IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Long-term adaptive tracking via complementary trackers

ISSN 1751-9659
Received on 4th September 2018
Revised 27th March 2019
Accepted on 8th May 2019
E-First on 18th June 2019
doi: 10.1049/iet-ipr.2018.6142
www.ietdl.org

Weicong Dai[1,2] ✉, *Longxu Jin*[1], *Guoning Li*[1]

[1]*Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130000, Jilin, People's Republic of China*

[2]*University of Chinese Academy of Sciences, Beijing, 100049, Beijing, People's Republic of China*

✉ *E-mail: daiweicong16@mails.ucas.ac.cn*

Abstract: In recent years, kernelised correlation filter-based trackers have been employed to manage short-term tracking problems and help long-term trackers achieve excellent accuracy and robustness under challenging conditions, such as geometry/photometry changes, heavy occlusion, fast motion, motion blur, and out-of-camera view. Nonetheless, the inherent boundary effects and risky update strategy of correlation filters constrain the performance of short-term tracking, which limits the performance of long-term trackers. Moreover, the complicated redetection module leads to high-computational cost, which results in the long-term trackers to run at a low speed, thereby significantly restricting their applications. In the present work, the authors propose to employ complementary trackers in designing an efficient long-term tracker. Furthermore, a sigmoid penalty coefficient is proposed to update the tracking model with an adaptive learning rate that adjusts the learning rate while the target encounters appearance variation. Finally, they propose a novel redetection method that combines a redetection classifier with a short-term component to redetect the target while satisfying the explicit condition. The long-term tracker proposed in this study is proven to perform real-time speed of more than 65 frames per second and state-of-the-art accuracy by the experimental result on several challenging benchmarks.

1 Introduction

Generic visual object tracking is one of the fundamental problems in computer vision applications and has a broad developmental prospect in video surveillance, robotics, and autonomous vehicle navigation [1]. The task of generic object tracking can be summarised as follows. Given a single target specified by a bounding box in the first frame, the location where the target will appear in all other frames in sequences is estimated. Although visual object tracking has long been proposed, its application in many practical applications remains a challenge.

In recent years, the emergence of correlation filters has caused significant progress in visual object tracking, which augments the samples to promote the discriminative ability and achieve a high-computational efficiency due to the circulant matrix. However, despite its recent advancements, visual object tracking remains a great challenge due to the large appearance variation caused by deformation, fast motion, illumination change, heavy occlusion, motion blur, out-of-plane rotation, and out-of-camera view [2, 3].

First, the boundary effects [4] have demonstrated significant influence on the tracking performance of correlation filter-based trackers. The poor performance of such trackers significantly limited the performance of correlation filter-based long-term trackers. To enhance the performance of correlation filters without considerably increasing the computation cost, we develop a colour-based tracker that complements kernelised correlation filters (KCFs) and mitigates the boundary effects (Fig. 1).

Second, the correlation filter-based trackers update the model frame-by-frame while tracking the target. However, the vast majority of state-of-the-art trackers [5–7] only update the tracking model with a constant learning rate. When the target encounters a significant deformation, the inaccurate tracking detection of the tracker produces corrupted training samples. The constant learning rate may lead to tracking failure due to model drifts [8] caused by updating the tracking model with corrupted samples. In addition, although targets are correctly detected, occlusions and background clutter will also lead to a decrease in the discriminative ability of the trackers. Therefore, adjusting the learning rate with the sample quality change is important. In this work, we address this problem by updating the tracking model with a sigmoid penalty coefficient that relates to the past and current maximum values of the response map.

Finally, the correlation filter-based trackers update the model to adapt target appearance variations rapidly with high learning rates, which results in the trackers only remembering the samples on the latest dozens of frames (Fig. 2). The tracking model is rapidly corrupted and causes tracking failure while it is updated with noisy training samples over a period of time. Hence, the correlation filter-based trackers can only address short-term tracking problems. The correlation filter based-trackers have adopted a long-term component to construct long-term tracker in recent works [5, 9, 10]. The long-term component aims to learn the appearance of the confident result conservatively and redetect the target to continue tracking while experiencing short-term component tracking failure. However, some critical problems concerning long-term trackers still remain. Primarily, the existing long-term trackers [5] activate the online trained classifier when the confidence scores are less than the given threshold and accept the result of the redetection classifier as the result of the long-term tracker. The existing criteria that control the long-term components are indistinct, which results in limited performance of the long-term trackers. Furthermore, existing long-term trackers [9, 10] run at a low speed due to their complicated structure and high-computational expenses of the long-term component, which significantly limits the real-world application of the long-term trackers. As previously mentioned, strengthening and simplifying the redetection module of the long-term trackers is critical. Hence, we propose a redetection strategy, named assistant redetection, which conservatively trains the online support vector machine (SVM) classifier [11] and the short-term component to redetect the target. The proposed long-term tracker activates the short-term component again after the SVM classifier to redetect the target precisely while tracking failure occurs.

In the present study, we propose long-term complementary adaptive tracker (LCAT). The proposed tracker focuses on the research that merges different complementary trackers to overcome the drawbacks of the correlation filter and realise long-term

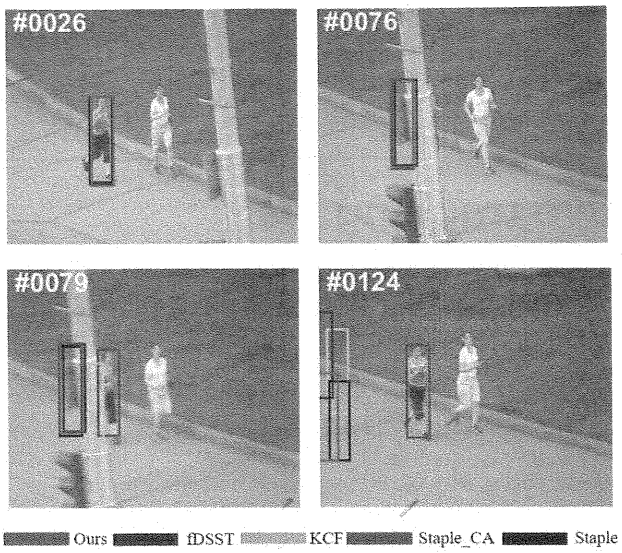

Fig. 1 *Effectiveness of assistant redetection in LCAT on sequence Jogging1. In comparison with some of the state-of-the-art trackers under the challenging circumstance of fast deformation, out-of-plane rotation, and heavy occlusion on Jogging1 sequence, our LCAT successfully redetect the target after a heavy occlusion, which leads to better robustness*

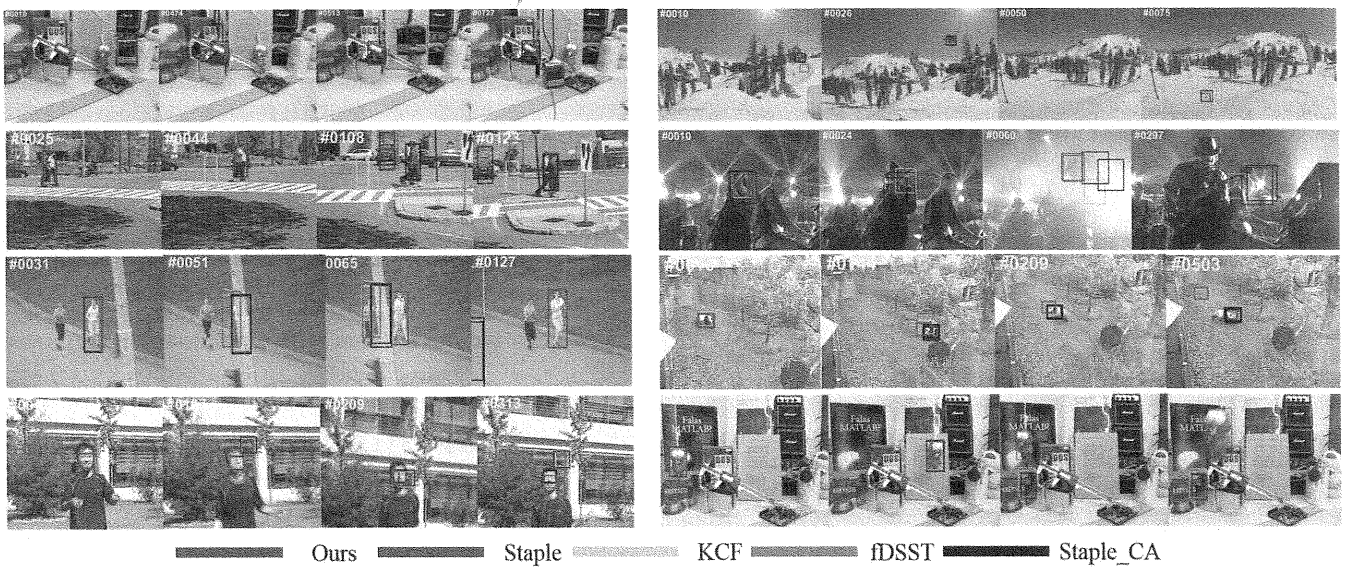

Fig. 2 *Qualitative comparison between LCAT and selected state-of-the-art trackers under challenging scenarios. From the tracking results of the Staple, fDSST, Staple_CA, KCF and LCAT, our proposed method demonstrates robustness under challenging scenarios*

tracking skilfully. This tracker not only establishes a remarkable result that exceeds the number of complex state-of-the-art trackers, including convolutional neural network (CNN)-based tracker and correlation trackers that employ CNN features, but also runs at 65 frames per second (FPS) on CPU.

Contributions. This study addresses the previously mentioned problem and proposes a novel long-term tracker by decomposing it into short-term, long-term, and scale-estimate components. The main contribution of our work can be summarised into four parts. First, we adopt complementary trackers to address the tracking problem and achieve a favourable performance. Second, we investigate the relationship between the response map of the tracker and the learning rate by proposing a sigmoid penalty coefficient to learn adaptively the tracking model. Third, we propose to employ past and current average peak-to-correlation energies (APCEs) [12] and maximum values from the response map to control the activity of the long-term component. Finally, we establish a simplified long-term component by combining the short-term component (complementary trackers) with the redetection SVM classifier to save computational cost and improve accuracy. Our long-term tracker effectively alleviates the model update problems, which often results in model drift, and robustly performs in various challenging video datasets.

2 Related works

2.1 Tracking-by-detection and correlation filters

Tracking-by-detection, as well as the development of machine learning in computer vision, became the most well-known and powerful tracking framework due to its high performance and efficiency. In comparison with generation models, tracking-by-detection usually employs a binary classifier to discriminate an object from its surrounding. Recently, several authoritative benchmark datasets [2, 3, 13–16] that contain a large number of challenging video sequences have been proposed, which significantly accelerate the development of the object tracking.

The following section introduces some recently proposed representative tracking-by-detection trackers. The Struck method [17] employs samples in training the SVM with structured labels to predict the target position, which achieved favourable results several years ago. Minimum output sum of squared error (MOSSE) [18] is the first tracker to employ a correlation filter on visual object tracking, which creates a correlation filter that can obtain the maximum value when it works on the target. The high-computational efficiency and performance of MOSSE have attracted considerable attention from the visual object tracking community. Henriques *et al.* proposed KCF [7] based on their prior

IET Image Process., 2019, Vol. 13 Iss. 9, pp. 1569-1577

work on circulant structure of tracking-by-detection with Kernels (CSK) [19], in which they employed a circulant matrix to obtain densely circular samples in the frequency domain and combined fast Fourier transform to train the ridge regression classifier with high efficiency. The KCF replaced the raw-pixel features of the CSK by the multichannel histogram of orientation gradients (HOG) [20] features and employed the kernel trick to enhance the discriminative ability further without significantly increasing the computational cost. The KCF achieved state-of-the-art in VOT14 challenge. Furthermore, discriminative scale space tracker (DSST) [21] and scale adaptive with multiple features tracker (SAMF) [22] were proposed in 2014 to solve the scale variation in video sequences. SAMF estimates the scale variation by training the KCF to search the target around the latest estimate position on various sizes of the image patch. The DSST trains a scale correlation filter and estimates the scales on the estimate position from the translation correlation filter. Although the KCF achieves excellent performance, it still has a number of critical problems due to circular samples that are weak approximations of the real samples. Thus, circular samples cannot truly represent the samples in real-world circumstances. The spatially regularised correlation filter (SRDCF) [4] was proposed to address this problem by employing spatial regularisation that penalises the filter coefficients. However, the SRDCF cannot be used in real time because it is difficult to optimise. The multi-expert entropy minimization (MEEM) [11] tracker combines multiple SVM classifiers with different adaptive learning rates and obtains tracking outputs according to a minimum entropy criterion. Bertinetto *et al.* [23] proposed Staple, which solves two independent ridge regression problems, to obtain a discriminative correlation filter and colour-based models rather than employing trackers of the same types. The two trackers are combined in a simple way but achieve a favourable result in the VOT16 challenge.

CNNs show outstanding performance in various fields, thereby attracting researchers' attention to extend the method to visual object tracking. Novel and powerful CNN features [24] have been introduced to correlation filter-based trackers [25–27], which significantly promotes the robustness and accuracy of the correlation filters. However, given that the complexity of CNN features is extremely high, it cannot run on the CPU at real time. Thus, CNN and SRDCF methods are inappropriate for real-time applications due to their high computational cost.

2.2 Long-term trackers

Long-term tracking is one of the problems to be solved in object tracking. In comparison with short-term trackers, long-term trackers are equipped with a redetection module to redetect the target, which helps stabilise the tracking result. The tracking-learning-detection (TLD) [28] is one of the most notable long-term trackers. As its name suggests, TLD consists of tracking, learning, and detection modules, which promote one another. The detector module will be activated while a tracking failure occurs to reinitialise the tracker. Correlation filters have been adopted to address long-term tracking problems in several recent works. Ma *et al.* [5] proposed a novel long-term correlation tracker (LCT), which achieves excellent performance by training a traditional translation correlation filter and an additional correlation filter to remember the excellent sample for confidence estimation that is employed to control the random forest classifier. Hong *et al.* [9] introduced cognitive psychology principles to visual tracking and proposed the MUlti-Store tracker (MUSTer). The framework of the MUSTer is based on the Atkinson–Shiffrin memory model. The MUSTer employs correlation filters for short-term tracking while employing scale-invariant feature transform keypoint and random sample consensus estimation for long-term tracking.

3 Our approach

In this section, the approach is divided into four parts. Section 3.1 presents the short-term component of our proposed tracker. Section 3.2 briefly describes the scale estimation approach used in our tracker. Section 3.3 presents the proposed adaptive learning rate model. Finally, Section 3.4 describes the proposed assistant redetection.

3.1 Short-term component

Similar to Staple, our short-term component is formulated by combining two complementary trackers that are sensitive to complementary elements. This complementary tracker generally works accurately and efficiently under relatively stable tracking scenarios.

3.1.1 Kernelised correlation filter (KCF): The KCF fully utilises the circulant matrix, utilises all the cyclic shift samples of the base sample, and converts the matrix multiplication to the Hadamard product based on Fourier transform. The circulant matrix transforms a computational cost of $\mathcal{O}(n^3)$ to nearly $\mathcal{O}(n\log n)$. In addition, the context information around the target is considered to enhance the discriminative ability of the KCFs.

Obtaining a KCF from an image patch x includes the target and the circular samples x_i of x by solving the ridge regression problem, is shown as follows:

$$\min_{w} \sum_{i} (f(x_i) - y_i)^2 + \lambda \| w \|^2. \tag{1}$$

The objective of solving the ridge regression problem is to obtain the optimal correlation filter w, such that $f(z) = w^{\mathrm{T}}z$ minimises the squared error over samples x_i with their soft regression label y_i. Traditionally, the classifier densely obtains samples around the target positions, assigning 0 to negative samples and 1 to positive samples. Different from the binary classifier, the soft regression label y is a Gaussian function. The value of the centred sample is 1, and other cyclic shifts vary from 0 to 1 based on the Euclidean distance to sample x; λ is a regularisation item that alleviates over-fitting. A cosine window is employed on the features extracted from the samples to avoid boundary clutter.

For the KCF, the solution w can be written as a linear combination of the training samples, as shown as follows:

$$w = \sum_{i} \alpha_i \varphi(x_i), \tag{2}$$

where $\varphi(x)$ denotes the mapping of a linear problem to a non-linear feature space by kernel trick and x denotes the feature extract from the samples.

The dual space coefficient α is defined as

$$\hat{\alpha} = \mathcal{F}(\alpha) = \frac{\mathcal{F}(y)}{\mathcal{F}(\varphi(x)\cdot\varphi(x)) + \lambda}, \tag{3}$$

where ^ denotes a corresponding symbol in the Fourier domain. In the new frame, we obtain a response for all the cyclic versions of the image patch z

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

where $\odot$ denotes the Hadamard product. The new position of the target can be found by locating the maximum value of the response map. k^{xz} is defined as the kernel correlation. In this study, we adopt the Gaussian kernel to enhance the performance of the ridge regression classifier. For the Gaussian kernel correlation, k^{xz} can be written as

$$k^{xz} = \exp\left(-\frac{1}{\sigma^2}(\| x \|^2 + \| z \|^2 - 2\mathcal{F}^{-1}(\hat{x}^* \odot \hat{z})\right) \tag{5}$$

The multichannel can be treated with a series of a single channel, such as $x = [x_1, \ldots, x_c]$, to handle a multichannel case. Given the linearity of the discrete Fourier transform, k^{xz} can be rewritten as the sum of the result of each channel

$$k^{xz} = \exp\left(-\frac{1}{\sigma^2}(\| x \|^2 + \| z \|^2 - 2\mathcal{F}^{-1}\left(\sum_C \hat{x}^* \odot \hat{z}\right)\right). \quad (6)$$

Equation (6) allows us to use a strong multichannel feature to improve the performance of the correlation filter-based trackers.

3.1.2 Colour-based tracker: In this section, we employ the Bayes rule [29] in constructing a discriminative colour-based object model to track the target efficiently. In the present work, we employ red–green–blue (RGB) colour histogram to obtain the pixel-wise scores on search region, which make the colour-based tracker strong at rapid deformation. To distinguish object O from the background on input image I, we employ the Bayes classifier at location x to obtain pixel-wise scores on search region

$$P(x \in O|F, B, c_x) \simeq \frac{P(c_x|x \in F)P(x \in F)}{\sum_{A \in \{F,B\}} P(c_x|x \in A)P(x \in A)}, \quad (7)$$

where F denotes a rectangular target region, B represents the surrounding region of the target, c_x is the pixel x that belongs to the cth bin of the RGB histogram, and $H_A^I(c)$ is the number of cth bin in the RGB histogram H of region $A \in I$. Pixel-wise scores can be simplified by estimating from the RGB histograms. Equation (7) can be rewritten as

$$P(x \in O|F, B, c_x) = \frac{H_F^I(c_x)}{H_F^I(c_x) + H_B^I(c_x)}. \quad (8)$$

In (8), pixel x belongs to region B. The response map of the colour-based tracker is obtained by efficiently applying the integral image on the pixel-wise scores on search region.

3.1.3 Complementary tracker: Owing to the HOG features and boundary effects of correlation filters, KCFs are weak at abrupt motion and fast deformation. The colour-based tracker is based on colour histogram and without boundary effects. The goal of constructing a complementary tracker involves fully utilising the advantages of a colour-based tracker to complement the disadvantages of KCFs and alleviate boundary effects to a certain extent.

The response map of the complementary tracker can be obtained by linearly combining the response map of the two complementary trackers.

$$\text{response} = (1 - \alpha)\text{response\_cf} + \alpha \cdot \text{response\_p}, \quad (9)$$

where α is the merge parameter between the response of the KCF response_cf and the colour-based tracker response_p. In this study, α is obtained through testing the OTB100 benchmark.

3.2 Fast scale space correlation filter

In the proposed long-term tracker LCAT, we employ a fast scale space correlation filter f_{scale} [6] to estimate the scale variation of the target. The fast scale space correlation filter employs the linear kernel to save computational cost.

Let $H \times W$ be the target size in the current frame. The image patch J_n of size $a^n H \times a^n W$ around the estimated position is extracted. a denotes the scale step between the two adjacent feature layers from the image patch. $S = 33$ is the size of the fast scale space correlation filter f_{scale} in the scale correlation filter

$$n \in \{[-(s-1)/2, \ldots, (s-1)/2]\}.$$

In this work, the fast scale space correlation filter reduces the feature dimensionality from ~1000 to 17 without losing the information by the standard principal component analysis method. Moreover, the sub-grid interpolation is employed on the scale correlation scores to interpolate the output scores from 17 to 33. The fast scale space correlation filter estimates and updates the target scale at the new target location from the proposed long-term tracker. The fast scale space correlation filter is linearly updated with a constant learning rate to adapt the scale changes of the target.

3.3 Adaptive learning rate

In most existing trackers, a constant learning rate is widely employed to update the tracking model at each frame. The model is updated to reduce the weight of the old samples in the tracking model. In the proposed tracker, we update the KCF and the colour histogram from background and foreground independently based on the tracking result in high learning rate. For the KCF, the updated formulation is listed as follows:

$$\tilde{\alpha}_t = (1 - \eta_{\text{cf}}) \cdot \tilde{\alpha}_{t-1} + \eta_{\text{cf}} \cdot \hat{\alpha}_t, \quad (10)$$

$$\tilde{x}_t = (1 - \eta_{\text{cf}}) \cdot \tilde{x}_{t-1} + \eta_{\text{cf}} \cdot \hat{x}_t. \quad (11)$$

In the colour-based tracker, the colour histogram updated online is as follows:

$$\tilde{b}_{\text{hist},t} = (1 - \eta_{\text{p}}) \cdot \tilde{b}_{\text{hist},t-1} + \eta_{\text{p}} \cdot b_{\text{hist},t}, \quad (12)$$

$$\tilde{f}_{\text{hist},t} = (1 - \eta_p) \cdot \tilde{f}_{\text{hist},t-1} + \eta_p \cdot f_{\text{hist},t}, \quad (13)$$

where η_{cf} and η_{p} are the learning rates of the correlation filter and colour-based tracker, respectively; f_{hist} is the RGB colour histogram of the target; b_{hist} is the RGB colour histogram of the search region; and $\sim$ denotes the corresponding symbol that is employed to detect the target position after the second frame.

However, the constant learning rate will update the tracking model with the same learning rate despite how terrible the tracking result is, which may result in tracking failure once the target is detected inaccurately. The response map is the feedback of the tracker, which is the accurate method to evaluate the quality of the tracking result. The main problem in the model update is the indistinct relationship between the sample quality and the learning rate. Recently, a wide variety of sigmoid functions have been broadly employed as activated functions in neural networks. The sigmoid function is often employed as the learning curve of a complex system when the specific mathematical model is lacking [30]. Thus, we propose a sigmoid penalty coefficient to connect the maximum value of the response map and the learning rate. For the KCF, the learning rate can be written as

$$\eta_{\text{cf}} = \frac{1.8}{1 + \exp(5(\text{Mean\_cf} - \text{Max\_cf}))} \cdot b, \quad (14)$$

where Mean_cf represents the previous mean maximum value of the response map in the KCF and Max_cf is the maximum value of the response map in the current frame. Similar to KCF, the learning rate of the colour histogram can be rewritten as

$$\eta_{\text{p}} = \frac{1.8}{1 + \exp(5(\text{Mean\_p} - \text{Max\_p})} \cdot c, \quad (15)$$

where Mean_p represents the previous mean maximum value of the colour response map; Max_p is the maximum value of the colour response map in the current frame; and b, c is the constant learning rate of the KCF and colour-based tracker, respectively.

In the model update stage, the KCF and colour-based tracker update the tracking model with the constant learning rate in the first dozen of frames to obtain the stable previous mean maximum value of the response map that is employed to construct the penalty coefficient. This approach avoids special circumstances that result in the previous mean maximum value of the response map producing drastic fluctuation at the beginning.

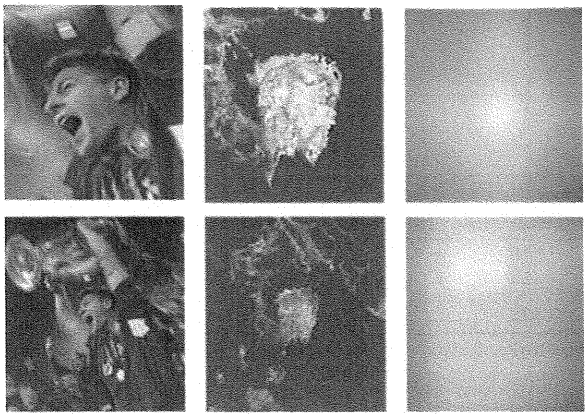

Fig. 3 *Search region, corresponding per-pixel scores, and histogram response on soccer sequences. A large search region significantly reduces the accuracy of the colour-based tracker*

Input: Initial target bounding box
Output: Estimated target bounding box
Repeat
1: Crop an image patch z centered at the last location and extract features.
2: Crop an image patch p centered at the last location and compute color histogram.
// Translation estimation
3: Compute kernelized correlation response and color-based response, merge trackers.
4: Estimate target position and compute $APCE$ and $Max\_response$.
// Assistant re-detection
5: **if** ($APCE<\beta\cdot Mean\_APCE$) or ($Max\_response<\beta\cdot Mean\_response$)
6: Activate redetection classifier and find the position pos_SVM.
7: Repeat 1–4 at pos_SVM, obtain $ACPE\_r$ and $Max\_response\_r$.
8: **if** ($APCE\_r>\gamma\cdot Mean\_APCE$) and($Max\_response\_r>\gamma\cdot Mean\_response$)
9: Accept the result of assistant redetection.
10: **end**
11: **end**
// Scale estimation
12: Crop different sizes of image patch and construct fast scale space correlation filter.
13: Compute current scale.
// update model
14: **if** ($APCE>\theta\cdot Mean\_APCE$) and ($Max\_response>\theta\cdot Mean\_response$)
15: Update translation filter and color histogram.
16: **if** ($APCE>\lambda\cdot Mean\_APCE$) and ($Max\_response>\lambda\cdot Mean\_response$)
17: Train SVM classifier.
18: **end**
19: **end**
20: Update fast scale space correlation filter.
21: **until** end of video sequence

Fig. 4 *Algorithm 1: Brief outline of LCAT*

3.4 Assistant redetection

A practical robust tracker should be equipped with a redetection module to recover the target after the tracking failure. Different from previous long-term trackers [5, 9, 10, 31, 32], we propose a novel redetection strategy to realise the long-term tracking with high accuracy and high speed. In our approach, we fully utilise our short-term component to save computational cost and improve accuracy. The SVM classifier used in our tracker acts as an auxiliary of the short-term component. In this way, we transform the long-term tracking problem into a novel problem, which extends the search region of the short-term component.

The inherent boundary effects of the correlation filter-based trackers constrain the search region in a fixed size, which weakens the correlation filter-based tracker for fast motion and occlusion. Particularly, our short-term component is the combination of the KCF and the colour-based tracker, which is limited by the simplicity of the colour-based tracker, thereby reducing the search region further. Fig. 3 shows a large search region that result in a significant decrease in the accuracy of the colour-based tracker. We alleviate this problem by proposing an assistant redetection to extend the search region after the tracking failure. The assistant redetection reuses the short-term component work on the discriminative result of the SVM classifier while satisfying the activated criterion. The redetection of the SVM classifier can be considered a method that extends the search region and weakens the boundary effects to some extent. The result of the short-term component is employed to compute the confidence scores to decide whether to activate the redetection module, adopt the assistant redetection result, update the tracking model, or train the classifier.

In this section, we introduce a novel criterion to distinguish the sample quality, which allows the tracker to adjust the threshold related to redetection adaptively. The APCE [12] criterion is expressed as follows:

$$\mathrm{APCE} = (R_{\max} - R_{\min})^2 / \mathrm{Mean}\left(\sum (R - R_{\min})^2\right). \quad (16)$$

The ideal tracking result is the accuracy while the response map is unimodal. The APCE shows the undulated degree of the response map, which indicates the confidence of the current tracking result.

In the present work, we employ two criteria, namely, APCE and maximum value of response map $R_{\max}$ to distinguish the confidence of the current tracking result. A pervasive criterion for long-term tracking is proposed on the basis of the APCE and $R_{\max}$. The redetection module will be activated despite either the APCE or $R_{\max}$ in the current frame is less than their previous mean values with a specific ratio β. The ratio β is the threshold that distinguishes the reliability of the tracking result of the short-term component. The result of the assistant redetection may also result in tracking failure. To ensure that the result of the assistant re-detection is sufficiently accurate, the result of the assistant redetection will be accepted only if the APCE and $R_{\max}$ of the assistant redetection are higher than the previous mean values with a specific ratio γ. We train the SVM classifier while the APCE and $R_{\max}$ are higher than their previous mean values with a certain high ratio θ to guarantee that the SVM classifier is trained with the correct samples. In case the tracking models are updated with inaccurate detection, the tracking model will be rapidly corrupted. Hence, we employ a high-confidence update strategy to eliminate the corrupted samples that contain a few useful messages. In other words, if the APCE and $R_{\max}$ are less than their previous mean values with a specific ratio λ, then the tracking model will not update and detect the target in the next frame. In this way, we prevent the tracking model from experiencing a heavy model drift to strengthen the discriminative ability of the short-term component.

Algorithm 1 summarises the proposed long-term trackers (Fig. 4).

4 Implementation

This section further describes the proposed methods considering the feature and kernel trick used in the proposed tracker and the classifier employed in the redetection module. The parameters in the proposed LCAT are also presented.

4.1 Feature

We employ the first 27 channels of the fast histogram of oriented gradient (fHOG) [33] as features and further augment the features with raw greyscale pixel values. The cell size of the fHOG feature is set to 4. In addition, the same feature is extracted from the pixel-wise scores map of the search region as a supplement to enhance the feature in the KCF. In summary, the features employed in the translation correlation filter are 56 channels. The features employed in the fast scale space correlation filter are the first 31 channels of the fHOG features. The feature used in the colour-based tracker is the RGB colour histogram. The number of bins for the RGB histogram is set to 32.

4.2 Kernel trick

In the present work, the Gaussian kernel $\left[k(x,x') = \exp\left(-|x-x'|^2/\sigma^2\right)\right]$ is selected to enhance the correlation filter, where $\sigma = 0.5$.

4.3 Redetection classifier

The redetection classifier employed in the present work is the online soft margin SVM. The training samples of the SVM

Table 1 Threshold of assistance re-detection

	β	γ	θ	λ
value	0.4	0.6	0.8	0.4

Table 2 Tracking results of different version of LCAT

Tracker	Learning rate	Assistant re-detection	OTB2013		OTB2015		FPS
			Precision	Success	Precision	Success	
baseline	constant	No	0.819	0.604	0.811	0.600	*77.9*
LCAT_SN	sigmoid	No	**0.861**	**0.640**	**0.844**	**0.628**	77.1
LCAT	sigmoid	Yes	*0.895*	*0.661*	*0.875*	*0.646*	65.9

Tracking result of baseline, LCAT_SN, and LCAT on OTB2013 and OTB100. The italic values denote the best in three versions of trackers, whereas the bold values denote the second best. The mean FPS is obtained by estimation on OTB2013.

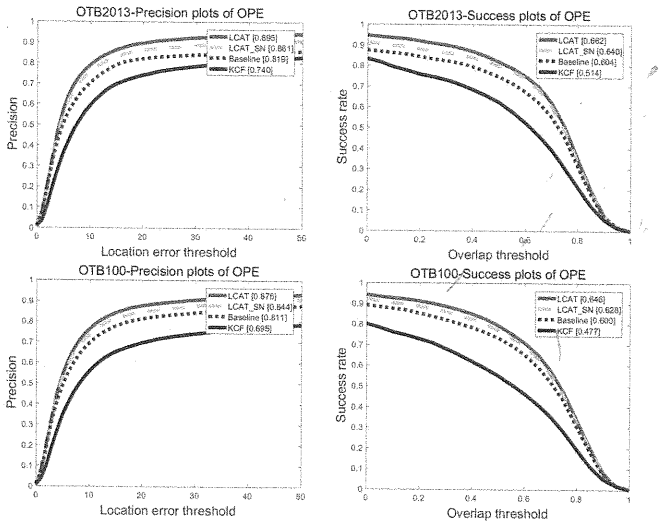

Fig. 5 *Precision and success plots on OTB2013 and OTB100; The order of the trackers in the plots are ranked by scores. The title of the plot includes the corresponding benchmark and evaluation method. The result demonstrates our methods are useful in both benchmarks*

classifier are considered positive when the overlap between the target and the samples is >0.9, whereas the samples are considered negative when their overlap is <0.5. The feature of the SVM classifier is LAB colour histogram.

4.4 Set-up

The merge factor α of the two trackers is set to 0.25. The regularisation item of the KCF is set to 1^{-4}. The scale step a is set to 1.02. In the model update stage, $b = 0.02$ and $c = 0.04$. The target is resized to the standard area of 150×150 pixels to save computational cost when the target is extremely large. Table 1 lists the previously mentioned threshold in the redetection module of the LCAT. The various thresholds are used to control the long-term tracking component.

5 Experimental results

Here, we evaluate the proposed method on the five challenging classic visual tracking benchmarks, namely, OTB2013 [2], OTB100 [3], TC128 [13], UVA123 [14], and UVA20L [14]. The evaluation of the trackers employs two criteria, namely, (i) success scores, which indicate the area under each success plots of trackers; and (ii) precision scores, which represent the percentage of the successfully tracked frames whose distance between the position of tracking result and the position of the annotation is <20 pixels. The success scores are more accurate than the precision scores. Furthermore, the three types of methods used to evaluate the trackers are as follows. The one-pass evaluation (OPE) is the traditional method in evaluating the tracker throughout a video sequence with correct initialisation in the first frame, and the temporal robustness evaluation (TRE) and the spatial robustness evaluation (SRE) are employed to analyse the robustness of the trackers with various initialisations.

We initially analyse the LCAT with the adaptive learning rate and assistant redetection on OTB2013 and OTB100. OTB2013 and OTB100 are common tracking benchmarks. OTB2013 contains 50 challenging video sequences, whereas OTB100 extends OTB2013 and contains 100 videos. Then, we compare the proposed LCAT with some state-of-the-art trackers on OTB2013, OTB100, TC128, UAV20L, and UVA123 benchmarks. All the results are compared under the same conditions. The parameters in the LCAT are the same in the following test.

Our experiment is conducted on MATLAB R2016a with I7-8700 3.20 GHz CPU with 16 GB RAM.

5.1 Component analysis

Various versions of the proposed tracker are tested on OTB2013 and OTB100 to validate the effect of the proposed methods. Particularly, we constrain the dimension of the RGB colour histogram to one for some greyscale video sequences on OTB2013 and OTB100. This approach causes the proposed LCAT to reach a low performance than the other benchmark that consists entirely of colourful video sequences.

The short-term component of the proposed tracker is denoted as baseline. The only employed sigmoid penalty coefficient is denoted as LCAT_SN. The LCAT employs sigmoid penalty coefficient and assistant redetection. Table 2 shows the tracking results and their structure. Fig. 5 shows the precision and success plots.

As shown in Fig. 5, the proposed LCAT substantially outperforms KCF with an average relative improvement of 32.1% on the success plots of OTB2013 and OTB100. All the proposed methods significantly improved the tracking result according to the experimental results. Table 2 indicates that among the presented trackers, the proposed LCAT shows the best accuracy in the OPE of OTB2013 and OTB100. The proposed LCAT significantly outperforms the short-term component with the aid of the sigmoid penalty coefficient and assistant redetection. The LCAT achieves a relative improvement of 9.4% in the success plot in comparison with the baseline on OTB2013 and achieves a relative improvement of 7.6% on OTB100. In addition, the lack of a redetection module results in the poorer performance of the LCAT_SN compared with the LCAT. The result of the running speed of the long-term trackers only slightly decreases with the significant increase of the performance due to the assistant redetection employed with the high-confidence update strategy and a small number of samples to train the SVM classifier.

As shown in Fig. 6, the sigmoid penalty coefficient adaptively adjusts the learning rate while the target encounters an appearance variation. The learning rate in Faceocc1 video sequences increases, whereas the target is distinct and decreases as it encounters an occlusion. Although the learning rate will be adaptively adjusted by sample quality, the corrupted sample will be learned in a relatively low learning rate, which may result in model drift. Hence, we adopt the high-confidence update strategy to eliminate corrupted samples and obtain samples with useful information.

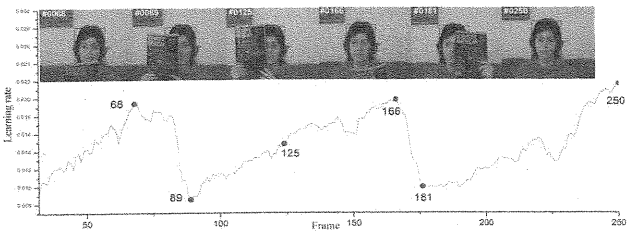

Fig. 6 *Illustration of the employed sigmoid penalty coefficient that adjusts the learning rate in Faceocc1 sequences on LCAT_SN, the learning rate of KCF, and the corresponding training samples in some frames*

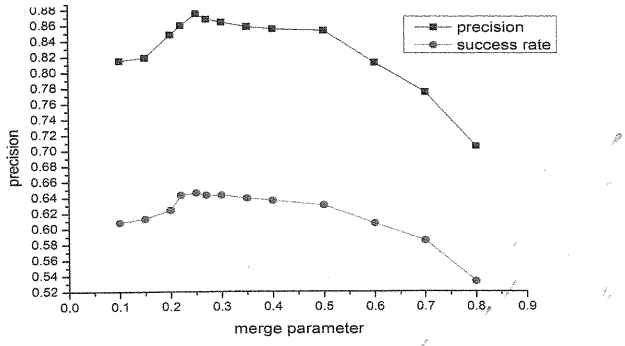

Fig. 7 *Merge parameter in relation to the success rate and precision on OTB100. The black and red points are obtained experimentally*

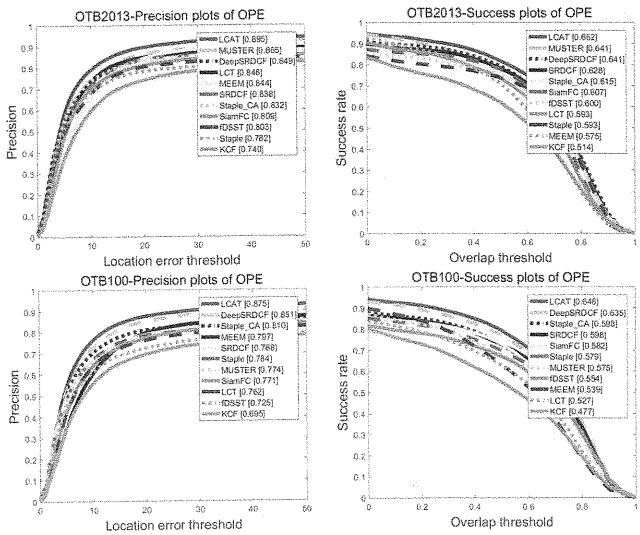

Fig. 8 *Success and precision plots of OPE on OTB2013 and OTB100. Our method outperforms the second-best tracker (i.e. DeepSRDCF) with 1.7% in the success plot of OTB100*

5.2 Merge parameter experiments

In (7), the response map of the two complementary trackers is merged by parameter α. This section explores the influence of the merge parameter in the LCAT on the OTB100 benchmark. Fig. 7 shows that the merge parameter α significantly influences the performance of the LCAT. The best performance of the LCAT is achieved at $\alpha = 0.25$. The merge parameter also indicates that the result of the KCF is more reliable than the colour-based tracker.

5.3 State-of-the-art comparison

5.3.1 One-pass evaluation (OPE): In this section, we evaluate the LCAT with other ten most related and state-of-the-art trackers, namely, MEEM, fast discriminative scale space tracking (fDSST), KCF, Staple, Staple_CA, SRDCF, LCT, DeepSRDCF, SiamFC, and MUSTer on OTB2013 and OTB100. The LCAT is proven to have a state-of-the-art performance.

Among the trackers, KCF, fDSST, SRDCF, DSST, and DeepSRDCF are the correlation filter-based trackers. Staple, Staple_CA, and MEEM are the trackers combined with multiple trackers. LCT and MUSTer are the long-term trackers based on KCFs. LCAT is the long-term tracker based on complementary trackers. SiamFC is a CNN-based tracker. DeepSRDCF is a correlation filter-based tracker that employs features extracted from CNNs.

Fig. 8 illustrates the precision and success plots of the 11 trackers. The result shows that the OTB100 is more challenging because almost all the trackers obtain a lower performance compared with OTB2013. The proposed LCAT performs significantly better than all the other state-of-the-art trackers, including the CNN-based and correlation filter-based trackers, with a relative average improvement of 11.6% with respect to the Staple tracker on the success plots of OTB2013 and OTB100. The proposed LCAT also obtains average progress on the SRDCF with a relative improvement of 6.7%. In comparison with the other long-term trackers, the proposed LCAT obtains an average improvement of 10.1% relative to the LCT and 8.3% relative to the MUSTer on the success plots of OTB2013 and OTB100. Moreover, in all the related trackers previously mentioned, the trackers that can run at real time on the CPU are Staple (107.6 FPS), fDSST (136.9 FPS), Staple_CA (62.3 FPS), KCF (232.8 FPS), and LCT (35.2 FPS). The proposed LCAT (65.9 FPS) not only runs at a significantly higher frame rate than LCT but also approximately close to the Staple_CA.

5.3.2 Attribute-based evaluation: The video sequences in the OTB2013 and OTB100 are annotated with 11 challenging attributes in the tracking problem. These challenging attributes are convenient for the evaluation of the tracker's performance under various challenging aspects. For detailed analyses, the proposed LCAT is also evaluated with other state-of-the-art trackers on 11 challenging attributes on OTB100. However, only the results of the eight main challenging attributes are reported, as shown in Fig. 9. From the figure, the result shows that the proposed LCAT significantly outperforms the second-best tracker relative to the deformation (8.6%), illumination variation (7.3%), occlusion (5.2%), and out-of-camera view (4.7%). The LCAT is robust to the deformation and illumination variation. The correlation filter-based trackers struggle in the cases of occlusion, deformation, out-of-camera view, fast motion due to the boundary effects and relatively risky update strategy. The proposed methods significantly enhanced the performance in comparison with the correlation-based trackers in these attributes. On the success plots of scale variation, although our LCAT is the second-best tracker, LCAT significantly outperforms the third-best tracker (i.e. SRDCF) with 6.6% and close to the best tracker (i.e. DeepSRDCF).

5.3.3 Robustness to different benchmarks: We evaluate the proposed LCAT with MEEM, fDSST, Staple, Staple_CA, SRDCF, LCT, DSST, KCF, and MUSTer on the TC128, UAV123, and UAV20L benchmarks to further validate the effect of the proposed methods. TC128 is a dataset with 128 colour sequences, which show the benefits of colour information for tracking. The proposed LCAT is encoded by colour information to a great extent, which may achieve a substantial improvement on the TC128. The UAV123 dataset is an aerial video dataset for the target tracking of unmanned aerial vehicles (UAVs). This dataset contains 123 fully annotated aerial videos with >110K frames from the low-altitude aerial perspective. The applications of the UAVs include crowd surveillance, obstacle avoidance, and localisation, which need to be addressed in real time. In the present experiment, the down-sampled version of UAV123 is employed to evaluate the proposed tracker. The down-sampled version is down-sampled to ten FPS, which increases the displacement between the two adjacent frames and results in additional challenges. Particularly, the correlation filter-based trackers that are limited by the boundary effects will further achieve a limited performance. UAV20L dataset is a long-term tracking benchmark derived from UAV123 dataset. Fig. 10 shows the result of the proposed tracker and the state-of-the-art trackers in TC128, UAV123, and UAV20L.

On the success plot of TC128, the top three trackers that are composed of a colour-based tracker fully demonstrate the effect of

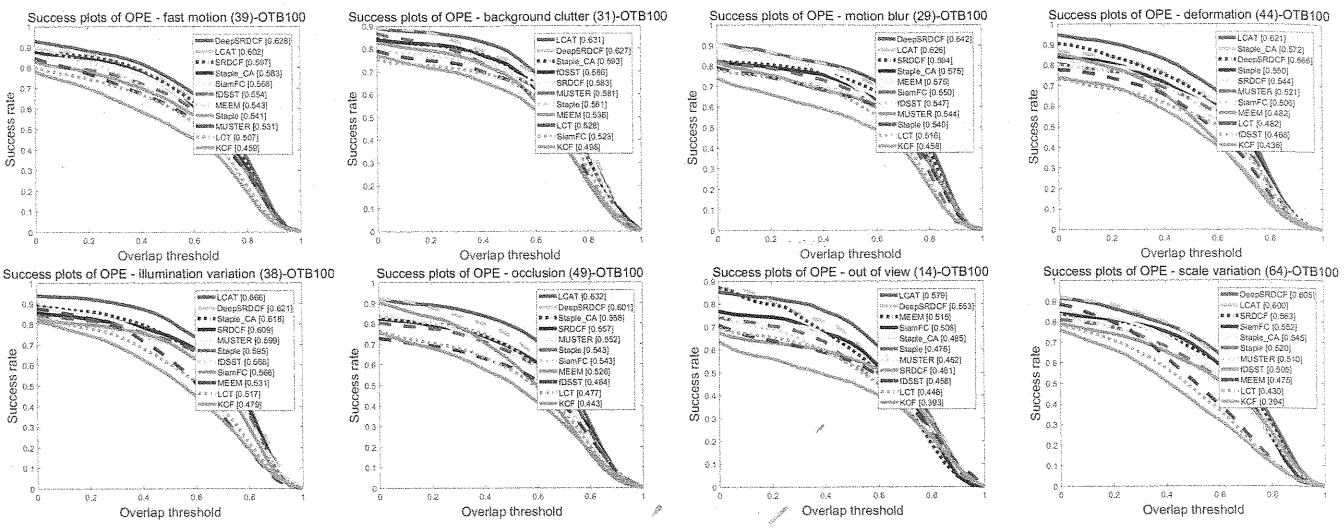

Fig. 9 *Success plots showing the performance of the 8 challenging tracking attributes (i.e. fast motion, background clutter, motion blur, deformation, illumination variation, occlusion, out-of-camera view, and scale variation) on OTB100 that contains 100 videos*

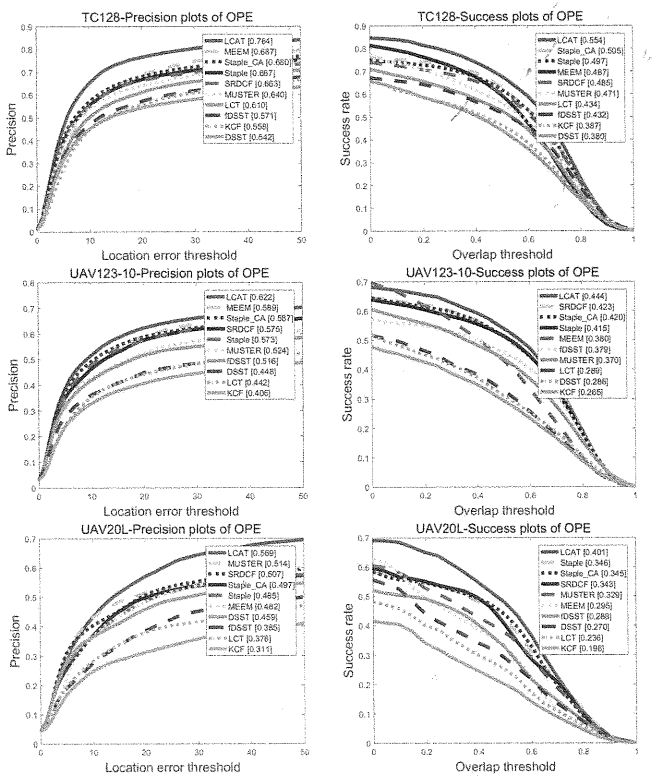

Fig. 10 *Success and precision plots of OPE on TC128, UAV123, and UAV20L. The LCAT shows state-of-the art performance compared with the related trackers in these challenging benchmarks*

colour information for tracking. Among the compared trackers, with the aid of the assistant redetection, the proposed LCAT is the best tracker in UAV123 datasets. SRDCF is the second-best tracker on the success plot of UAV123 due to its handle boundary effects by spatial regularisation. MEEM is also the second-best tracker based on the precision plots of TC128 and UAV123; however, it is inferior to the success plot because it cannot address scale variation. On the down-sampled version of UAV123, which is limited by the boundary effects, the standard correlation filter-based trackers, namely, LCT, DSST, and KCF, show a significant inferior performance than the other trackers. The LCAT reaches a relative substantial improvement of 43.1% with respect to the KCF and 27.6% with respect to the LCT on the success plot of TC128. The LCAT achieves a relative substantial improvement of 67.5 and 53.6% on UAV123. On the UAV20L dataset, our LCAT is the best tracker and far exceeds the rest of trackers. On the success plots of UAV20L, LCAT obtains progress on the second-best tracker with a relative improvement of 15.9%. In comparison with the other long-term trackers, LCAT obtains an average improvement of 16.3% relative to the MUSTer and 60.2% relative to the LCT on the success and precision plots of UAV20L. The results on UAV20L also show that the proposed redetection strategy displays more favourable robustness on long-term tracking than existing redetection strategy.

5.3.4 Initialisation robustness evaluation: Object tracking is greatly influenced by initialisation. The SRE and TRE are employed in analysing the robustness of each tracker on OTB100 to evaluate the robustness of the proposed method. The result on the SRE represents the sensitivity of the tracker at the start of the noisy initialisation. The TRE evaluates the sensitivity of the tracker at the start with various frames on the same sequences. Fig. 11 shows that the LCAT is the best tracker in SRE and TRE, which indicates that the proposed method is robust to different spatial and temporal initialisations.

Although the LCAT is the best tracker in SRE and TRE, the results of the proposed method show that the LCAT relatively weakens at spatial noisy and different temporal initialisations, which is similar to other long-term trackers. The TRE evaluates trackers numerous times from various starting frames in the video sequence, which results in the advantage of redetection badly weakened and exhibit a relative inferior performance in the TRE. The SRE evaluates the trackers under different shifting or scaling of target initialisation, inaccurate initialisation of the target result in the proposed short-term component, and the SVM classifier trained by inaccurate samples. The error of the short-term component and the SVM classifier will be accumulated during tracking, which leads to a weak LCAT and a relatively low SRE. Therefore, the merits of the long-term trackers cannot be fully reflected by the SRE and TRE.

6 Conclusion

In this study, we propose the novel LCAT to address the long-term visual object tracking problem. The baseline method is a combination of the KCF and the colour-based tracker, which alleviate the boundary effects of the correlation filter-based trackers and achieve a real-time excellent performance. The feedback with the learning rate of the tracker by the sigmoid penalty coefficient is examined to prevent the tracking model from drifting. Then, a high-confidence update strategy is employed to avoid updating corrupted samples. Moreover, we establish a simplified long-term tracker by employing assistant redetection, which combines the short-term components with the SVM classifier in constructing a long-term component that handles tracking failure to effectively realise the long-term tracking. The extensive experiment results of the four benchmarks show that the

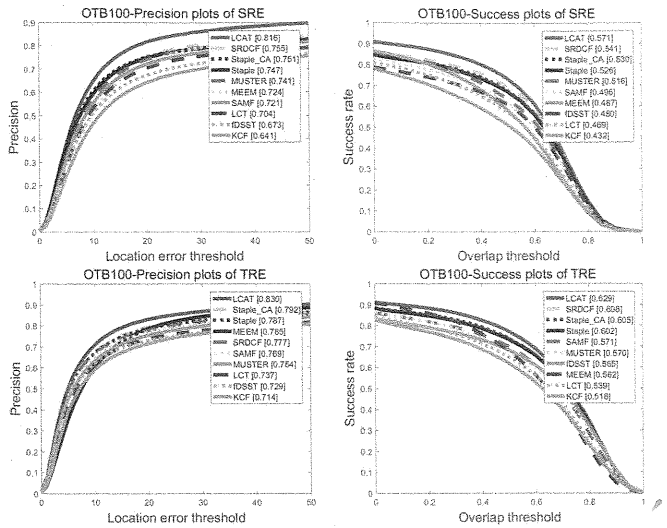

Fig. 11 *Success and precision plots of SRE and TRE and OTB100 benchmarks. These plots show the comparisons of temporal and spatial robustness between the related state-of-the-art trackers and the proposed approach. In both cases, LCAT demonstrates superior performance*

proposed LCAT significantly outperforms the related state-of-the-art trackers relative to its efficiency, accuracy, and robustness when running at a fairly high frame rate. The results show that the proposed LCAT is a preferable choice for real-word applications that must work in real time with high accuracy.

7 References

[1] Yilmaz, A.: 'Object tracking: a survey', *ACM Comput. Surv.*, 2006, **38**, (4), p. 13

[2] Wu, Y., Lim, J., Yang, M.H.: 'Online object tracking: a benchmark'. IEEE Conf. on Computer Vision and Pattern Recognition IEEE Computer Society, Portland, OR, USA, 2013, pp. 2411–2418

[3] Wu, Y., Lim, J., Yang, M.H.: 'Object tracking benchmark', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (9), p. 1834

[4] Danelljan, M., Hager, G., Khan, F.S., *et al.*: 'Learning spatially regularized correlation filters for visual tracking', Santiago, Chile, 2016, pp. 4310–4318

[5] Ma, C., Yang, X., Zhang, C., *et al.*: 'Long-term correlation tracking'. Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 5388–5396

[6] Danelljan, M., Hager, G., Khan, F.S., *et al.*: 'Discriminative scale space tracking', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **39**, (8), pp. 1561–1575

[7] Henriques, J.F., Rui, C., Martins, P., *et al.*: 'High-speed tracking with kernelized correlation filters', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (3), pp. 583–596

[8] Matthews, I., Ishikawa, T., Baker, S.: 'The template update problem', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004, **26**, (6), pp. 810–815

[9] Hong, Z., Chen, Z., Wang, C., *et al.*: 'MUlti-Store tracker (MUSTer): a cognitive psychology inspired approach to object tracking'. Computer Vision and Pattern Recognition IEEE, Boston, MA, USA, 2015, pp. 749–758

[10] Zhu, G., Wang, J., Wu, Y., *et al.*: 'Collaborative correlation tracking'. British Machine Vision Conf., Swansea, UK, 2015, pp. 184.1–184.12

[11] Zhang, J., Ma, S., Sclaroff, S.: 'MEEM: robust tracking via multiple experts using entropy minimization', Zurich, Switzerland, 2014, Vol. 8694, pp. 188–203

[12] Wang, M., Liu, Y., Huang, Z.: 'Large margin object tracking with circulant feature maps'. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, 2017, pp. 4800–4808

[13] Liang, P., Blasch, E., Ling, H.: 'Encoding color information for visual tracking: algorithms and benchmark', *IEEE Trans. Image Process. A Publ. IEEE Signal Process. Soc.*, 2015, **24**, (12), p. 5630

[14] Mueller, M., Smith, N., Ghanem, B.: 'A benchmark and simulator for UAV tracking'. European Conf. on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 445–461

[15] Smeulders, A.W.M., Chu, D.M., Cucchiara, R., *et al.*: 'Visual tracking: an experimental survey', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **36**, (7), pp. 1442–1468

[16] Zajc, L.C., Lukezic, A., Leonardis, A., *et al.*: 'Beyond standard benchmarks: parameterizing performance evaluation in visual object tracking'. IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 3343–3351

[17] Hare, S., Saffari, A., Torr, P.H.S.: 'Struck: structured output tracking with kernels'. IEEE Int. Conf. on Computer Vision IEEE, Ontario, Canada, 2012, pp. 263–270

[18] Bolme, D.S., Beveridge, J.R., Draper, B.A., *et al.*: 'Visual object tracking using adaptive correlation filters'. Computer Vision and Pattern Recognition IEEE, San Francisco, CA, USA, 2010, pp. 2544–2550

[19] Henriques, J.F., Rui, C., Martins, P., *et al.*: 'Exploiting the circulant structure of tracking-by-detection with kernels', *Lect. Notes Comput. Sci.*, 2012, **7575**, (1), pp. 702–715

[20] Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition CVPR 2005 IEEE, San Diego, CA, USA, 2005, pp. 886–893

[21] Danelljan, M., Häger, G., Khan, F.S., *et al.*: 'Accurate scale estimation for robust visual tracking'. British Machine Vision Conf., Nottingham, UK, 2014, pp. 65.1–65.11

[22] Li, Y., Zhu, J.: 'A scale adaptive Kernel correlation filter tracker with feature integration'. European Conf. on Computer Vision, Cham, 2014, pp. 254–265

[23] Bertinetto, L., Valmadre, J., Golodetz, S., *et al.*: 'Staple: complementary learners for real-time tracking'. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 1401–1409

[24] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition', *Comput. Sci.*, 2014, arXiv:1409.1556

[25] Ma, C., Huang, J.B., Yang, X., *et al.*: 'Hierarchical convolutional features for visual tracking'. IEEE Int. Conf. on Computer Vision. IEEE Computer Society, Santiago, Chile, 2015, pp. 3074–3082

[26] Danelljan, M., Hager, G., Khan, F.S., *et al.*: 'Convolutional features for correlation filter based visual tracking'. IEEE Int. Conf. on Computer Vision Workshop IEEE Computer Society, Santiago, Chile, 2015, pp. 621–629

[27] Danelljan, M., Robinson, A., Khan, F.S., *et al.*: 'Beyond correlation filters: learning continuous convolution operators for visual tracking'. European Conf. on Computer Vision, Cham, 2016, pp. 472–488

[28] Kalal, Z., Mikolajczyk, K., Matas, J.: 'Tracking-learning-detection', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (7), pp. 1409–1422

[29] Possegger, H., Mauthner, T., Bischof, H.: 'In defense of color-based model-free tracking'. Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 2113–2120

[30] Gibbs, M.N., Mackay, D.J.C.: 'Variational Gaussian process classifiers', *IEEE Trans. Neural Netw.*, 2002, **11**, (6), pp. 1458–1464

[31] Facehugger, F. P.: 'The ALIEN tracker applied to faces'. Proc. of the European Conf. on Computer Vision, Florence, Italy, 2012

[32] Supancic, J.S., Ramanan, D.: 'Self-paced learning for long-term tracking'. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013

[33] Felzenszwalb, P., Girshick, R., McAllester, D., *et al.*: 'Object detection with discriminatively trained part-based models', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (9), pp. 1627–1645