Learning correlation filter with fused feature and reliable response for real-time tracking

Lin, Bin; Xue, Xizhe; Li, Ying; Shen, Qiang

Published in:
Journal of Real-Time Image Processing

DOI:
10.1007/s11554-022-01195-2

Publication date:
2022

Citation for published version (APA):
Lin, B., Xue, X., Li, Y., & Shen, Q. (2022). Learning correlation filter with fused feature and reliable response for real-time tracking. *Journal of Real-Time Image Processing*, *19*(2), 417-427. https://doi.org/10.1007/s11554-022-01195-2

Journal of Real-Time Image Processing manuscript No.
(will be inserted by the editor)

Learning correlation filter with fused feature and reliable response for real-time tracking

Bin Lin[1,*] · Xizhe Xue[1,*] · Ying Li[1] · Qiang Shen[2]

Received: date / Accepted: date

Abstract Object tracking is a key component of machine vision system and getting much attention in different walk of life. Recently, correlation filters have been successfully applied to visual tracking. However, how to design effective features and deal with model drifts remain open issues for online tracking. This paper tackles these challenges by proposing a real-time correlation tracking algorithm (RCT) based on two ideas. First, we propose a method to fuse features in order to more naturally describe the gradient and color information of the tracked object, and introduce the fused feature into a background-aware correlation filter to obtain the response map. Second, we present a novel strategy to significantly reduce noise in the response map and therefore ease the problem of model drift. Systematic comparative evaluations performed over multiple tracking benchmarks demonstrate the efficacy of the proposed approach. The results show that the proposed RCT significantly improves the performance compared to the baseline tracker while still maintaining a real-time tracking speed of 26 fps in MATLAB implementation.

Keywords Visual tracking · Real-time tracking · Correlation filter · Fused feature · Model drift

✉ Ying Li
lybyp@nwpu.edu.cn

Bin Lin
binlin@mail.nwpu.edu.cn

Xizhe Xue
xuexizhe@mail.nwpu.edu.cn

Qiang Shen
qqs@aber.ac.uk

[1] School of Computer Science, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, Northwestern Polytechnical University, Xi'an 710072, PR China

[2] Department of Computer Science, Faculty of Business and Physical Sciences, Aberystwyth University, Aberystwyth SY23 3DB, U.K.

[*] Bin Lin and Xizhe Xue contributed equally to this work.

1 Introduction

Visual tracking plays an active role in a wide range of applications, including surveillance systems, driverless vehicles, robotics, human-computer interaction and so on. The task of object tracking involves estimating the states (positions and scales) of the target in subsequent frames, with initial state given in the first frame. Recent years have witnessed significant developments in visual tracking, where an enormous amount of research effort has gone into tasks such as short-term single-object tracking. However, many challenges remain, such as target deformation, rotation, scale variation, occlusion, and imbalanced training samples [3]. Furthermore, object tracking is usually only a single component in a complete machine vision system, thus the real-time capabilities of tracking algorithms are of paramount importance for the whole pipeline to work online. Nevertheless, the current top-ranking trackers are mostly based on deep learning technology and are neither memory-efficient nor real-time capable.

Correlation filters have recently been introduced for visual tracking and have been shown to achieve high speed as well as robust performance. Thanks to the learning of a correlation operator which is formulated as a ridge problem can be accelerated by fast Fourier transform (FFT) in the frequency domain, the correlation filter-based trackers (CFTs) can perform real-time

tracking. In general, extracting powerful features is extremely crucial for CFTs. Gradient [16], color [4,12], and deep features [24,32] which extracted from convolutional neural networks (CNNs) are widely used in CFTs. However, how to best utilize different features jointly for real-time tracking remains an open issue. There is another tough problem for most CFTs, that is the trackers can not maintain tracking robustness in the subsequent frames once the model drift occurs. Model drift means that the object appearance model gradually drifts away from the object due to the accumulated errors during online tracking. Existing works [17,18,20] have aimed to prevent model drift through modifying the training strategy rather than improving the underlying model-based predictions themselves.

In this paper, we propose a robust correlation tracking method (RCT) via the exploitation of feature fusion and reliable response. A fused feature herein describes the gradient and color information conjunctively in a more natural way as compared to existing approaches [21,22] which directly concatenates features together. The novel fused feature is then embedded into a correlation filter that is background-aware (in the sense that the filter is capable of learning from real, negative examples densely extracted from the background). For alleviateing the model drift issue, an adaptive optimization strategy is introduced to remove the untrusted part of the response map that is caused by deformation or other challenging factors, so as to improve the predictions by obtaining and manipulating a more reliable response map which leading to an enhanced tracking result. The flowchart of the proposed approach is shown in Fig. 1.

We evaluate the proposed tracker on the OTB [29, 30] and Temple-Color [23] datasets. The results demonstrate that our method obtains a very competitive accuracy level in comparison with the state-of-the-art trackers, but does so with a real-time tracking speed of 26 fps on a standard desk-top CPU.

The remainder of this paper is organized as follows. Sect. 2 introduces the related works, and Sect. 3 describes the details of the proposed approach. Experiments and analysis are conducted in Sect. 4, and Sect. 5 concludes this study.

2 Related works

We discuss correlation filter-based tracking methods closely related to this work in this section. For the other visual tracking approaches, readers are referred to comprehensive review [19,26,29].

2.1 Feature representation in correlation tracking

Bolme et al. [5] proposed one of the seminal correlation tracking methods based on the minimum output sum of squared errors (MOSSE), which can perform online tracking at an astonishing tracking speed of ∼700 fps. In MOSSE, raw pixels are directly used for tracking. Unfortunately, noises brought by raw images extremely limit its tracking performance. Over the years, gradient and color features have been successfully applied in CFTs. The kernelized correlation filter tracker (KCF) [16] employs the famous histogram of oriented gradient (HOG) [8] feature to improve the accuracy of the tracker. Also, color features like color names (CN) [12] and global color histogram [4] are investigated to reinfore color-video tracking for CFTs. Li et al. [22] proposed a scale adaptive with multiple features tracker (SAMF), which fuses HOG and CN within correlation tracking framework, to further boost the tracking performance. After recognizing the success of deep learning on a wide range of visual-recognition tasks, a number of tracking methods based on deep features and correlation filters have been developed [24,32]. For instance, Ma et al. [24] utilized hierarchical CNN features to exploit semantic information of the target object with a state-of-the-art performance. However, extracting CNN features from each frame, and training or updating correlation filters with high dimensions is computationally expensive. Therefore for correlation tracking, such an approach often leads to poor real-time performance.

2.2 Robustness to model drift

Model drifts lead to inaccurate model-based predictions. In addressing this problem, Kalal et al. [18] proposed an approach that decomposed the ultimate task of tracking into subtasks of tracking, learning and detection (TLD), where tracking and detection reinforce each other. Zhang et al. [31] introduced the fuzzy logic [2] to alleviate model drift by formulating tracking as a fuzzy classification problem. Inspired by the KCF and TLD trackers, Li et al. [20] proposed a scale adaptive kernelized correlation filter tracker, termed as SKCF, which estimates an accurate scale and models the distribution of correlation response with Gaussian constraint during the process of re-detection. However, the circulant shifted samples in such CFTs suffer from periodic repetitions on boundary positions, thereby leading to model drift and significantly degrading the tracking performance. Spatial regularization methods have since been suggested to alleviate the unwanted boundary effects. For example, using the alternating direction method of multipliers (ADMM) [6], Galoogahi et al. [14] resolved a

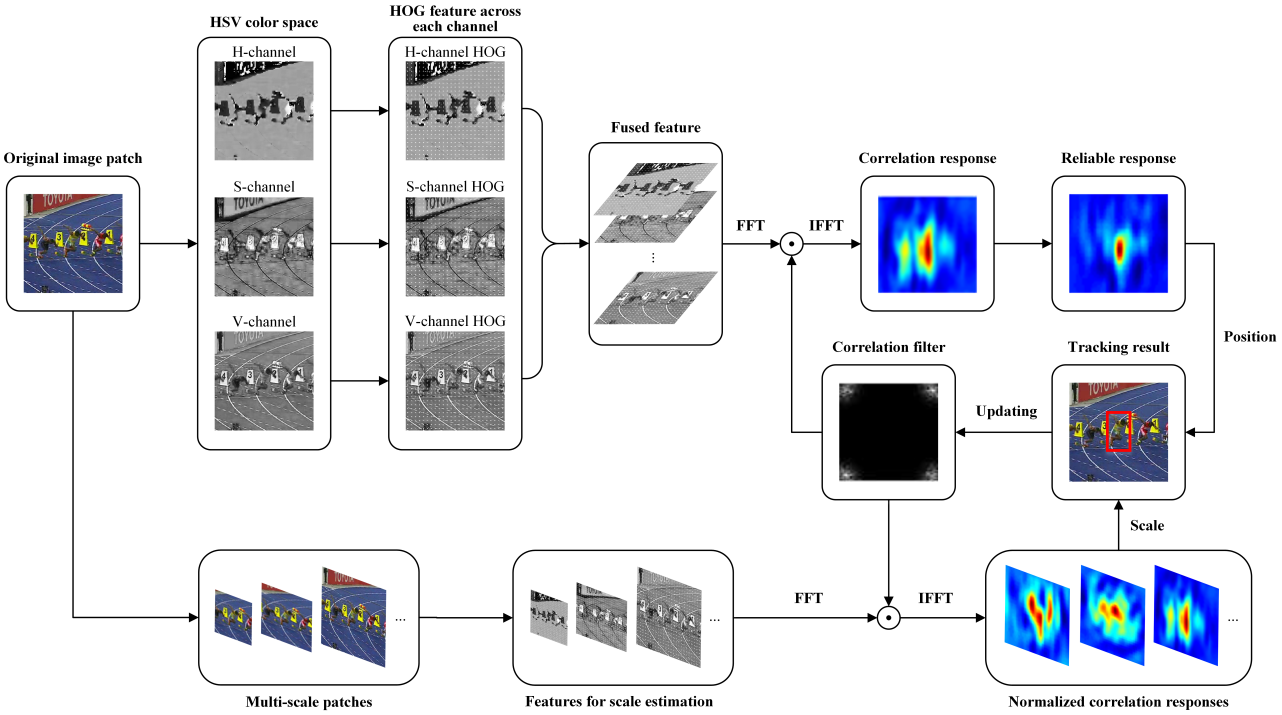


Fig. 1 Flowchart of the proposed approach. The operator $\odot$ is the Hadamard (element-wise) product

constrained optimization problem for single-channel filters. Somewhat differently, the SRDCF formulation [10] allows correlation filters to be trained on a significantly larger set of negative training samples, without corrupting the positive samples, where a spatial regularization component is introduced to the training process to penalize the correlation filter coefficients in relation to their spatial location. Recently, Varfolomieiev et al. [27] combines the channel-independent calculation with the spatial regularization to suppress the background filter component. Unlike previous CFTs, in which negative examples are restricted to circular shifted patches, BACF [13] utilizes a correlation filter whose spatial size is much smaller than that of the training samples; real negative training examples, densely extracted from the background are utilized. To avoid drifting for real-time UAV tracking, Huang et al. [17] tried to repress aberrances during the training phase.

Compared with the existing methods, our proposed tracker has several merits. First, while RCT may be viewed as an (improved) approximation to the work of [13] on multiple training samples, the filter works more efficiently owing to the use of a more reliable response map. Second, with the introduction of fused features, the RCT tracker can learn more robust features than the previous work, thereby leading to superior tracking performance.

3 Proposed approach

We aim to develop a robust tracking algorithm that is adaptive to significant changes without being prone to drifting. We first propose a fused feature mechanism which describes the gradient and color information in an integrated way. Then, a background-aware correlation filter based on the exploitation of fused features is designed to obtain a response map. Furthermore, the mask obtained according to the value of the response map will be multiplied with a given original image to form a more reliable response map, which help alleviate possible model drifts.

3.1 Multi-channel fused features

Inspired by the duplicity theory of vision [15], we construct a more natural feature representation to fuse different types of features. In our setting, instead of concatenating the color and gradient features directly, we first transform the original image patch into HSV (Hue, Saturation, Value) color space, which is based more upon how colors are organized and conceptualized in human vision. In such a color space, brightness and colorfulness are absolute measures, which usually describe the spectral distribution of light entering the eye. Benefiting from this, our fused feature performs robust to the illumination variation. Secondly, HOG

gradient information is extracted from each channel of the HSV color space, separately. Finally, all the HOG features are concatenated to form our proposed fused feature, in the form of a 93-dimensional matrix. Hence, for terminology, we think of the output feature as the combination of fused-(input)-features or, as a (singular) fused feature. Without losing generality and for conciseness, we term the resultant feature descriptor a fused feature.

3.2 Correlation tracking through fused feature

In this subsection, we introduce our fused feature into background-aware correlation filter [13] to construct a better correlation tracking framework. We utilize a correlation filter with a spatial size which is smaller than the size of training examples to reduce the boundary effects. Denote x_k as the fused feature vector of a cardinality $x_k \in \mathbb{R}^T$, respectively. We consider $y \in \mathbb{R}^T$ as the desired correlation output corresponding to a given sample x_k. A correlation filter w with the dimensionality of D (where $T >> D$) is then learned by solving the following minimization problem as that:

$$E(w) = \sum_{j=1}^{T} ||y_j - \sum_{k=1}^{K} w_k^{\top} \mathrm{P} x_k[\Delta\tau_j]||^2 + \lambda \sum_{k=1}^{K} ||w_k||_2^2, \tag{1}$$

where λ is a regularization parameter, P is a binary matrix, and $\mathrm{P}x_k[\Delta\tau_j]$ generates all circular shifts of size D from the entire image patch over all $j = [0, ..., T-1]$ steps. The transpose operator $^{\top}$ on a complex vector or matrix gives the conjugate transpose.

Note that the Eq. (1) can be readily transformed into frequency domain in order to improve the computational efficiency. We introduce $\hat{g} = [\hat{g}_1^{\mathrm{T}}, \cdots, \hat{g}_K^{\mathrm{T}}]^{\mathrm{T}}$ as an auxiliary variable. The trained filter in the frequency domain will be written as:

$$\begin{aligned} &E(w, \hat{g}) = ||\hat{y} - \hat{X}\hat{g}||_2^2 + \lambda ||w||_2^2, \\ &\text{s.t.} \quad \hat{g} = \sqrt{T}(\mathrm{F}\mathrm{P}^{\top} \otimes \mathrm{I}_K) w \end{aligned} \tag{2}$$

where $\hat{X}$ is denoted by $\hat{X} = [diag(\hat{x}_1)^{\top}, ..., diag(\hat{x}_K)^{\top}]^{\top}$, $\mathrm{I_K}$ is the $K \times K$ identity matrix, and $\otimes$ denotes the Kronecker product. In particular, $\hat{A}$ represents the FFT of a signal A, where F is the orthonormal $T \times T$ matrix of complex basis vectors, mapping any T-dimensional vectorized signal to its Fourier domain.

By directly employing the augmented Lagrangian method (ALM) [13], we can solve Eq. (2) and obtain the required correlation filter $\hat{g}^{(f-1)}$, where f is the current frame number.

3.3 Object location by reliable response

Representing the response value of every pixel, the response map $r^{(f)}$ in frame f can be computed by applying the filter $\hat{g}^{(f-1)}$ that has been updated in the previous frame as:

$$r^{(f)} = \mathcal{F}^{-1}(\sum_{k=1}^{K} \hat{x}_k^{(f)} \odot \hat{g}_k^{(f-1)}), \tag{3}$$

where $\odot$ denotes the Hadamard product, and $\mathcal{F}^{-1}$ is the inverse FFT (IFFT) transform.

Due to the challenges typically faced in performing real-world tracking tasks, such as deformation and rotation, the similarity between the target and modeling template may be decreased, leading to great risk of model drift or locating mistakenly. Therefore, the response map r obtained by Eq. 3 can be regarded as an original (coarse) response. How to remove a lot of potentially misleading redundant information (responses to similar objects) contained in the original response map then? As shown in Fig. 2, when noise exists, the position with the maximum value in the response map does not necessarily correspond to the real target. In this case, simply taking the position with the highest response as the target position is rather unreliable. Through a large number of experiments, empirically we find that the response peak of the real target often changes gradually, while the response peak of the disturbed object is usually very steep and looks very abrupt. Accordingly, in order to exclude the anomalies, we first try to identify the target proposals which are associated with a relatively high value in the response map. In order to achieve this goal, we exploit a threshold α, which divides the response map $r^{(f)}$ into two parts. The pixels with a gray value greater than α belong to the target proposal set A and the remaining ones are deemed to attribute to the background part B. The number of pixels contained in the two parts is represented with $N_{A,\alpha}$ and $N_{B,\alpha}$ respectively. We vary α from 0 to 255, each time, $N_{A,\alpha}$ and $N_{B,\alpha}$ are counted to calculate the ratio of the target proposals in the patch, which is denoted as $Q_{\alpha}^{(f)}$, f is the frame index, such that:

$$Q_{\alpha}^{(f)} = \frac{N_{A,\alpha}^{(f)}}{N_{A,\alpha}^{(f)} + N_{B,\alpha}^{(f)}}. \tag{4}$$

Repeat Eq. (4) until the difference between the $Q_{\alpha}^{(f)}$ and $Q^{(1)}$ (initial ratio of the target area in the patch) is less than the error range threshold β as:

$$|Q_{\alpha}^{(f)} - Q^{(1)}| < \beta. \tag{5}$$

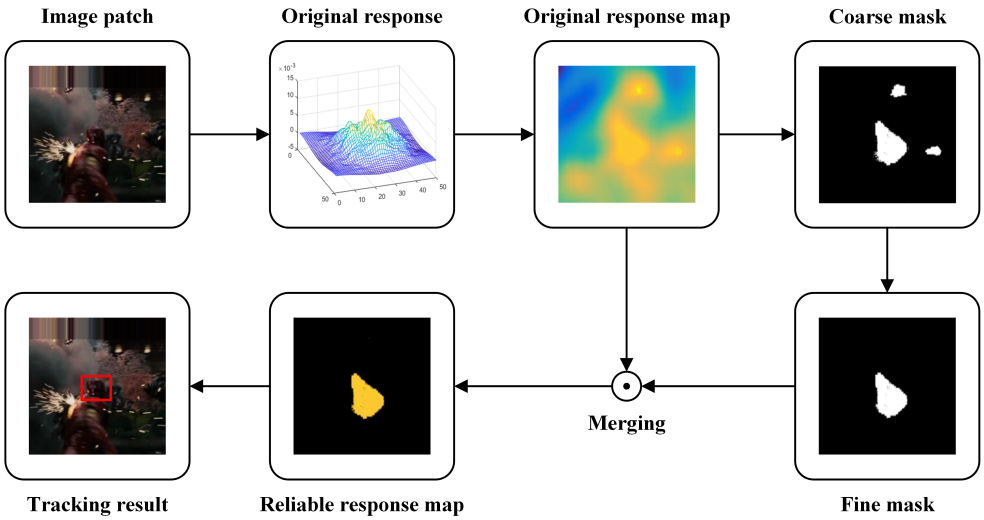


Fig. 2 Object location by our reliable response map. We improve the predictions by manipulating the reliable response map which is obtained by merging a coarse-to-fine mask, leading to an enhanced tracking result

When Eq. (5) is satisfied, the grey value of pixels in the set A is reset to 255, while each of the rest pixels is set to 0. From this, a number of connected domains are obtained. Then, any connected domain whose pixel area is less than a fixed threshold μ is deleted to form the fine mask matrix $\mathrm{M}^{(f)}$. By merging $\mathrm{M}^{(f)}$ with the original response map $r^{(f)}$, the reliable response map $\tilde{r}^{(f)}$ results. Finally the position with the maximum value in the reliable response map $\tilde{r}^{(f)}$ is treated and recognized as the target location.

3.4 Model updating and scale estimation

To obtain a robust approximation, at frame f, we use an online updating strategy which is formulated as:

$$\hat{x}^{(f)}_{model} = (1-\eta)\hat{x}^{(f-1)}_{model} + \eta\hat{x}^{(f)}, \tag{6}$$

where $\hat{x}^{(f)}_{model}$ and $\hat{x}^{(f-1)}_{model}$ represent the newly updated template model and old one respectively, η is the learning rate.

In order to be adaptable to any change of the scale of a target, the filter is applied on multiple resolutions of the searching area where tracking takes place [22]. This returns S correlation outputs with different scales, where S is the number of scales. The scale with the maximum correlation output is used to update the object location and the subsequent scale. To sum up, Algorithm 1 recapitulates the whole method.

Algorithm 1: Tracking algorithm

Data: Frames I_f, initial target location p_1 (f means the number of the current frame)
Result: Target location p_f

1 **repeat**
2 Crop an image region from I_f at the last location and extract its fused-feature vector x_f;
3 Compute the optimum correlation filter (via Eq (2)) and obtain the original response map;
4 Construct the mask to yield a reliable response map;
5 Detect the target location p_f via the reliable response map;
6 Estimate the scale of the target and update the tracking model (as summarized in Sect. 3.4);
7 **until** end of video sequence.

4 Experimental results

In order to present an objective evaluation regarding the performance of the proposed approach, we examine our RCT tracker on three standard datasets, including OTB50 [30], OTB100 [29], and Temple-Color128 (TC128) [23]. Both the general capability and the special scenarios-handling ability are tested. The experiments are performed in Matlab R2016b on an Intel i7 3.0GHz CPU with 16G RAM. In all the experiments carried out, we use the same parameter values for all image sequences. We employ HOG features with 4×4 cells to obtain the fused feature. The regularization factor is empirically set to 0.001 and the number of scales is set to 5 with a scale-step of 1.01. A 2D Gaussian function with bandwidth of $\sqrt{wh/16}$ is used to define the correlation output for an object of size $[h, w]$. The learning rate η of the correlation filter is 0.013. The pixel area threshold μ is set to 105, and the error range β is set to be within 0.07.

We compare our tracker with a range of excellent trackers, including: BACF [13], ECO_HC [9], ARCF [17], AutoTrack [21], KCC [28], Staple_CA [25], MCPF [32],

SRDCFdecon [11], BIT [7]. Different metrics may be used for evaluation depending on preferred perspectives, amongst which one-pass evaluation (OPE) is arguably the most commonly used evaluation method. OPE runs a tracker on each sequence once: it initializes a tracker using the ground truth object state in the first frame, and reports the average precision or success rate of all subsequent results. Having recognized this, OPE is also used herein to comparatively evaluate the present work. Center location error (CLE) is obtained through the Euclidean distance between the center of the groundtruth and estimated bounding box. Overlap precision (OP) is computed as the fraction of frames in the sequence where the intersection-over-union (IOU) overlap between the groundtruth box and the tracker prediction is higher than a threshold, and area under curve (AUC) score is the average of the success rates corresponding to the sampled OP thresholds. The trackers are ranked by the distance precision (DP) score with a CLE threshold of 20 pixels in the precision plots, and by the AUC score depicted by the success plots.

4.1 Evaluation on OTB datasets

We implement the one-pass experiment on the OTB benchmark datasets. Fig. 3 shows the precision and success plots on OTB50 and OTB100, respectively. The DP and AUC scores of all compared trackers are shown in the legend. Overall, the RCT tracker performs well on these two evaluation metrics. It ranks the first in the success plots and second in the precision plots among all competing algorithms. Our RCT approach employs the "background-aware" mechanism from the BACF tracker, but achieves a remarkable gain on the baseline. BIT is a tracker that extracts low-level biologically-inspired features while imitating an advanced learning mechanism to combine generative and discriminative models for target location. Our RCT improves on BIT by 15.72% in terms of AUC score on OTB50, and by 14.52% on OTB100. This testifies to the extraordinary performance of the fused features embedded in the correlation tracking framework. ECO_HC is a famous correlation filter-based tracking algorithm, our method also obtains superior results than ECO_HC in both of the DP and AUC scores. Note that the MCPF tracker outperforms RCT by 2.66% on OTB50, and 2.65% on OTB100 in terms of DP scores, yet its AUC scores are lower than our tracker with 1.99% and 1.62% respectively.

Table 1 AUC scores of the proposed trackers versus other state-of-art trackers. The first, second and third best methods are shown in color (The red ranks first, green means second and blue means third)

Trackers	AUC scores (%)			
	OTB50	OTB100	TC128	Average
BACF	57.39	62.03	49.56	56.33
ECO_HC	57.17	62.49	54.69	58.10
AutoTrack	52.83	58.69	52.16	54.56
ARCF	55.50	60.66	51.94	56.03
KCC	51.12	56.46	49.00	52.19
Staple_CA	54.21	59.83	50.55	54.86
MCPF	57.69	62.70	54.06	58.15
SRDCFdecon	55.23	61.97	48.21	55.14
BIT	43.96	49.80	44.93	46.23
RCT	59.68	64.32	53.63	59.21

4.2 Evaluation on TC128 dataset

TC128 is a comprehensive color-video tracking benchmark. The results of ten trackers on the 128 sequences are summarized in Fig. 4. As can be seen from these results, for both of the precision plots and success plots, our tracker obtains the third place and performs reliably. Compared to baseline BACF, RCT has a significant advantage of 4.48% in DP score and 4.07% in AUC score, which is in light of the robust fused features and reliable response. MCPF utilizes deep features which is extracted from the pre-trained convolutional neural network and obtains the first place in terms of DP score. Due to an adaptive decontamination of the training set and a conservative model update strategy, ECO_HC also performs better than RCT and ranks the first in terms of AUC score on this dataset.

For further overall comparison, in Table 1, we summarize the AUC scores of all compared trackers from the experimental results on the three datasets. It shows that our RCT tracker achieves the highest AUC score of 59.21% on average, outperforms all handcrafted features-based trackers and, including even MCPF, which utilizes deep features (and hence involves substantially more computation). Moreover, our approach just uses the simple BACF as baseline, it should be noticed that ECO_HC can further enhance its performance with our framework.

4.3 Attribute-based performance

As in the OTB datasets, all the image sequences are annotated with 11 attributes which cover various challenging factors in visual tracking, including scale variation (SV), occlusion (OCC), illumination variation (IV), motion blur (MB), deformation (DEF), fast motion (FM),

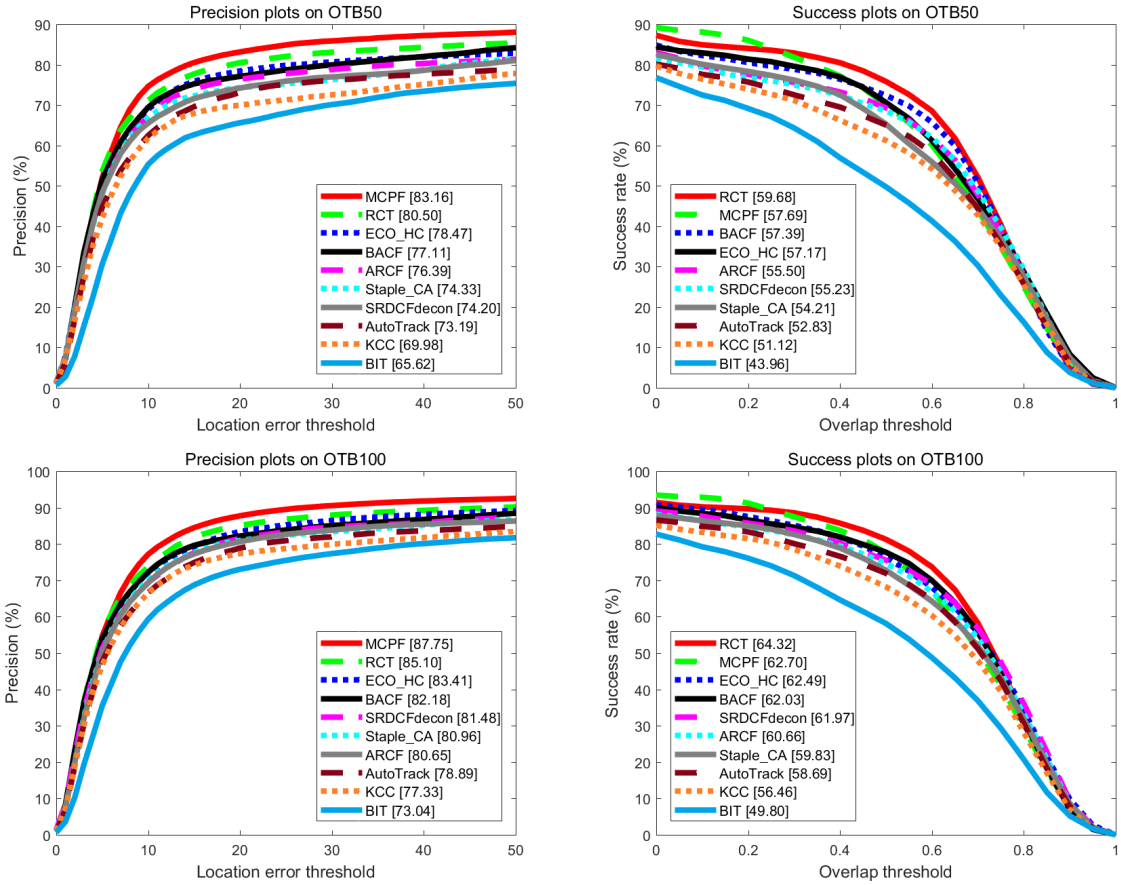


Fig. 3 Results of the proposed tracker and other compared trackers on OTB dataset

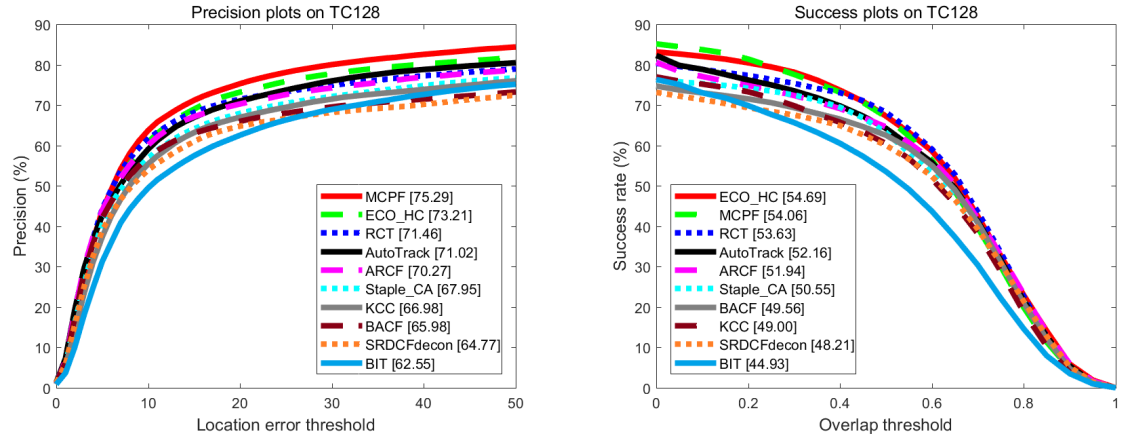


Fig. 4 Results of proposed tracker and other compared trackers on TC128 dataset

out-of-plane rotation (OPR), background clutters (BC), out-of-view (OV), in-plane rotation (IPR) and low resolution (LR). Fig. 5 shows the results of six representative attributes (FM, IV, MB, OCC, OPR and SV) over the OTB50 benchmark to testify the excellent attribute-performance of our RCT tracker in terms of AUC scores. It shows that the proposed method performs robustly against other state-of-the-art trackers in most challenging scenes.

Fig. 6 shows a qualitative comparison of our method with several state-of-the-art trackers including MCPF, ECO_HC, BACF and ARCF in challenging situations. The example frames are from the *DragonBaby*, *Soccer* and *Rubix* sequences, respectively. Obviously, our approach performs well as compared to the others. Se-

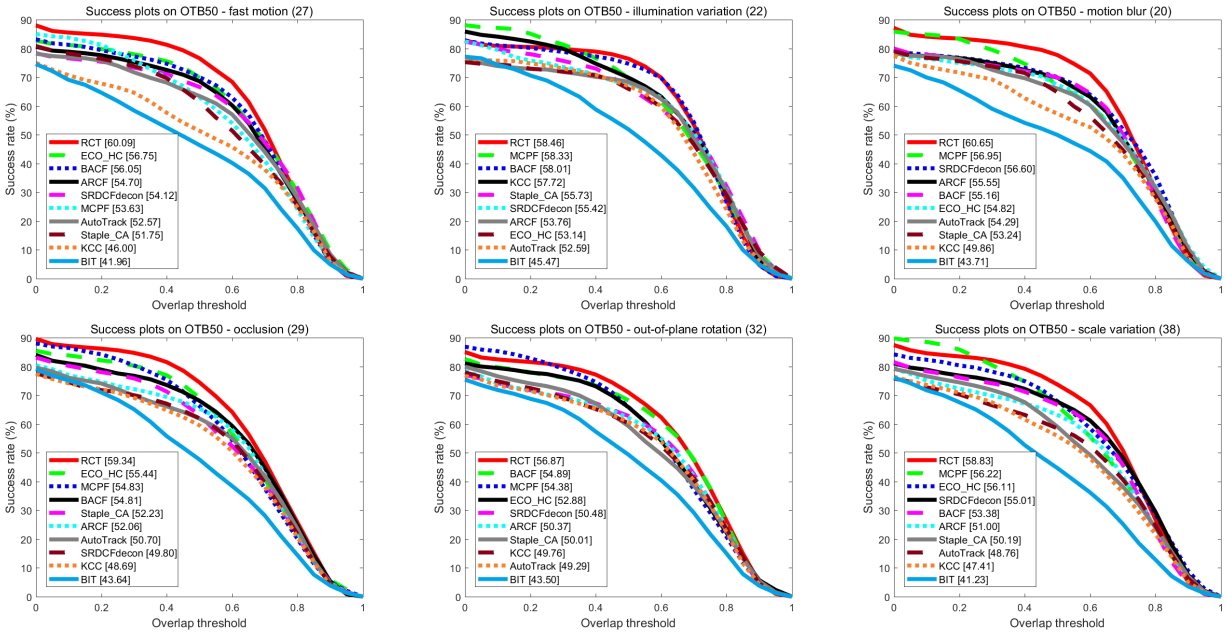


Fig. 5 Results of proposed tracker and other compared trackers on annotated challenging attributes

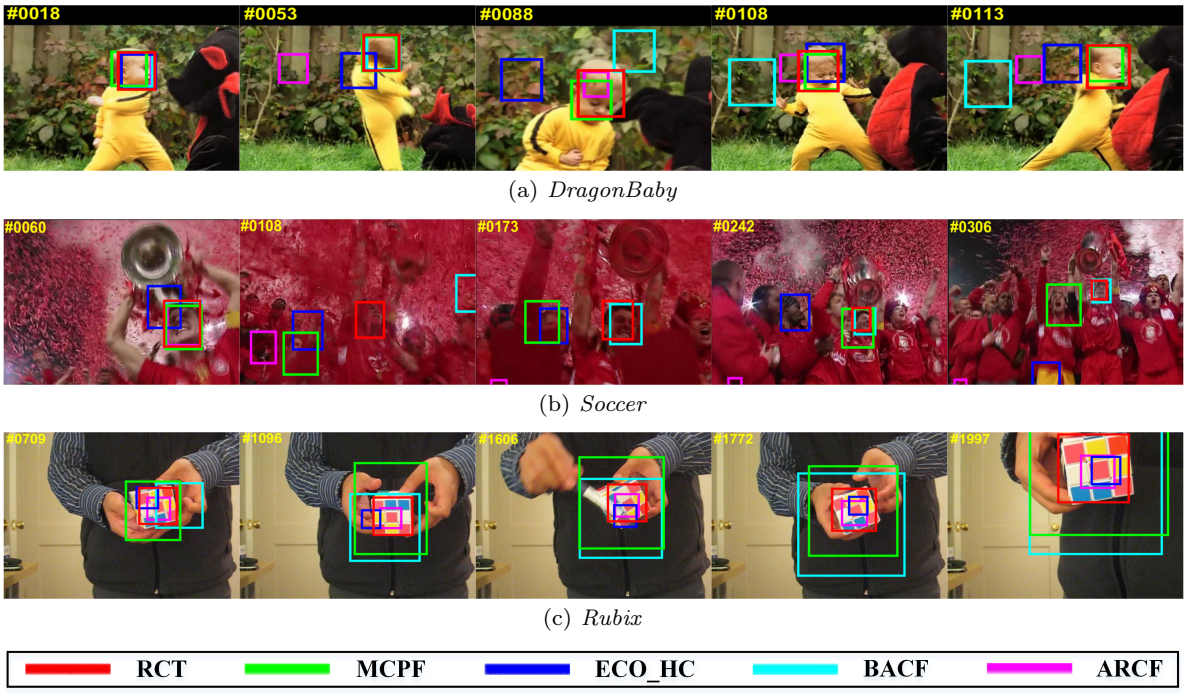


Fig. 6 Tracking results of RCT in qualitative comparison with state-of-the-art algorithms

quences with fast motions (*DragonBaby*), illumination variations (*Soccer*), scale variations (*Rubix*), in-plane and out-of-plane rotations (*DragonBaby*, *Soccer*, *Rubix*) can be succesfully tackled by our method without model drifts. Videos with motion blurs (*DragonBaby*, *Soccer*) and occlusions (*Soccer*) also benefit from our strategy of reliable response. It should be noted that for the *DragonBaby* and *Rubix* sequences, only our RCT tracker still keep estimating both the position and scale of the target accurately. To sum up, the proposed tracking algorithm can perform robustly in various tracking scenes and alleviate model drifts effectively.

Table 2 Tracking speed comparison over OTB50 benchmark

Trackers	Average fps	Real-time
BACF	47.19	Y
ECO_HC	65.02	Y
AutoTrack	64.82	Y
ARCF	14.27	N
KCC	100.92	Y
Staple_CA	59.94	Y
MCPF	0.51	N
SRDCFdecon	5.68	N
BIT	85.53	Y
RCT	26.45	Y

Table 3 Tracking performance comparison of the proposed tracker and its key components over OTB50 benchmark

Trackers	Averge fps	DP/AUC scores
Baseline	47.19	77.11/57.39
Baseline+FF	31.24	78.35/58.59
Baseline+RR	36.99	77.98/57.69
Baseline+FF+RR	26.45	80.50/59.68

4.4 Real-time performance

In addition to robust demands in challenging scenes, real-time performance is another essential requirement for online visual tracking. We present the tracking speed comparison over the OTB50 benchmark by the average FPS in Table 2. It can be shown that the ARCF, SRDCFdecon, MCPF trackers (<25 fps) can not meat the real-time requirement. They generally need to solve a complicated model formulation or extract deep features with a time-consuming procedure, which may limit their use in many real-time applications. On Intel core i7-9700 hardware environment, our RCT tracker operates at a real-time speed of 26.45 fps without using multi-threading or GPU. The tracking speed can be further improved by optimizing the code. Even so, RCT runs more than 50 times faster as compared to MCPF, which operates on an high-end NVIDIA GTX 1080Ti GPU with a measured tracking speed of 0.51 fps.

To verified the real-time performance and effectiveness of key components in our tracker, we also report the comparison results of the average tracking speed and DP/AUC scores over the OTB50 benchmark in Table 3. The basic notions are as follows: (1) 'Baseline' denotes the original BACF; (2) 'Baseline+FF' means the baseline tracker with our fused features; (3) 'Baseline+RR' stands for the baseline tracker with the desiged scheme of reliable response; (4) 'Baseline+FF+RR' is our final tracker RCT. From Table 3, we can see that both of the two modules operate efficiently without degrading the real-time performance of the baseline tracker. Further, they contribute to the substantial improvement on tracking accuracy over the baseline method.

5 Conclusion

In this paper, we have proposed a real-time correlation filter-based tracking method via the use of multi-channel fused features and reliable response maps. The correlation filter that utilizes multi-channel fused features leads to a significant improvement in tracking performance while dealing with challenging factors such as illumination variation and rotation. We have also proposed a novel strategy to obtain a more reliable response map, thereby locating the target through it. This allows our tracker to reduce the probability of incorrect locating when target occlusion and motion blur exist severely, so as to achieve the effect of suppressing model drift. Comparative experimental investigations have proven, both quantitatively and qualitatively that our approach has comparable performance with that of the state-of-the-art tracking methods. In particular, the proposed approach achieves a significant improvement in overall tracking performance compared to the baseline BACF. Meanwhile, our tracker is still able to maintain a real-time tracking speed of 26 fps. Our method still has shortcomings to be improved. For example, our tracker can not confirm staple tracking when facing long-term occlusion. Future work will involve investigating more powerful fused features with low dimension and more efficient tracking framework to deal with long-term occlusions for real-time applications. Besides, our method can be generalized to other areas of computer vision, such as human appellative [1].

Conflict of interest

The authors declare that they have no conflict of interest.

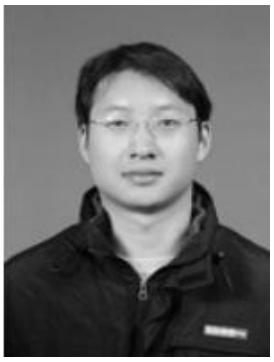
Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61871460; in part by the Shaanxi Provincial Key Research and Development Program under Grant 2020KW-003; in part by the Natural Science Foundation of Guangxi under Grant 2019GXNSFBA245056; and in part by the Sêr Cymru II Strategic Partner Acceleration Award Programme, U.K., under Grant 80761-AU201.

References

1. Ashraf, S., Aslam, Z., Sehrish, S., Afnan, S., Aamer, M.: Multi-biometric sustainable approach for human appellative. Computational Research Progress in Applied Science and Engineering **6**(3), 146–152 (2020)

2. Ashraf, S., Muhammad, D., Shuaeeb, M., Aslam, Z.: Development of shrewd cosmetology model through fuzzy logic. International Journal of Research in Engineering and Applied Sciences **5**(3), 93–99 (2020)
3. Ashraf, S., Saleem, S., Ahmed, T., Aslam, Z., Muhammad, U.: Conversion of adverse data corpus to shrewd output using sampling metrics. Visual Computing for Industry, Biomedicine, and Art **3**, 1–19 (2020)
4. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.: Staple: Complementary learners for real-time tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1401–1409. IEEE (2016)
5. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2544–2550. IEEE (2010)
6. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3**(1), 1–122 (2010)
7. Cai, B., Xu, X., Xing, X., Jia, K., Miao, J., Tao, D.: BIT: biologically inspired tracker. IEEE Transactions on Image Processing **25**(3), 1327–1339 (2016)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893. IEEE (2005)
9. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: ECO: efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6931–6939 (2017)
10. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision (CVPR), pp. 4310–4318. IEEE (2015)
11. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1430–1438 (2016)
12. Danelljan, M., Shahbaz Khan, F., Felsberg, M., Van de Weijer, J.: Adaptive color attributes for real-time visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1090–1097. IEEE (2014)
13. Galoogahi, H.K., Fagg, A., Lucey, S.: Learning background-aware correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1144–1152. IEEE (2017)
14. Galoogahi, H.K., Sim, T., Lucey, S.: Correlation filters with limited boundaries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4630–4638. IEEE (2015)
15. Hecht, S., Schlaer, S., Smith, E.L.: Intermittent light stimulation and the duplicity theory of vision. In: Cold Spring Harbor Symposia on Quantitative Biology, vol. 3, pp. 237–244. Cold Spring Harbor Laboratory Press (1935)
16. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 583–596 (2015)
17. Huang, Z., Fu, C., Li, Y., Lin, F., Lu, P.: Learning aberrance repressed correlation filters for real-time uav tracking. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2891–2900. IEEE (2019)
18. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE transactions on pattern analysis and machine intelligence **34**(7), 1409–1422 (2012)
19. Kristan, M., Matas, J., Leonardis, A., Vojíř, T., Pflugfelder, R., Fernández, G., Nebehay, G., Porikli, F., Čehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(11), 2137–2155 (2016)
20. Li, C., Liu, X., Su, X., Zhang, B.: Robust kernelized correlation filter with scale adaption for real-time single object tracking. Journal of Real-Time Image Processing **15**(3), 583–596 (2018)
21. Li, Y., Fu, C., Ding, F., Huang, Z., Lu, G.: Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11920–11929. IEEE (2020)
22. Li, Y., Zhu, J.: A scale adaptive kernel correlation filter tracker with feature integration. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 254–265. Springer (2014)
23. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: algorithms and benchmark. IEEE Transactions on Image Processing **24**(12), 5630–5644 (2015)
24. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3074–3082. IEEE (2015)
25. Mueller, M., Smith, N., Ghanem, B.: Context-aware correlation filter tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1396–1404. IEEE (2017)
26. Smeulders, Wm, A., Chu, D., M., Cucchiara, R., Calderara, S., Dehghan, A.: Visual tracking: an experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1442–1468 (2014)
27. Varfolomieiev, A.: Channel-independent spatially regularized discriminative correlation filter for visual object tracking. Journal of Real-Time Image Processing **18**(3), 233–243 (2021)
28. Wang, C., Zhang, L., Xie, L., Yuan, J.: Kernel cross-correlator. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 4179–4186. AAAI (2018)
29. Wu, Y., Lim, J., Yang, M.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(9), 1834–1848 (2015)
30. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2411–2418. IEEE (2013)
31. Zhang, S., Zhao, S., Sui, Y., Zhang, L.: Single object tracking with fuzzy least squares support vector machine. IEEE Transactions on Image Processing **24**(12), 5723–5738 (2015)
32. Zhang, T., Xu, C., Yang, M.H.: Multi-task correlation particle filter for robust object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4819–4827. IEEE (2017)

Authors' biography



Bin Lin received the M.S. degree in Chongqing University of Posts and Telecommunications, Chongqing, China, in 2012. He has been working as a lecturer at Guilin University of Technology, Guilin, China from 2013. He is currently pursuing the Ph. D. degree in Computer Science at Northwestern Polytechnical University, Xi'an, China. His research interests include visual tracking and deep learning techniques.

Xizhe Xue received the B.E. degree from Northwestern Polytechnical University, Xi'an, China in 2018. She is currently pursuing the Ph.D. degree in School of Computer Science, Northwestern Polytechnical University. Her research interests include visual tracking, digital image processing and deep learning techniques.

Ying Li received the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2002, Since 2003, she has been with the School of Computer Science, Northwestern Polytechnical University, Xi'an, where she is currently a Full Professor. Her current research interests include computation intelligence, image processing, and pattern recognition. She has published extensively in the above areas.

Qiang Shen received the Ph.D. degree in computing and electrical engineering from Heriot-Watt University, Edinburgh, U.K., in 1990, and the D.Sc. degree in computational intelligence from Aberystwyth University, Aberystwyth, U.K., in 2013. He holds the established chair of computer science and is the Pro Vice-Chancellor for the Faculty of Business and Physical Sciences, Aberystwyth University. He has authored two research monographs and 400+ peer-reviewed papers. Dr. Shen was a recipient of an Outstanding Transactions Paper Award from the IEEE.