variations, and the affine manifold can better locate the target, we propose a new tracking algorithm by using affine transformation and convolutional features. Furthermore, a standard discriminative correlation filter is used to develop the effect of convolutional features, and is more efficient than other methods used for CNN Networks. We conducted experiments for evaluating the proposed trackers performance on different videos with out-of-plane rotation, scale variation, deformation, in-plane rotation. All the analysis results show an outstanding performance of the proposed trackers.

REFERENCES

[1] Y. Wu, J. Cheng, J. Wang, H. Lu, J. Wang, H. Ling, E. Blasch, and L. Bai, "Real-time probabilistic covariance tracking with efficient model update," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2824–2837, May 2012.

[2] Z. H. Khan and I. Y.-H. Gu, "Tracking visual and infrared objects using joint Riemannian manifold appearance and affine shape modeling," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Barcelona, Spain, Nov. 2011, pp. 1847–1854.

[3] Z. H. Khan and I. Y.-H. Gu, "Bayesian online learning on Riemannian manifolds using a dual model with applications to video object tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Barcelona, Spain, Nov. 2011, pp. 1042–1409.

[4] L. Liu, D. Jing, and J. Ding, "Adaptive extraction of fused feature for panoramic visual tracking," in *Proc. IEEE 3rd Int. Conf. Image Vis. Comput. (ICIVC)*, Jun. 2018, pp. 21–25.

[5] T. Wakahara, Y. Kimura, and A. Tomono, "Affine-invariant recognition of gray-scale characters using global affine transformation correlation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 384–395, Apr. 2001.

[6] Y. Yamashita and T. Wakahara, "Affine-transformation and 2D-projection invariant k-NN classification of handwritten characters via a new matching measure," *Pattern Recognit.*, vol. 52, no. 4, pp. 459–470, Apr. 2016.

[7] J. Kwon and F. C. Park, "Visual tracking via particle filtering on the affine group," in *Proc. Int. Conf. Inf. Automat.*, Jun. 2008, pp. 198–216.

[8] J. Kwon, K. M. Lee, and F. C. Park, "Visual tracking via geometric particle filtering on the affine group with optimal importance functions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 991–998.

[9] Y. Rathi, N. Vaswani, A. Tannenbaum, and A. Yezzi, "Tracking deforming objects using particle filtering for geometric active contours," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1470–1475, Aug. 2007.

[10] K. Zhang, Q. Liu, J. Yang, and M.-H. Yang, "Visual tracking via Boolean map representations," *Pattern Recognit.*, vol. 81, pp. 147–160, Sep. 2018.

[11] K. Zhang, X. Li, H. Song, Q. Liu, and W. Lian, "Visual tracking using spatio-temporally nonlocally regularized correlation filter," *Pattern Recognit.*, vol. 83, pp. 185–195, Nov. 2018.

[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] J. Yang, K. Zhang, and Q. Liu, "Robust object tracking by online Fisher discrimination boosting feature selection," *Comput. Vis. Image Understand.*, vol. 153, pp. 100–108, Dec. 2016.

[14] W. Chen, K. Zhang, and Q. Liu, "Robust visual tracking via patch based kernel correlation filters with adaptive multiple feature ensemble," *Neurocomputing*, vol. 214, pp. 607–617, Nov. 2016.

[15] H. Song, Y. Zheng, and K. Zhang, "Robust visual tracking via self-similarity learning," *Electron. Lett.*, vol. 53, no. 1, pp. 20–22, 2017.

[16] K. Zhang, L. Zhang, and M. Yang, "Fast compressive tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2002–2015, Oct. 2014.

[17] D. A. Ross, J. Lim, R.-S. Lin, M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 125–141, 2008.

[18] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[19] 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.

[20] 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.

[21] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," in *Proc. ICCV*, Nov. 2011, pp. 263–270.

[22] X. Jia, H. Lu, M.-H. Yang, "Visual tracking via coarse and fine structural local sparse appearance models," *IEEE Trans. Image Process*, vol. 25, no. 10, pp. 4555–4564, Oct. 2016.

[23] C. Bao, Y. Wu, H. Ling, H. Ji, "Real time robust L1 tracker using accelerated proximal gradient approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1830–1837.
[24] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, p. 13, 2006.
[25] 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.
[26] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognit.*, vol. 76, pp. 323–338, Apr. 2018.
[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, Jun. 2014, pp. 580–587.
[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[29] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. CVPR*, 2014, pp. 1717–1724.
[30] J. Gao, H. Ling, W. Hu, and J. Xing, "Transfer learning based visual tracking with Gaussian processes regression," in *Proc. ECCV*, 2014, pp. 188–203.
[31] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. CVPR*, Jun. 2010, pp. 2544–2550.
[32] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. ECCV*, 2012, pp. 702–715.
[33] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Proc. ECCV*, 2014, pp. 127–141.
[34] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. BMVC*, 2014.
[35] M. Danelljan, F. S. Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. CVPR*, Jun. 2014, pp. 1090–1097.
[36] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proc. CVPR*, Jun. 2015, pp. 5388–5396.
[37] 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. CVPR*, Jun. 2015, pp. 749–758.
[38] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. ICCV*, Dec. 2015, pp. 4310–4318.
[39] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in*Proc. ICCV*, Dec. 2015, pp. 3074–3082.
[40] 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. ICCV*, 2017.
[41] B. C. Hall *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. Springer, 2003.
[42] V. Svensson, "Curvatures of lie groups," M.S. Thesis, Lund Univ., Stockholm, Sweden, 2009.
[43] V. N. Boddeti, T. Kanade, and B. V.K. V. Kumar, "Correlation filters for object alignment," in *Proc. CVPR*, Jun. 2013, pp. 2291–2298.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.
[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, Jun. 2009, pp. 248–255.
[46] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. CVPR*, Jun. 2013, pp. 2411–2418.
[47] 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.
[48] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. C. Zajc, T. Vojir, G. Hager, A. Lukezic, A. Eldesokey, and G. Fernandez, "The visual object tracking VOT2017 challenge results," in *Proc. ECCV Workshops*, Oct. 2017, pp. 1949–1972.
[49] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking," in *Proc. ECCV*, 2012, pp. 864–877.

YINGHONG XIE received the M.S. degree in software and its theory from Northeastern University, China, in 2005, and the Ph.D. degree in pattern recognition and artificial intelligence from Northeastern University, China, in 2014.

Since 2005, she has been working in Shenyang University, where she is currently an Associate Professor with the Information and Engineering Institute. From 2014 to 2016, she held a postdoctoral position with Tianjin University. She was a Visiting Scholar with the University of Michigan–Dearborn from 2017 to 2018. She is the first author of more than 20 articles, and hosts natural science foundation of China, in 2015. Her main research interests include video image processing and pattern recognition.

Dr. Xie is a member of the provincial intelligent building committee. She received the Second Prize of the Provincial Natural Science Academic Achievement in 2014.

JIE SHEN served as an editorial board member for two international journals; an organizer for eight international conferences; an associate editor of two international conference proceedings; a program committee member for 20 international conferences; a session chair for 13 international or national conferences; a board member for three international- or national-level technical committees; and a member for various committees at department and campus levels within the University of Michigan–Dearborn. He is the Editor-in-Chief of *International Journal of Modelling and Simulation*, which is an EI-indexed, peer-reviewed research journal in the field of modeling and simulation.

Dr. Shen author's awards and honors include the Frew Fellowship (Australian Academy of Science), the I. I. Rabi Prize (APS), the European Frequency and Time Forum Award, the Carl Zeiss Research Award, the William F. Meggers Award, and the Adolph Lomb Medal (OSA).

CHENGDONG WU is currently the Vice President of the Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China, where he is also the Director of the Institute Artificial Intelligence, a Professor, and the Doctoral Tutor. He has long been involved in automation engineering, artificial intelligence, and teaching and researching in robot navigation. He is an Expert of Chinese modern artificial intelligence and robot navigation. He is also a Special Allowance of the State Council.