Multimedia Tools and Applications (2019) 78:27271–27290
https://doi.org/10.1007/s11042-019-07828-2

A multiple feature fused model for visual object tracking via correlation filters

Di Yuan[1,2] · Xinming Zhang[2] · Jiaqi Liu[3] · Donghao Li[1]

Received: 4 October 2017 / Revised: 11 May 2019 / Accepted: 23 May 2019 /
Published online: 17 June 2019

Abstract

Common tracking algorithms only use a single feature to describe the target appearance, which makes the appearance model easily disturbed by noise. Furthermore, the tracking performance and robustness of these trackers are obviously limited. In this paper, we propose a novel multiple feature fused model into a correlation filter framework for visual tracking to improve the tracking performance and robustness of the tracker. In different tracking scenarios, the response maps generated by the correlation filter framework are different for each feature. Based on these response maps, different features can use an adaptive weighting function to eliminate noise interference and maintain their respective advantages. It can enhance the tracking performance and robustness of the tracker efficiently. Meanwhile, the correlation filter framework can provide a fast training and accurate locating mechanism. In addition, we give a simple yet effective scale variation detection method, which can appropriately handle scale variation of the target in the tracking sequences. We evaluate our tracker on OTB2013/OTB50/OBT2015 benchmarks, which are including more than 100 video sequences. Extensive experiments on these benchmark datasets demonstrate that the proposed MFFT tracker performs favorably against the state-of-the-art trackers.

Keywords Visual tracking · Multiple feature fusion · Adaptive weighting · Correlation filter · Scale detection

✉ Di Yuan
yuandi@stu.hit.edu.cn

Xinming Zhang
xinmingxueshu@hit.edu.cn

Jiaqi Liu
jiaqi_liu1993@163.com

Donghao Li
lidh@hit.edu.cn

[1] School of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China

[2] School of Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China

[3] College of Finance and Statistics, Hunan University, Changsha, China

1 Introduction

Visual object tracking is one of the most fundamental and challenging research problems in the computer vision area for its numerous applications in human-computer interaction, video surveillance, driverless vehicle, etc. Despite having achieved enormous signs of progress over the past decades, object tracking remains more challenges for designing a tracker which can handle critical situations (such as illumination variation, scale variation, deformation, occlusion, etc) perfectly.

There are many different tracking frameworks that attempt to improve tracking performance in different ways. Sparse representation based trackers by finding the best candidate with minimal reconstruction error using target templates [3, 4, 26]. DCFs-based trackers approximate the dense sampling scheme by generating a circulant matrix, of which each row denotes a circular shift of a base sample [12, 14, 23, 24, 60]. Deep learning based trackers often use CNN features and neural network structure to improve tracking performance [34, 37, 45, 47, 56]. Saliency depicts as an evaluation mechanism have been introduced into detection and tracking tasks in recent years [27, 28, 49]. In [49], Wang et al. present a salience-based tracking method, which can estimate object salience and environment salience of extracted visual features for robust visual tracking. For visual tracking, the appearance model is a crucial factor for object representation. Various feature descriptors with effective appearance models have been proposed in numerous literatures [21, 22, 43, 48, 51]. Single feature descriptor has been widely used in appearance based visual tracking models [10, 20, 24, 42, 44] for their computational convenience. The single feature is easily disturbed by noise, however, can not describe the appearance of the object target clearly. Due to different features can provide complementary information [18, 40, 52, 63], this paper is desirable to combine multiple feature descriptors to improve visual tracking performance.

Recently, several visual tracking methods based on multiple feature fusion have been established. The famous ensemble tracking [2] combines the HOG and RGB by using Adaboost algorithm. Ma et al. [41] using multiple feature fusion via weighted entropy to do data-adaptive visual tracking problem. A multi-view correlation filters tracker for enhancing the robustness of the tracker has been proposed in literature [33]. There are also some multiple feature fusion methods under the semi-supervised learning framework [62] and sparse representation-based framework [25]. Although those methods achieved some success, all of them are either limited by a larger computational cost or produce an unsatisfactory tracking performance.

To relieve these problems, we propose a novel multiple features fused tracking method into a correlation filter framework. This fusion method uses a simplest but effective adaptive weighted average of each feature, and the weight adaptively determined by the maximum response value of each feature. By using the complementarity among different features under different tracking scenarios, our model can eliminate the disadvantage of single feature easily affected by noise, which can enhance the ability to represent the appearance of a target. Based on the correlation filter framework, we can get the central coordinate of the target by finding the maximum response value in the response map for each feature. And then, through the adaptive weighted average calculation of the coordinates of the target center in each feature, the specific position of the target can be obtained. Meanwhile, the correlation filter framework provides a fast calculation mechanism, which can increase the speed of our tracking method.

The main contributions of this paper are summarized below:

- Based on the correlation filter framework, we propose a novel multiple feature fused model for visual object tracking. This model can adaptively combine these advantages of different features perfectly, and handle the disadvantage of a single feature which is susceptible to noise interference effectively. Meanwhile, the correlation filter framework is efficient for multiple feature fusion operation.
- We present a simple but effective scale variation detection mechanism based on the different response value between adjacent frames. This mechanism can enhance the robustness of our tracker in scale variation tracking sequences.
- On OTB evaluation benchmarks, our proposed algorithm achieves robust and promising tracking performance.

The rest of this paper is organized as follow. We first review some related works in Section 2, and present an adaptive weighted multiple features fused tracker via the correlation filter framework in Section 3, which includes the fundamental introduce about the KCF tracker, the adaptive weighted multiple features fused model, and the scale evaluation mechanisms. Section 4 describes the implementation details, the evaluation of our approach on comprehensive benchmark datasets and the comparison with some correlative and representative trackers. Finally, we give a brief conclusion about our work in Section 5.

2 Related works

As an extensive review on multi-feature learning and correlation filter framework beyond the scope of this paper, we review the works related to our approach including multiple feature based trackers and correlation filter based trackers.

2.1 Multiple feature based trackers

To deal with the limitations of one single feature in object visual tracking, several multiple feature fusion based visual tracking methods are established [15, 32, 33, 35, 46, 53, 54]. Galoogahi et al. [15] propose a multi-channel detector/filter in the frequency domain, which can improve the tracking performance obviously. In [46], Tang et al. derive a multi-kernel correlation filter based tracker which fully takes advantage of the invariance-discriminative power spectrums of various features to further improve the performance. Yin et al. [53] propose a generic likelihood map fusion framework to combine some different features into a fused soft segmentation which is suitable for mean-shift tracking. Li et al. [33] give a multi-view correlation model to enhances the robustness of the tracker. Qi et al. [45] suggest a hierarchical CNN based tracking framework, which takes full advantage of different features and uses an adaptive Hedge way to hedge these trackers into a stronger one. Literature [30] formulate the tracking problem as some basic observation and motion models corresponding to several features. The multiple basic models are constructed by SPCA and each of them is integrated with an interactive Markov Chain Monte Carlo scheme. These trackers achieved some good or robust performances, however, brought high computational costs.

2.2 Correlation filter based trackers

Correlation filter has been widely used in object detection, recognition, etc. Since Bolme et al. [7] propose the MOSSE tracker, correlation filters have been studied as a robust and efficient method to object visual tracking problem. Most improvements for the MOSSE tracker include the incorporation of kernel skill and HOG features [23, 59], the color name features [5, 12], the sparse representation [58], adaptive scale [10, 31, 35], and the integration of deep features [47, 56]. Henriques et al. [23] propose a CSK tracker and it can provide pretty performance and high calculation speed. In literature [24], the KCF method further improves the efficiency of CSK tracker by using HOG features and using kernel skill to transform the non-linear regression problem into linear regression. In [58], Zhang et al. exploit circulant structure property of target template to improve sparse representation based trackers. Yuan et al. [55] suggest a particle filter re-detection model in correlation filter framework, which can effectively reduce the occurrence of target loss by the tracker. In [11], a spatially regularized correlation filters method have been proposed to learn the filters from training examples with large spatial supports. Some local patches or parts based correlation filters trackers also have been developed [36, 38, 39] to improve the robustness of the trackers. Li et al. [36] introduce a reliable patch to exploit local contexts for tracking task. Liu et al. [38] propose a part based structural correlation filter to preserve the target structure for visual tracking. Although the correlation filter based trackers get better performance at current benchmarks and remaining computationally efficient, one single feature has its limitations and easily interfered by the noise, which cannot locate the object target accurately. In this paper, we propose an adaptive multi-features fused tracker via the correlation filter framework. Compared with these single feature correlation filter based trackers, our tracker exploits multiple features to enhance the robustness in dealing with various changes of the moving target and selects more discriminative features to ensure the tracking accuracy.

3 The adaptive weighted multiple features fused tracking method

According to use a single feature, the target tracking is always easy to be disturbed by noise, so that the tracking performance cannot reach the ideal state. In order to achieve a pretty tracking performance, we propose a novel multiple features fused tracker in correlation filters framework in this section.

Correlation filter based tracker using the information of image I and a filter w to get the center coordinate $x(i, j)$ of object target. The image is obtained from the m-th feature, and the target center coordinates are denoted as $x_m(i, j)$. In general, according to the Bayes formula, we know that:

$$P(x|I) = \int P(x|B)P(B|I)dB \approx \sum_{m=1}^{M} \omega_m P(x|B_m), \tag{1}$$

where M represents the number of features, ω_m demotes confidence in characteristic likelihood distributions, $\omega_m = P(B_m|I)$, and $\sum \omega_m = 1$.

3.1 Kernelized correlation filter tracker

The KCF [24] tracker is a representative of tracking by detection. It trains a classifier with all sub-windows of an image by dense sampling. Using kernel trick can make the data matrix of samples become highly structured. Meanwhile, using a fast Fourier transform can make the convolution of two images computed by an efficient element-wise product in the Fourier domain.

The KCF tracker uses a filter w, which is trained on an image patch x of $M \times N$ pixels with HOG features to model the appearance of the target. Let the circular shifts of $x_{m,n}$, (m,n) $\in \{0, 1, ...M-1\} \times \{0, 1, ..., N-1\}$ as training samples for the filter with Gaussian function label $y_{m,n}$. Minimizing the error between the training sample $x_{m,n}$ and the regression label $y_{m,n}$, we can get the filter w as:

$$w = \arg\min_{w} \sum_{m,n} |\langle \phi(x_{m,n}), w \rangle - y(m,n)|^2 + \lambda_1 \|w\|^2, \tag{2}$$

where ϕ represents the mapping to kernel space, $\langle . \rangle$ denotes the inner product, and λ is a regularization parameter. Since the label $y_{m,n}$ is not binary, the filter w learned from the training samples contains the coefficients of Gaussian ridge regression.

Using Fast Fourier Transform (FFT) to compute the Gaussian ridge regression problem, the objective function can rewrite as $w = \sum_{m,n} \alpha_(m,n)\phi(x_{m,n})$, thus (2) can be acquired by:

$$\alpha = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(y)}{\mathcal{F}(k^x) + \lambda}\right), \tag{3}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denotes denotes FFT and its inverse transformation (IFFT), respectively. In Fourier domain, the kernel correlation $k^x = \kappa(x_{m,n}, x)$ is computed by Gaussian kernel. The vector α contains all the $\alpha_{m,n}$ coefficients.

In the tracking process, an image patch z with the same size of x is cropped out in the new frame. And then, the response score can be calculated by:

$$f(z) = \mathcal{F}^{-1}(\mathcal{F}(k^z) \odot \mathcal{F}(\alpha)), \tag{4}$$

where $f(z)$ denotes the response map of patch z, $\odot$ denotes the element-wise product, $k^z = \kappa(z_{m,n}, \hat{x})$ and the $\hat{x}$ is the learned target appearance. When get $f(z)$, we can find the position of maximum response value in the map and let the position as the center coordinate of the target. Train new filter and update parameters according to the current position. After that, the steps are repeated so that the target tracking can be achieved in the entire sequence.

Although, with the circulate data matrices and the efficient element-wise product, the KCF achieves a fast and satisfactory performance. The single feature has its limitation in dealing with various changes in tracking sequences yet. In order to obtain a robust and pretty performance, we propose a novel multiple feature fused model in the efficient KCF tracking framework.

3.2 Adaptive weighted multiple features fused model

A multiple features fused model should exploit the complementary information of different features. The selection of features and fusion method directly affect the performance of the multiple feature fused model. For the correlation filter, the different features are suitable for fusion, due to the maximum response value which is used to determine the coordinates of the target. The simplest tracking instance of t-th frame can be seen in Fig. 1.

We propose to unify different feature under a probabilistic framework. For each feature t (where $t = 1, 2, ..., k$), its probability distribution is p_{ij}^t and $\sum p_{ij}^t = 1$, (i,j) $\in \{1, 2, ...M\} \times \{1, 2...., N\}$, where $M \times N$ is the size of an image patch. By using the correlation filter framework for visual tracking, we can get the coordinates (i, j) of the center of object target. Next, we choose these centers coordinates from the response map of each feature to determine the center coordinates of the target. For the sake of simplicity, we acquire the final coordinates of the target by the adaptively weighted average of each coordinate

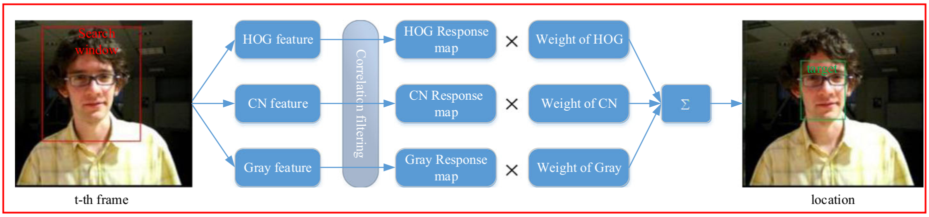


Fig. 1 The tracking framework of our approach. For t-th frame, different features are extracted from the original image within the preset search window(in our approach, we exploit HOG, CN, Gray features to do fusing). With correlation filtering, we can obtain the response maps of the three features. Different response maps show different maximum response values, due to each feature has different discriminatory powers in a variety of tracking scenarios. After the three response maps are fused with the fusion model, a more accurate and discriminative response map can be obtained, and the object target can be located accurately

position [24]. After multiple features fused, the center coordinates of the target are showed as follows:

$$\begin{aligned} &p_i = \lambda_1 * p_{i1} + \lambda_2 * p_{i2} + ... + \lambda_n * p_{in}, \\ &s.t. \sum \lambda_j = 1, \end{aligned} \tag{5}$$

where p_i denotes the i-th frame's center coordinates, p_{ij} is the i-th frame's center coordinates of j-th feature, λ_j is the corresponding weighting factors of j-th feature.

Traditionally, a good feature can obtain large response values in correlation filtering relatively, so the quality of the feature is very significant for determining the final target position. Based on this opinion, we adopt the maximum response values of each feature in the correlation filter to adaptively acquire its corresponding weight to get the target position, and the corresponding weight can calculate as:

$$\lambda_j = \frac{mR_j}{\sum mR_j}, \tag{6}$$

where mR_j denotes the maximum response value of j-th feature.

Since the weights obtained by simple weighted averaging can lead to excessive positional weights determined by the interference characteristics, we employ a simple penalty term $\frac{1}{\eta * mR_j}$ to solve this problem:

$$\begin{aligned} \lambda_j' &= \frac{mR_j}{\frac{1}{\eta * mR_j} + \sum mR_j}, \\ \lambda_j'' &= \frac{\lambda_j'}{\sum \lambda_j'}, \end{aligned} \tag{7}$$

where λ_j'' denotes the modified adaptive weighting factor of j-th feature, η denotes the penalty term coefficient. The purpose of the penalty item $\frac{1}{\eta * mR_j}$ is to obtain a large weight for the feature with a large response value, and to obtain a small weight for the feature with a small response value, respectively.

In this approach, we select the features from edge, color and intensity, which correspond to features of HOG [24], Color Names [9] and gray value. The HOG features are robust to illumination and deformation, which obtains excellent results in human detection and tracking [19, 24]. Color Names and gray value are robust to motion blur, which gives good results in image retrieval [9]. Given t-th frame image and the correlation filters F, we can get the center coordinates of the target with each feature: $p_c = F_c(I)$, $p_h = F_h(I)$, $p_g = F_g(I)$. After the fusion of color feature and HOG feature, the corresponding coordinates of the center are:

$$p_{ch} = \lambda_c * F_c + \lambda_h * F_h, \tag{8}$$

where p_{ch} denotes the center coordinates by the fusion of color features and HOG features.

After adaptively weighting, the noise of a single feature is filtered out by the response of another feature, so that the original nature of the target can be better represented. The other kinds of fusion are given by: $p_{cg} = \lambda_c * F_c + \lambda_g * F_g$, $p_{gh} = \lambda_g * F_g + \lambda_h * F_h$, correspondingly. From 5 to 8, we can obtain the objective function of the adaptive weighted multiple feature fusion model:

$$p_i = \sum \frac{F_{ji}(I) m R_{ji}}{\left(\frac{1}{\eta * m R_{ji}} + \sum m R_{ji}\right) \sum \frac{m R_{ji}}{\frac{1}{\eta * m R_{ji}} + \sum m R_{ji}}}. \tag{9}$$

From the previous description, we can see that the algorithm of multiple feature fusion only needs an adaptive weighted average of the selected features. From the intuitive point of view, the corresponding noise of target center position of each feature achieves good filtering. To fuse the maximum response values of each response map of corresponding features, we can determine the ultimate goal of the target position. In Fig. 1, it is obvious that the center point after fusion is more robust. Therefore, the model based on adaptive weighted multiple feature fusion can effectively improve the robustness of the algorithm.

3.3 Scale evaluation mechanism

Based on the correlation filter tracking framework, by finding the maximum value of the response map in each image frame, we can only obtain the center position of the object target. In visual tracking, scale change is one of the most common challenging aspects, however, which influences the tracking performance directly. In this section, we give a simple but effective scale evaluation mechanism based on the multiple feature fused model.

For correlation filter based framework, the initial target size set as $size_1 = [h_1, w_1]$. And then, we use the relationship of maximum response value between the current t-th frame and previous $t-1$-th frame to determine the size of the target in the current frame. In the case of not affecting the result, simply determine the magnitude of the maximum response value of the adjacent frame to determine the direction of the target change. By the property of the Gauss function, we can see that the other conditions remain constant, there is a negative correlation between the target size and the maximum response value. If the maximum response value of the current t-th frame is higher than the previous $t-1$-th frame, then the target size decreases. If the maximum response value of the current t-th frame is lower than the previous $t-1$-th frame, then the target size increases. If the maximum response value of the current t-th frame is the same as the previous $t-1$-th frame, then the target size remains unchanged. Using the change of the target size corresponding to the three features is taken as the weight to get the size change of the target, and the rate of change is expressed by c. So, for the size of the target in t-th frame can be determined by $t-1$-th frame target's size:

$$\begin{aligned} size_t &= size_{t-1} * c', \\ c' &= \frac{1}{3} * (c_c + c_h + c_g), \end{aligned} \tag{10}$$

where $size_t$ denotes the size of target in t-th frame, c_c, c_h and c_g denotes the rate of change of color feature, HOG feature and gray feature, respectively.

The scale change of the target will not be too obvious between the adjacent two frames, so the simple scale reduction is used to update the target scale.

4 Experiments

In order to evaluate the proposed tracker objectively and comprehensively, we test our tracker on a standard visual tracking benchmark. First, we introduce the algorithm flow and the experimental environment and details. And then, we give the details and standards of the experimental evaluation. Finally, the performance of our tracker is validated mainly based on the OTB2013/OTB50/OBT2015[50, 51] benchmarks, which contains more than 100 test video sequences.

Our method is implemented in MATLAB and the experiments run on a PC with an Intel Core-i3-4170 CPU (3.70 GHz) and 8 GB RAM.

4.1 Implementation details

In this section, we give a description of the whole tracking process and parameter settings. First, we can extract different features from the given initial bounding box in the first frame and trains the corresponding filters. And then, we run the tracker iteratively on each frame in the tracking sequences. In each iteration, we can determine the appropriate scale size and locate the center position of the target through the multiple features fused model, successively. Finally, we update the correlation filter models in a linear way. The whole process of our method can be seen in Algorithm 1.

Algorithm 1 The multiple feature fused tracking framework(MFFT).

```
Inputs: the initial target bounding box $b_1$ , the target size $sz$, the search window $sz$-$window$,the penalty term coefficient $\eta$, the initial tracking frame $I_1$ , the model learning rate $\gamma$.
Outputs: The position and scale of target in each frame.
Extract the target features from $I_1$ with area $b_1$;
Train the initial models $mod_1$ of with (3) and $b_1$;
if $t \leq T$ then
    (where $t$ is the number of current frame, $T$ is the totally number of the tracking frames)
    Evaluate the scale change and get the optimal scale factor $c'$ with (12);
    Crop out the search window with $sz$-$window * c'$ from the current frame $I_t$ and extract the features from the search window;
    Compute the correlation filter response maps of each feature with (4);
    Calculate objectifunction of multiple feature fusion model with (8), Eq. (10) and Eq. (11);
    Get the target position of the current frame $t$ and the current target size with (12);
    Get the current correlation filter model $mod_{t'}$ with the current target and update the correlation model of each feature as $mod_t = (1-\gamma)mod_{t-1} + \gamma mod_{t'}$;
end if
```

Parameters setting as follows: The search window size is twice of the target set as sz-$window = 2 * sz$, the scale change ratios are set as $c = [0.98, 1, 1.02]$ which depend on the scale change ratios of different features, and the penalty term coefficient set as $\eta = 25$.

We use the same parameters about the correlation filter as in [24] and the same parameters about the features as in [33].

4.2 Evaluation criterion

We use central location error (CLE) and Pascal VOC overlap ratio (VOR) to evaluate the effectiveness of our proposed tracking algorithm [8, 13].

Central position error (CLE): the mean Euclidean distance between the target center location coordinates determined by the algorithm and the true values of the artificial markers. The mean central position error of all frames in the sequence is used to evaluate the overall performance. In order to rank the performance of each tracking algorithm, authors usually used 20 pixels as the threshold with central position error are to measure the score. Pascal VOC overlap rate (VOR) can calculate as:

$$VOR = \frac{Area\{B_R \cap B_G\}}{Area\{B_R \cup B_G\}}, \tag{11}$$

where B_R denotes the bounding-box of the tracking result, B_G denotes the real bounding-box of the tracking target.

Under the VOR framework, we choose the number of those frames which VOR larger than the threshold value θ as the successful frames. The success plot figure shows the success threshold varies from 0 to 1 of the ratio of the success number of frames. By comparing the area under the success rate curve (AUC), the corresponding algorithms are sorted accordingly.

In order to evaluate the performance of the algorithm, we use three classes of evaluation indexes given by OTB2013 [51]: one-pass evaluation (OPE), temporal robustness evaluation (TRE), and spatial robustness evaluation (SRE). The OPE means for each test sequence in the evaluation set, only run the tracking algorithm one time, and statistics are over a certain threshold percentage of heart error and overlap. The TRE means runs 20 times with different start frames on each video sequence. The SRE means runs 12 times with different spatial perturbations.

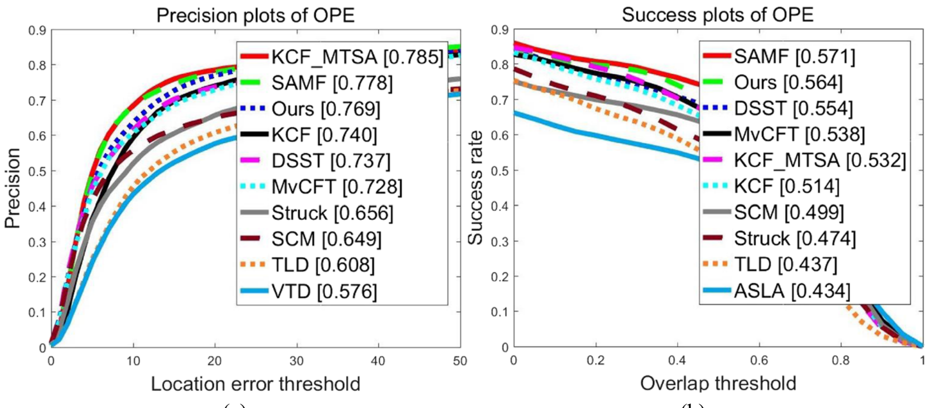
4.3 Evaluation with OTB benchmarks

Datasets The OTB2013 [51] dataset contains 51 different video sequences and categorizes these sequences with 11 attributes. OTB2015 [50] dataset is an extension of the OTB2013 dataset, which contains 100 different video sequences. OTB50 is a collection of 50 most challenging sequences selected from the OTB2015 dataset. The 11 attributes including: out-of-plane rotation (OR), in-plane rotation (IR), occlusion (OCC), scale variation (SV), illumination variation (IV), background cluster (BC), deformation (DEF), fast motion (FM), motion blur (MB), out of view (OOV) and low resolution (LR). Each of them has different sequences.

Baseline evaluation We compared our tracker with all the 29 trackers in OTB2013 benchmark including: Frag [1], MIL [3], ASLA [26], TLD [29], Struck [17], L1APG [4], CSK [23], SCM [61], etc. Besides five other representative trackers DSST [10], KCF [24], MvCFT [33], SAMF [35] and KCF_MTSA [6], are also compared with our tracker. The KCF [24] tracker is basically using a kernelized correlation filter operating on the HOG features. The DSST [10] tracker is extending the MOSSE tracker through the robust scale estimation and obtaining the top rank in performance by outperforming 19 state-of-the-art trackers on OTB and 37 state-of-the-art trackers on VOT2014. The MvCFT [33] tracker based on correlation filters proposed a multi-view model under a unified probabilistic framework. The SAMF tracker and the KCF_MTSA tracker are two widely used multi-feature fusion based trackers. The codes and settings are all the same with OTB2013 benchmark, which is widely approved. The comparison results are shown in Fig. 2. Compared with the KCF, MvCFT and DSST tracker, the tracking performance of our proposed algorithm is significantly improved. Moreover, compared with the SAMF and KCF_MTSA tracker, our proposed tracker is very closed to the best tracker both in precision and success plots. This demonstrates the effectiveness of our designed multi-feature fusion model.

Evaluation per attribute The success and precision plots of each attribute gives in Figs. 3 and 4. As we can see in Fig. 3, our tracker achieves the best performance in the attributes of MB, DEF, IV, OCC and OPR. It also achieves close to the best performance in other attributes. For Fig. 4, our tracker achieves the best performance in the attributes of FM, BC, DEF, IV, OCC and OPR. Meanwhile, it achieves the second or third performance in the attributes of MB, IR, OOV, LR and SV. These advantages benefit from the multiple feature fused model. For scale variation, our result is very close to the best result (DSST) who mainly considers the scale evaluations both in success and precision plots. It also shows the effectiveness of our scale estimation mechanism. For low resolution, our algorithm combines multiple features but the feature has a poor characterization ability, which caused the unsatisfactory results. Generally speaking, our proposed tracker achieves the best or close to the best results in almost all the attributes.

Robustness to initialization In order to give sufficient experimental contrast results to verify the robustness of our tracker, we evaluated it towards different spatial and temporal initialization using two robustness metrics: TRE and SRE. Fig. 5 shows the overall comparison performance on SRE and TRE. From Fig. 5b, we can see that our tracker achieves the second best performance on AUC success plots, which is close to DSST and better than KCF. On precision plots Fig. 5a, our MFFT tracker gets the best performance. From Fig. 5c,



(a) (b)

Fig. 2 Precision and success plots on OTB2013. **a** The precision plots; **b** The success plots. The numbers in the legend indicate the representative precision at 20 pixels for precision plots, and the average area-under-curve scores for success plots. To illustrate the problem, we only given the top 10 trackers

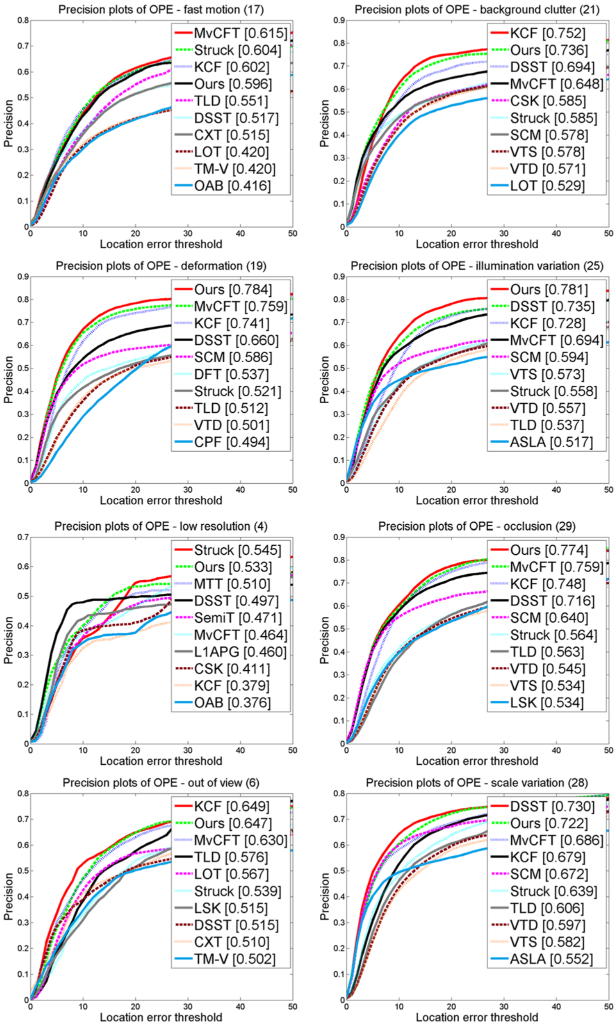


Fig. 3 The precision plots of fast motion (FM), background cluster (BC), motion blur (MB), deformation (DEF), illumination variation (IV), in-plane rotation (IR), low resolution (LR), occlusion (OCC), out-of-plane rotation (OR), out of view (OOV) and scale variation (SV)

d, we know that both the precision and the success plots show our tracker achieving the best performance. The results on TRE shows the robustness of our tracker on initialization in the first frame by shifting or scaling the ground truth. In summary, our MFFT tracker is effective and achieve a promising result in the visual tracking OTB2013 [51] benchmark.

Comparison to state-of-art trackers To put the tracking performance into perspective, we compare our tracker with the most recent state-of-the-art trackers including: 1) deep learning based trackers: HDT [45], CNT [56], CFNet-conv1 [47]; 2) correlation filter based trackers: KCF [24], DSST [10], CSK [23], STC [57]; 3) multi-feature based trackers: MvCFT [33], SAMF [35], KCF_MTSA [6]; and 4) representative trackers: TGPR [16], SCM [61],

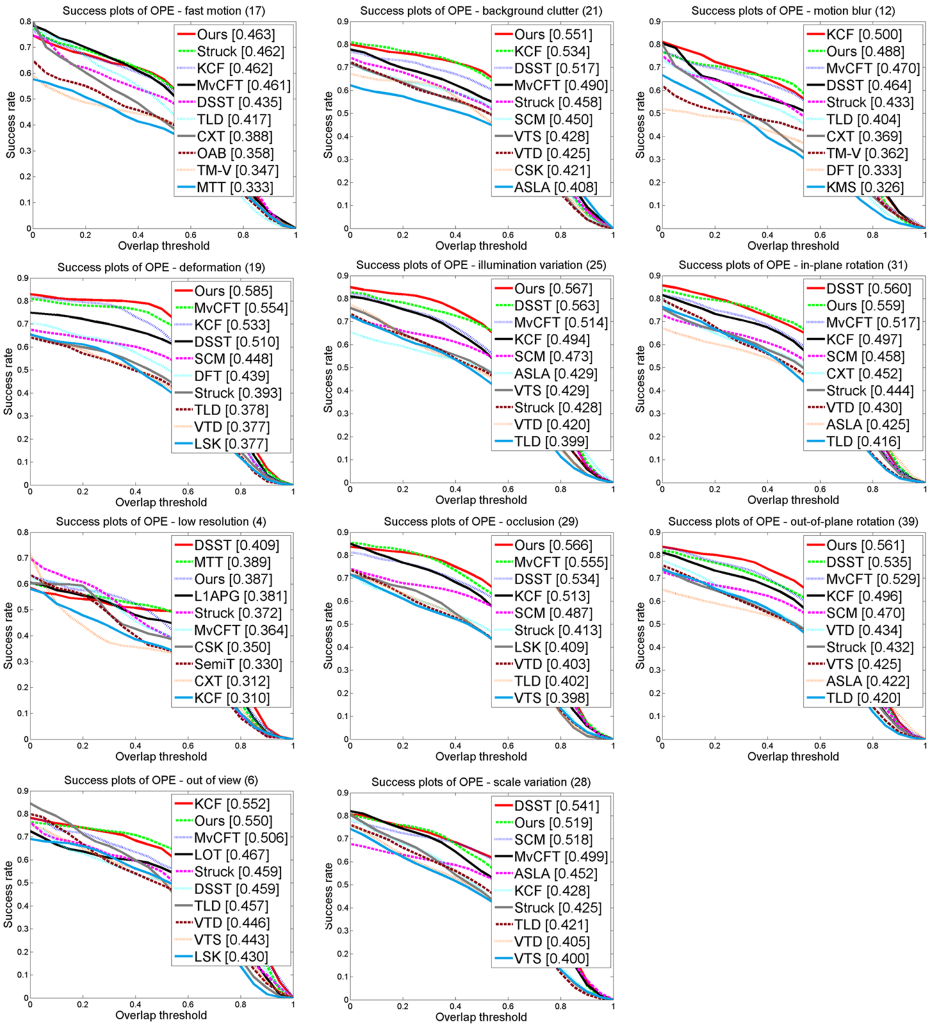
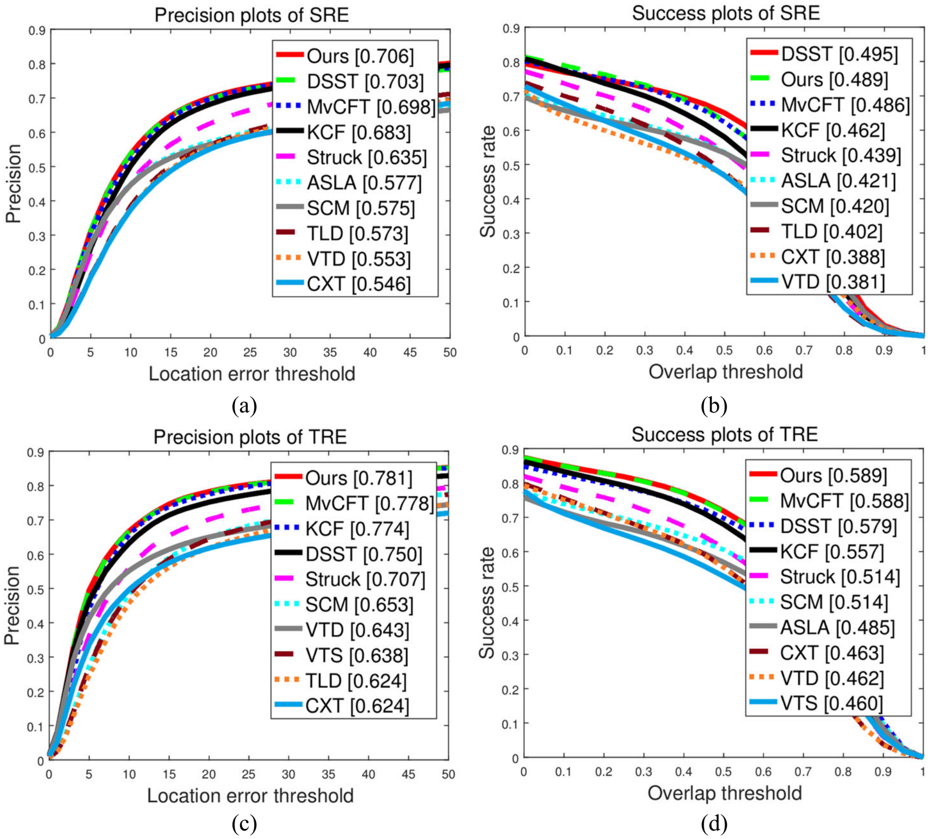


Fig. 4 The success plots of fast motion (FM), background cluster (BC), motion blur (MB), deformation (DEF), illumination variation (IV), in-plane rotation (IR), low resolution (LR), occlusion (OCC), out-of-plane rotation (OR), out of view (OOV) and scale variation (SV)

Struck [17]. We analyze the performance of our MFFT tracker with other nine state-of-the-art tracking algorithms under different attributes in OTB2015 [50]. Table 1 shows the comparison results on these 11 attributes. From this table, we can know that both in distance precision rates (DPR) and overlap success rates (OSR), our MFFT tracker achieves the best or closes the best results under all 11 attributes.

In addition, we use the OTB2013/OTB50/OTB100 datasets to a quantitative comparison of distance precision rate (%) (DPR) at a threshold of 20 pixels and overlap success rate (%) (OSR) at an overlap threshold of 0.5 in Table 2. From this table, we can see that our MFFT tracker achieves the best or close the best tracking results. Comparing with the correlation filter based trackers: KCF [24], DSST [10], CSK [23], and the representative trackers:



(a) (b) (c) (d)

Fig. 5 SRE and TRE precision and success plots on OTB2013. **a** The precision plots of SRE. **b** The success plots of SRE. **c** The precision plots of TRE. **d** The success plots of TRE. The numbers in the legend indicate the representative precision at 20 pixels for precision plots, and the average area-under-curve scores for success plots. To illustrate the problem, we only given the top 10 trackers

TGPR [16], SCM [61], Struck [17], our tracker have achieved better tracking performance than these trackers. Comparing with the multi-feature based trackers: MvCFT [33], SAMF [35], KCF_MTSA [6], our trackers have achieved similar tracking performance as these trackers. Even compared with the deep learning based trackers, the tracking performance of our tracker is also better than the CNT [56] tracker. These advantages are due to our adaptive weighted multi-feature fusion model. All the experimental results show that our MFFT tracker is comparable to other state-of-the-art trackers.

Qualitative comparison Our approach significantly improves the performance compared with the single feature based trackers in some complex cases. Figure 6 shows a qualitative comparison of our approach with some existing methods on some challenging tracking sequences. Whether the target scale changes (e.g., Car1 and Human6) or the target is occluded (e.g., Girl2 and Tiger2), our tracker can give a better tracking result than other trackers. Despite no explicit illumination variation handling component, our tracker performs favorably in cases with illumination variation (e.g., Human2).

Table 1 Average precision and success scores of our MFFT and KCF [24], DSST [10], MvCFT [33], CSK [23], CNT [56], TGPR [16], SCM [61], Struck [17], HDT [45] trackers on OTB2015 [50] dataset on 11 attributes including: background cluttered (BC), deformation (DEF), fast motion (FM), in-plane rotation (IPR), illumination variation (IV), low resolution (LR), motion blur (MB), occlusion (OCC), out-of-plane rotation (OPR), out-of-view (OV) and scale variation (SV)

Attribute	MFFT	KCF	DSST	MvCFT	CSK	CNT	TGPR	SCM	Struck	HDT
BC	72.9/55.2	71.2/49.7	70.4/52.1	67.0/50.8	58.2/41.8	62.4/49.0	59.3/42.8	57.7/46.2	54.7/42.6	84.4/57.8
DEF	66.5/49.0	61.7/43.6	57.0/43.4	64.8/46.8	45.3/33.8	52.4/39.8	63.0/45.5	52.4/40.2	52.7/38.3	82.1/53.4
FM	61.4/48.9	61.9/44.9	58.3/47.0	63.6/49.5	40.3/33.2	37.7/32.6	50.7/39.8	34.9/31.9	60.0/45.0	80.2/55.0
IPR	70.3/51.0	69.3/46.5	71.3/51.0	69.9/50.2	51.5/38.0	55.3/41.3	65.9/46.2	54.3/40.8	62.5/44.6	84.4/55.5
IV	76.7/57.9	70.7/47.4	72.6/55.9	69.2/52.0	48.2/36.8	56.7/46.2	63.3/45.2	59.7/48.7	54.9/42.0	82.0/53.5
LR	55.2/40.2	54.5/30.6	58.1/38.9	62.7/45.3	36.7/25.1	57.9/41.0	62.9/37.8	55.8/38.1	62.8/34.7	76.6/42.0
MB	67.3/53.5	61.7/45.7	59.7/49.1	61.2/47.4	38.4/32.3	36.9/35.8	50.8/40.9	31.6/30.8	58.0/45.1	79.4/56.3
OCC	68.2/50.7	62.1/43.8	60.9/46.0	65.8/48.4	43.1/33.2	55.4/43.4	59.4/42.9	54.9/42.2	52.4/38.7	77.4/52.8
OPR	71.2/52.0	67.0/45.0	66.5/48.1	67.3/49.3	48.4/35.5	57.6/43.6	64.2/45.5	56.9/43.1	59.3/42.4	80.5/53.3
OV	58.4/45.4	49.8/39.4	48.0/38.5	60.0/44.1	29.3/26.6	37.4/34.1	49.3/37.3	42.3/33.3	46.0/35.7	66.3/47.2
SV	68.0/50.0	63.9/39.9	66.4/48.5	65.8/48.0	45.7/32.7	53.1/41.7	59.1/39.9	56.5/43.6	59.7/40.3	81.1/48.9
average	71.9/53.6	69.1/47.5	69.4/52.0	69.8/51.7	51.8/38.4	58.4/45.4	64.3/45.8	57.2/44.5	63.4/45.9	84.8/56.4

The value format of each table cell is “DPR/OSR (%)”

Table 2 Comparisons with state-of-the-art tracking methods include: STC [57], KCF [24], DSST [10], MvCFT [33], SAMF [35], KCF_MTSA [6], CFNet-conv1 [47], HDT [45], CNT [56], TGPR [16], SCM [61] and Struck [17] on OTB2013 [51], OTB50 and OTB2015 [50] datasets

Dataset	Evaluation	MFFT	Correlation filters trackers			Muti-features trackers			Deep learning trackers			Representative trackers		
	Criterion	(Ours)	STC	KCF	DSST	MvCFT	SAMF	KCF_MTSA	CFNet-conv1	HDT	CNT	TGPR	SCM	Struck
OTB2013	DPR	76.9	54.7	74.0	73.7	72.8	77.8	78.5	77.6	88.9	72.3	76.6	64.9	65.6
	OSR	56.3	34.7	51.4	55.4	53.8	57.1	53.2	57.8	60.3	54.5	52.9	49.9	47.4
OTB50	DPR	63.4	43.1	61.1	62.7	62.9	64.3	64.8	65.3	80.4	50.1	61.2	48.1	52.9
	OSR	46.4	27.4	40.3	46.4	45.8	46.0	41.5	48.8	51.5	36.9	42.9	36.4	37.6
OTB2015	DPR	71.9	50.7	69.1	69.4	69.8	74.34	74.4	71.3	84.8	58.4	64.3	57.2	63.4
	OSR	53.6	31.9	47.5	52.0	51.7	53.5	49.7	53.6	56.4	45.4	45.8	44.5	45.9

Our MFFT tracker is almost outperforms existing approaches in DPR at 20 pixels and OSR at 0.5

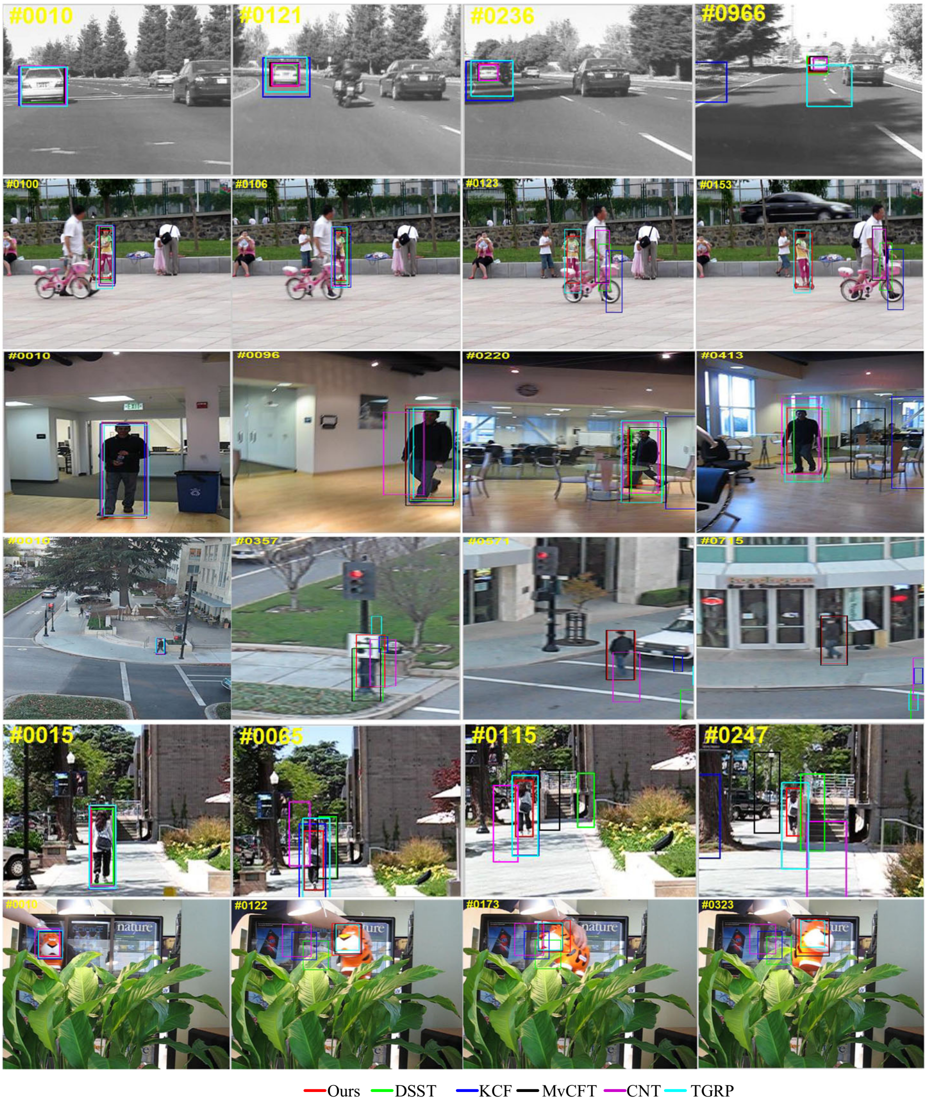


Fig. 6 Qualitative comparison of our approach with state-of-the-art trackers on the Car1, Girl2, Human2, Human6, Human7 and Tiger2 videos. Our approach provides consistent results in challenging scenarios, such as occlusions, fast motion, illumination variation, background clutter and target rotations

5 Conclusion

In this paper, we propose a multiple feature fused tracker in the correlation filter framework to achieve a pretty performance on OTB2013/OTB50/OTB2015 benchmarks. The multiple feature fused model can apply different features to deal with various changes of the target in the tracking sequences. This method can adaptively exploit the complementary information between different features to handle the weakness of a single feature that is easily susceptible to noise. And the correlation filter can provide an efficient fusing and tracking framework. Besides, we give a novel scale evaluation mechanism to deal with the moving target with scale change in the tracking sequences. The experiment results with different attributes show the competitive performance of our tracker.

Acknowledgment This research was supported by the Shenzhen Research Council (Grant No. JCYJ20160406161948211, JCYJ20160226201453085, JSGG20150331152017052, JCYJ20160531194006833), by the National Natural Science Foundation of China (Grant No. 61672183, 61272366, 61672444), by Science and Technology Planning Project of Guangdong Province (Grant No. 2016B090918047).

References

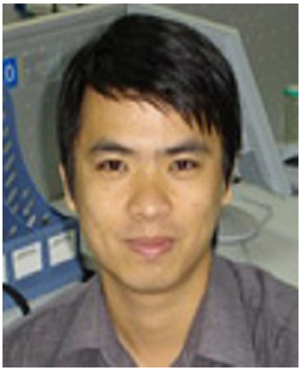
1. Adam A, Rivlin E, Shimshoni I (2006) Robust fragments-based tracking using the integral histogram. In: IEEE conference on computer vision and pattern recognition, pp 798–805
2. Avidan S (2007) Ensemble tracking. IEEE Trans Pattern Anal Mach Intell 29(2):261–271
3. Babenko B, Yang MH, Belongie S (2009) Visual tracking with online multiple instance learning. In: IEEE conference on computer vision and pattern recognition, pp 983–990
4. Bao C, Wu Y, Ling H, Ji H (2012) Real time robust $l1$ tracker using accelerated proximal gradient approach. In: IEEE conference on computer vision and pattern recognition, pp 1830–1837
5. Bertinetto L, Valmadre J, Golodetz S, Miksik O, Torr P (2016) Staple: complementary learners for real-time tracking. In: IEEE conference on computer vision and pattern recognition, pp 1401–1409
6. Bibi A, Ghanem B (2015) Multi-template scale-adaptive kernelized correlation filters. In: IEEE international conference on computer vision workshop, pp 613–620
7. Bolme DS, Beveridge JR, Draper BA, Lui YM (2010) Visual object tracking using adaptive correlation filters. In: IEEE conference on computer vision and pattern recognition, pp 2544–2550
8. Cehovin L, Kristan M, Leonardis A (2014) Is my new tracker really better than yours? In: Applications of computer vision, pp 540–547
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE conference on computer vision and pattern recognition, pp 886–893
10. Danelljan M, Hager G, Khan FS, Felsberg M (2014) Accurate scale estimation for robust visual tracking. In: British machine vision conference, vol 65, pp 1–11
11. Danelljan M, Hager G, Khan FS, Felsberg M (2015) Learning spatially regularized correlation filters for visual tracking. In: IEEE international conference on computer vision, pp 4310–4318
12. Danelljan M, Khan FS, Felsberg M, Weijer JVD (2014) Adaptive color attributes for real-time visual tracking. In: IEEE conference on computer vision and pattern recognition, pp 1090–1097
13. Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
14. Fan N, Li J, He Z, Zhang C, Li X (2019) Region-filtering correlation tracking. Knowl-Based Syst 172:95–103
15. Galoogahi HK, Sim T, Lucey S (2013) Multi-channel correlation filters. In: IEEE international conference on computer vision, pp 3072–3079
16. Gao J, Ling H, Hu W, Xing J (2014) Transfer learning based visual tracking with gaussian processes regression. European Conference on Computer Vision, 188–203
17. Hare S, Golodetz S, Saffari A et al (2016) Struck: Structured output tracking with kernels. IEEE Trans Pattern Anal Mach Intell 38(10):2096–2109
18. He Z, Chung AC (2010) 3-D b-spline wavelet-based local standard deviation (bwlsd): its application to edge detection and vascular segmentation in magnetic resonance angiography. Int J Comput Vis 87(3):235–265
19. He Z, Li X, You X, Tao D, Tang Y (2016) Connected component model for multi-object tracking. IEEE Trans Image Process 25(8):3698–3711
20. He Z, Yi S, Cheung Y-M, You X, Tang Y (2017) Robust object tracking via key patch sparse representation. IEEE Trans Cybern 47:354–364
21. He Z, You X, Tang Y (2008) Writer identification of chinese handwriting documents using hidden markov tree model. Pattern Recogn 41(4):1295–1307
22. He Z, You X, Zhou L, Cheung Y-M, Du J (2010) Writer identification using fractal dimension of wavelet subbands in gabor domain. Integrated Computer Aided Engineering 17(17):157–165
23. Henriques JF, Caseiro R, Martins P, Batista J (2012) Exploiting the circulant structure of tracking-by-detection with kernels. In: European conference on computer vision, pp 702–715

24. Henriques JF, Rui C, Martins P, Batista J (2014) High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell 37(3):583–596
25. Hong Z, Mei X, Prokhorov D, Tao D (2014) Tracking via robust multi-task multi-view joint sparse representation. In: IEEE international conference on computer vision, pp 649–656
26. Jia X, Lu H, Yang MH (2012) Visual tracking via adaptive structural local sparse appearance model. In: IEEE conference on computer vision and pattern recognition, pp 1822–1829
27. Jian M, Lam K, Dong J, Shen L (2015) Visual-patch-attention-aware saliency detection. IEEE Trans Cybern 45(8):1575–1586
28. Jian M, Qiang Q, Dong J, Yin Y, Lam KM (2018) Integrating qdwd with pattern distinctness and local contrast for underwater saliency detection ¡î. J Vis Commun Image Represent 53:31–41
29. Kalal Z, Mikolajczyk K, Matas J (2012) Tracking-learning-detection. IEEE Trans Pattern Anal Mach Intell 34(7):1409–1422
30. Kwon J, Lee KM (2010) Visual tracking decomposition. In: IEEE conference on computer vision and pattern recognition, pp 1269–1276
31. Li F, Yao Y, Li P, Zhang D, Zuo W, Yang MH (2017) Integrating boundary and center correlation filters for visual tracking with aspect ratio variation. In: IEEE international conference on computer vision workshop, pp 2001–2009
32. Li X, Liu Q, Fan N, He Z, Wang H (2019) Hierarchical spatial-aware siamese network for thermal infrared object tracking. Knowl-Based Syst 166:71–81
33. Li X, Liu Q, He Z, Wang H, Zhang C, Chen WS (2016) A multi-view model for visual tracking via correlation filters. Knowl-Based Syst 113:88–99
34. Li X, Ma C, Wu B, He Z, Yang M. (2019) Target-aware deep tracking, arXiv:1904.01772
35. Li Y, Zhu J (2014) A scale adaptive kernel correlation filter tracker with feature integration. In: European conference on computer vision, pp 254–265
36. Li Y, Zhu J, Hoi SCH (2015) Reliable patch trackers: robust visual tracking by exploiting reliable patches. In: IEEE conference on computer vision and pattern recognition, pp 353–361
37. Liu Q, Lu X, He Z, Zhang C, Chen W (2017) Deep convolutional neural networks for thermal infrared object tracking. Knowl-Based Syst 134:189–198
38. Liu S, Zhang T, Cao X, Xu C (2016) Structural correlation filter for robust visual tracking. In: IEEE conference on computer vision and pattern recognition, pp 4312–4320
39. Liu T, Wang G, Yang Q (2015) Real-time part-based visual tracking via adaptive correlation filters. In: IEEE conference on computer vision and pattern recognition, pp 4902–4912
40. Lu X, Lei H, Hao Z (2010) Automatic camshift tracking algorithm based on multi-feature. J Comput Appl 30(3):650–652
41. Ma L, Lu J, Feng J, Zhou J (2016) Multiple feature fusion via weighted entropy for visual tracking. In: IEEE international conference on computer vision, pp 3128–3136
42. Ma X, Liu Q, He Z, Zhang X, Chen WS (2016) Visual tracking via exemplar regression model. Knowl-Based Syst 106:26–37
43. Ou W, You X, Tao D, Zhang P, Tang Y, Zhu Z (2014) Robust face recognition via occlusion dictionary learning. Pattern Recogn 47(4):1559–1572
44. Ou W, Yuan D, Liu Q, Cao Y (2018) Object tracking based on online representative sample selection via non-negative least square. Multimed Tools Appl 77(9):10569–10587
45. Qi Y, Zhang S, Qin L, Yao H, Huang Q, Lim J, Yang MH (2016) Hedged deep tracking. In: IEEE conference on computer vision and pattern recognition, pp 4303–4311
46. Tang M, Feng J (2015) Multi-kernel correlation filter for visual tracking. In: IEEE international conference on computer vision, pp 3038–3046
47. Valmadre J, Bertinetto L, Henriques JF, Vedaldi A, Torr PHS (2017) End-to-end representation learning for correlation filter based tracking. In: IEEE conference on computer vision and pattern recognition, pp 2805–2813
48. Wang N, Shi J, Yeung DY, Jia J (2015) Understanding and diagnosing visual tracking systems. In: IEEE international conference on computer vision, pp 3101–3109
49. Wang Q, Tang S, Zhai D, Hu X (2018) Salience based object tracking in complex scenes. Neurocomputing 314:132–142
50. Wu Y, Lim J, Yang M-H (2015) Object tracking benchmark. IEEE Trans Pattern Anal Mach Intell 37(9):1834–1848
51. Wu Y, Lim J, Yang MH (2013) Online object tracking: a benchmark. In: IEEE conference on computer vision and pattern recognition, pp 2411–2418

52. Yi S, Lai Z, He Z, Cheung Y-M, Liu Y (2017) Joint sparse principal component analysis. Pattern Recogn 61:524–536
53. Yin Z, Porikli F, Collins RT (2008) Likelihood map fusion for visual object tracking. In: IEEE workshop on applications of computer vision, pp 1–7
54. Yuan D, Lu X, Li D, He Z, Luo N (2017) Multiple feature fused for visual tracking via correlation filters. In: International conference on security, pattern analysis, and cybernetics, pp 88–93
55. Yuan D, Lu X, Li D, Liang Y, Zhang X (2018) Particle filter re-detection for visual tracking via correlation filters. Multimed Tools Appl, pp 1–25
56. Zhang K, Liu Q, Wu Y, Yang MH (2016) Robust visual tracking via convolutional networks without training. IEEE Trans Image Process 25(4):1779–1792
57. Zhang K, Zhang L, Liu Q, Zhang D, Yang MH (2014) Fast visual tracking via dense spatio-temporal context learning. In: European conference on computer vision, pp 127–141
58. Zhang T, Bibi A, Ghanem B (2016) In defense of sparse tracking: circulant sparse tracker. In: IEEE conference on computer vision and pattern recognition, pp 3880–3888
59. Zhang T, Xu C, Yang MH (2017) Multi-task correlation particle filter for robust object tracking. In: IEEE conference on computer vision and pattern recognition, pp 4819–4827
60. Zhang T, Xu C, Yang MH (2019) Learning multi-task correlation particle filters for visual tracking. IEEE Trans Pattern Anal Mach Intell 41(2):365–378
61. Zhong W, Lu H, Yang MH (2012) Robust object tracking via sparsity-based collaborative model. In: IEEE conference on computer vision and pattern recognition, pp 1838–1845
62. Zhou Y, Rao C, Lu Q, Bai X, Liu W (2011) Multiple feature fusion for object tracking. In: Sino-foreign-interchange conference on intelligent science and intelligent data engineering, pp 145–152
63. Zhou Z, Wu D, Peng X, Zhu Z, Luo K (2014) Object tracking based on camshift with multi-feature fusion. J Softw 9(1):147–153

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Di Yuan graduated from Harbin Institute of Technology Shenzhen Graduate School, China in 2017. He is pursuing the Ph.D degree in Computer Science with the research in statute of Biocomputing, School of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, China. His current research interests include object tracking, machine learning and kernel methods.



Xinming Zhang received his MS degree in applied mathematics from Harbin Institute of Technology, Harbin, China, in 2003 and his PhD in general mechanics and mechanics basis from Harbin Institute of Technology, Harbin, China, in 2006. Currently, he is an associate professor at the School of Science at Harbin Institute of Technology (Shenzhen), Shenzhen, China. His current research interests include Inverse problem of differential equations, lattice Boltzmann method, and parallel algorithm.

Jiaqi Liu graduated from Jishou University, China in 2018. He is pursuing the Ph.D degree in Statistics, College of Finance and Statistics, Hunan University, China. His current research interests include experiment design, statistical computing and statistical learning.

Donghao Li is pursuing his master's degree in Computer Science with the research institute of Biocomputing, School of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, China. His current research interests include machine learning and computer vision.