Pattern Recognition 87 (2019) 216–225

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Multi-level context-adaptive correlation tracking

Peng Liu, Chang Liu, Wei Zhao*, Xianglong Tang

Department of Computer Science, Harbin institute of Technology, Harbin, China

ARTICLE INFO

Article history:
Received 13 March 2018
Accepted 9 October 2018
Available online 11 October 2018

MSC:
00-01
99-00

Keywords:
Visual tracking
Correlation filter
Context-adaptive tracker
Context pyramid

ABSTRACT

The discriminative correlation filter (DCF) has shown impressive performance in visual tracking. Context has two functions in DCF: addressing the disturbance in target locating, and supplying cues for locating the target within the context. To improve the context utilization, we introduce a multi-level context-adaptive tracking (MCAT) approach for DCF tracking. Firstly, a multi-level context representation—called a context pyramid—is proposed to exploit the relationship between the target and its context for better visual tracking. Secondly, for each level of the context pyramid, we control the effect of context in DCF learning and tracking using context-adaptive spatial windows. An accurate target model can thereby be learned, even when the background clutter is severe. Moreover, the target can be more easily tracked when the background is weakened by the spatial window. Thirdly, a robust prediction of the target position is obtained with the multi-level structure of the context pyramid. Experimental results showed that, with conventional hand-crafted features, our tracker provided state-of-the-art performance on OTB100 comparable to those of deep-learning-based trackers.

1. Introduction

Visual tracking plays a significant role in computer vision and can be widely applied in surveillance, navigation, and human-computer interaction. The most widely considered scenario is model-free single-object tracking, wherein the initial position and size of a target are given and should be predicted in each video frame. Although large datasets [1–4] and many outstanding researchers have enabled significant advancements, it remains challenging to successfully track any target in various video sequences. This is because several factors should be simultaneously considered, including target and background variations, such as deformation, pose variation, fast motion, background clutter, and occlusion.

Discriminative methods [8–10] have recently shown good performances. In these methods, visual tracking is formulated as a binary classification or regression problem to distinguish the target from the background. Tracking based on a correlation filter [11–13] is the most renowned approach because of the high precision and high speed. An online correlation filter is learned from the region of interest in the current frame and applied in the next frame to predict the target location with the maximum response.

Correlation filters can be efficiently implemented with the Fast Fourier transform (FFT); nevertheless, the bound effect [14,15] is unavoidable because the samples are regarded as periodic ones. The original correlation filters use a spatial window to restrain the features in the window boundary. The spatial window strategy can reduce the bound effect while also reducing the field of interest in the window. This is because the target located closer to the boundary will be more difficult to track. Another limitation is that the samples should have the same size in the regions of interest in learning and tracking. A rectangle window is typically utilized to represent the region of interest. However, a contradiction occurs. A tracker with a small window size will suffer from fast motion. Once the target moves beyond the window, the tracker fails. Furthermore, a large window size will suffer from background clutter. The learned filters are more influenced by the background and more distractors will be included by the window. However, existing trackers based correlation filters typically employ a fixed window size and fixed spatial window. It is difficult to achieve a good compromise between fast motions and background clutter.

In this paper, we introduce a context pyramid representation, which is shown in Fig. 1(a). With this representation, we combine the target with different context levels and jointly learn correlation filters for them. In other words, we utilize a three-dimensional (3D) window to train the 3D correlation filters. As mentioned above, the spatial window can influence the field of interest. Accordingly, we utilize a fixed window size for simplicity and apply

* Corresponding author at: Harbin institute of Technology, Department of Computer Science, Rm. 914 Xin Jishu Building, Harbin 150001, China.
E-mail addresses: pengliu@hit.edu.cn (P. Liu), magicallc@126.com (C. Liu), zhaowei@hit.edu.cn (W. Zhao), tangxl@hit.edu.cn (X. Tang).

https://doi.org/10.1016/j.patcog.2018.10.013
0031-3203/

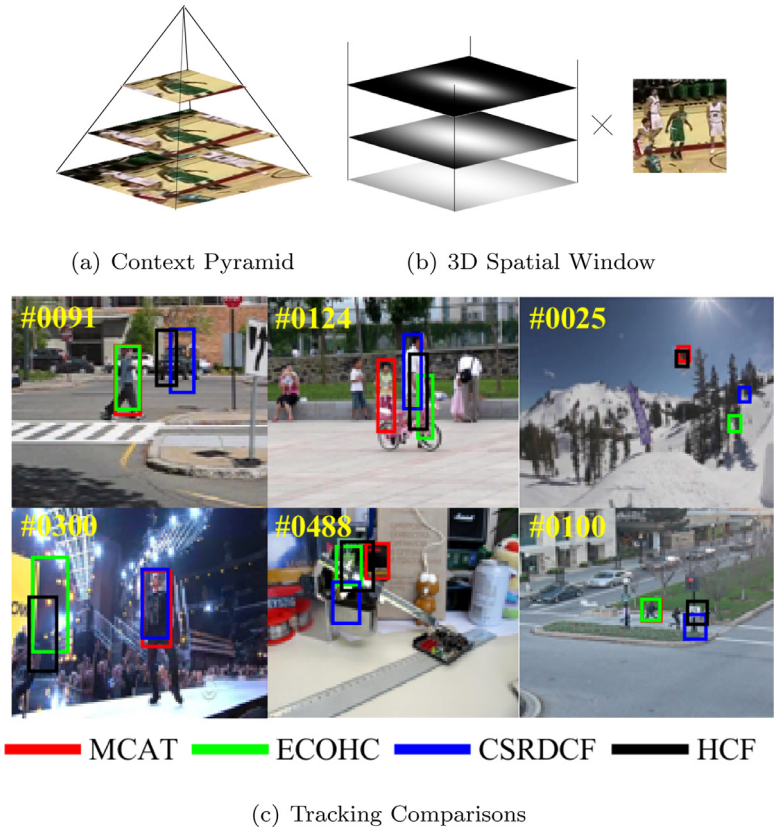

Fig. 1. (a) Context pyramid representation. (b) Three-dimensional (3D) spatial window model. (c) Comparison of the proposed multi-level context adaptive tracker (MCAT) with the renowned CF-based trackers: ECOHC [5], CSRDCF [6], and HCF [7].

a different spatial window for each level to achieve a 3D context representation as shown in Fig. 1(b).

The top levels of the pyramid contain minimal background content. The corresponding correlation filters provide a better description of the target and are reliable in ordinary tracking. The low levels of the pyramid contain varying degrees of backgrounds. The corresponding correlation filters explore more relationships between the target and background, and they are helpful in cases of occlusions and distractors. The whole pyramid will comprises a complete representation of the target and its context.

In addition, we propose a method of controlling the effect of backgrounds in different levels both in learning and tracking. Specifically, the 3D spatial windows in learning and tracking are carefully designed for the context pyramid. Furthermore, the correlation filters are jointly learned for the pyramid representation, and the target is tracked with adaptive prediction fusion. We evaluated the proposed multi-level context-adaptive tracking (MCAT) method on the tracking benchmark OTB2015 [2]. As shown in Fig. 1(c), with hand-crafted features, the proposed tracker demonstrates favorable performance against state-of-the-art methods, such as ECO-HC [5], CSRDCF [6], which are based on hand-crafted features, and HCF [7], which is based on deep convolution neural networks.

2. Related work

In this section, related tracking methods are discussed, primarily with respect to discriminative correlation filters.

Correlation filters were introduced to visual tracking by MOSSE [11], and a remarkably high speed was obtained using the fast Fourier transform (FFT). The feature representation was then extended into multi-channels, such as Colornames [13] and HOG [16]. In CSK [17] and the later KCF [12], the correlation filters model is interpreted with circular samples, and the tracking problem is translated into a regression problem. DSST [18] was proposed as a discriminative scale space filter and can efficiently estimate the target scale. SAMF [19] integrates HOG and CN features. Additionally, it simultaneously estimates the target scale and position in several scale samples. Many other improvements have been implemented in correlation filters, such as part-based [20–22], long-term [23], response adaptive [24], training set adaptive [25], and CNN based [5,7,26] methods.

Correlation filters have a bound effect, which means incorrect information exists in the circular samples. A detailed explanation is provided in [14]. The bound effect is inevitable on account of the FFT. Some methods have been proposed to reduce the bound effect. A Hann spatial window is used to weight the features in CF trackers, ranging from the original MOOSE [11] to the latest ECO [5]. However, it is minimally or not at all effective when the window size is enlarged. CFLB [15] uses a masking matrix to cut the circular samples, and it was improved into BACF [27] for multi-channel tracking. Moreover, CSRDCF [6] was proposed for spatially constrained correlation filters that constrain the filters to desired shapes. Both BACF and CSRDCF have good discrimination against backgrounds. However, they lack utilization of the background and perform poor in cases of occlusion. SRDCF [14] places a penalty weight on the filters to learn spatially regularized correlation filters. C-COT [28] learns continuous correlation filters for multi-resolution features based on SRDCF. ECO [5] introduces efficient convolution operators for C-COT. It has remarkable performance but causes errors in some cases of occlusion and background clutter with hand-crafted features.

Meanwhile, the context model is helpful in tracking [1]. The context means the target and its surrounding backgrounds. In a context-aware tracker [29], auxiliary objects are mined to help

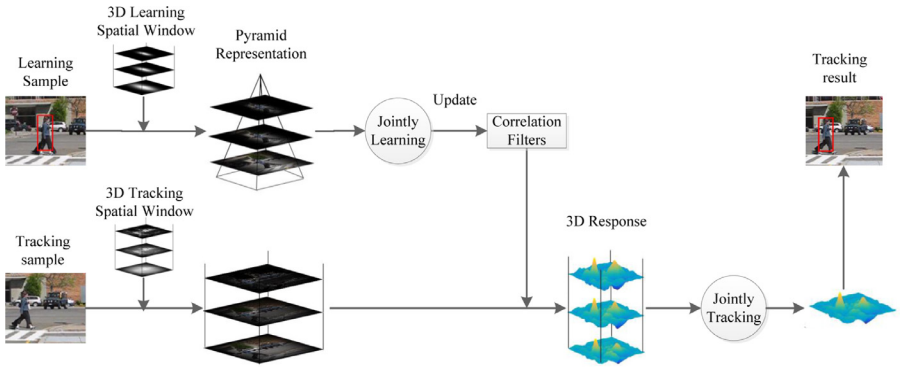

Fig. 2. Framework of the proposed MCAT method. In the learning stage, a 3D spatial window is weighted on the base learning sample to achieve a context pyramid representation. By jointly learning the multi-level representation, 3D correlation filters can be obtained to update the DCF models. In the tracking stage, another 3D spatial window is used to obtain the 3D context-adaptive sample. The multi-level responses are adaptively fused to produce robust prediction by jointly tracking.

track the target. Actually, the correlation filters regard the context as a whole and track it accordingly. Thus, the correlation filters can potentially utilize the background information. CACF [30] extracts surrounding backgrounds to jointly learn the filters, with the base sample. DAT [31] discriminates the target from the background by using a statistic color histogram. Staple [32] combines KCF with DAT for complementary learners. CSRDCF [6] utilizes DAT to segment the target before learning stage to reduce the effect of background. Moreover,background restriction has shown advantages in the learning stage. Inspired by SWCF [33], which iteratively optimizes the spatial window to achieve better performance, we propose a context-adaptive spatial window by restricting the backgrounds. We contend that background restriction can also be beneficial in the tracking stage.

Furthermore, multi-kernel, multi-template, and multi-task learning methods have made progress in correlation filters [34–37]. It is noteworthy that a small amount of background is utilized in smooth tracking, whereas a large amount of background is required to locate the target when it rapidly varies or is occluded. Accordingly, we introduce a multi-level context structure that is robust against the above various challenges by jointly learning and tracking the multi-level contexts.

3. Proposed approach

In this section, we firstly review the original discriminative correlation filters. Then, the context pyramid representation is proposed. From that point, the 3D context-adaptive spatial windows are introduced. Finally, we jointly learn and track with the pyramid in a multi-level structure. The framework of our method is deplicted in Fig. 2.

3.1. Discriminative correlation filters

Multi-channel discriminative correlation filters were introduced in [18]. In the learning stage, a base sample representation f is extracted at the position of the target with d feature channels. For each feature channel $l \in 1, \ldots, d$, f^l and h^l denote the corresponding feature and filter. The objective is to learn a correlation filter h by minimizing the error of the correlation response compared to the predefined output g,

$$\min_h ||g - \sum_{l=1}^{d} h^l * f^l||_2^2 + \lambda \sum_{l=1}^{d} ||h^l||_2{}^2, \tag{1}$$

where $*$ denotes the circular correlation operator, and λ is a regularization parameter. The output, g, typically has a Gaussian function [11].

The minimizer has a closed-form solution in the Fourier domain using the Parseval's formula,

$$H^l = \frac{\overline{G}F^l}{\sum_{k=1}^{d}\overline{F^k}F^k + \lambda}, l = 1, \ldots, d, \tag{2}$$

where the capital letters denote the discrete Fourier transform (DFT) of the corresponding quantities, such as $H = DFT(h)$, which hence defines the capital letters hereafter.

The filters are updated online in the approach introduced in MOOSE [11] with a forgetting factor η,

$$\begin{aligned} H_t^l &= \frac{A_t^l}{B_t}, l = 1, \ldots, d \\ A_t^l &= (1-\eta)A_{t-1}^l + \eta\overline{G}F_t^l, l = 1, \ldots, d \\ B_t &= (1-\eta)B_{t-1} + \eta\sum_{k=1}^{d}\overline{F_t^k}F_t^k. \end{aligned} \tag{3}$$

In the tracking stage, a new sample representation, z, is extracted from the region of interest. The new correlation response y can be obtained by fast tracking,

$$Y = \sum_{l=1}^{d} H^l Z^l, \tag{4}$$

where $Y = DFT(y)$, $Z = DFT(z)$.

The target is located where y is the maximum. In addition, the training sample f and tracking sample z are weighted with a fixed spatial window in the feature extraction stage to restrain the bound effect.

3.2. Context pyramid representation

Context is defined as the target along with its surrounding background. The context can indicate the latent relationship between the target and background. In DCF tracking, the context is learned and tracked as a whole. In the ordinary tracking process, the backgrounds should be restricted in the context because less background will contribute to a more precise target model. Nevertheless, in considering challenges, such as occlusion and target loss, the context that contains more background will be more beneficial to locating the target. Accordingly, we propose a context pyramid representation in which different levels contain varying degree of backgrounds.

An n-levels context pyramid representation for the sample in the region of interest is defined as $r = \{r_i, i = 1, \ldots, n\}$. A higher level will contain less background. However, it is time-consuming to extract features for each level. In addition, it is not suitable to

learn the pyramid in a joint framework because different levels have different sizes.

To jointly and conveniently learn the 3D filters, we utilize a fixed window size (W, H) for targets with size (w, h) and the extended spatial window model to realize the context pyramid. For each level i of the pyramid, the backgrounds are restricted to different degrees, and the sample is represented by weighting a base sample f with a spatial window$_i$ p_i,

$$r_i = p_i \odot f, i = 1, \ldots, n. \tag{5}$$

Here, p_i with a larger i will allocate lower weights to backgrounds to restrict their effect. Its calculation is introduced in the Section 3.3.

In this way, the feature extraction must be conducted only once, and all levels have a consistent sample size. Meanwhile, we can use the gradual weights to smoothly reduce the backgrounds. However, the spatial window for each level should be finely designed to meet the requirement of the pyramid structure.

Moreover, the original discriminative correlation filters can be regarded as a single-level context pyramid. The context pyramid model extends the representation of the sample, thereby providing a more complete relationship between the target and the surrounding background for tracking.

3.3. Context-adaptive spatial windows

In the context pyramid, different context levels contain varying degrees of backgrounds to describe different levels of the latent relationship. We utilize the spatial window strategy where greater weights are placed on the target and smaller ones are placed on the backgrounds to reduce them. The spatial window should be flexible and adaptive to different context levels to construct a good pyramid.

We introduce the 3D context-adaptive spatial windows. The backgrounds are reduced to varying degrees in different levels both in the learning and tracking stages. The spatial window strategy results in a feature transformation from the original sample to a background reduced domain. It is notable that the closet solution and fast tracking in DCF are preserved.

In the learning stage, the target is located in the center of the window. Hence, pixels farther from the center are more likely to be parts of the background. In the original DCF, the spatial window is a Hann window,

$$\begin{aligned} p_{cos} &= hann(W) * hann(H)', \\ hann(W) &= \frac{1}{2}\left(1 - \frac{cos(2\pi\,(0 : W-1)')}{(W-1)}\right). \end{aligned} \tag{6}$$

This Hann window is adaptive to the window size (W, H). However, it can neither adapt the aspect ratio of the target nor change the field of interest in the window. To learn a good relationship between the target and backgrounds, the background should be evenly distributed around the target. Meanwhile, the Hann window cannot adapt to the different context levels. A higher level i should have a wider field of interest.

Therefore, the Gaussian window is introduced: for each level i of the pyramid,

$$\begin{aligned} p_i &= gauss\left(W, \frac{h}{H}\theta_i\right) * gauss\left(H, \frac{w}{W}\theta_i\right)', \\ gauss(W, \theta) &= e^{-\frac{1}{2}(\theta\frac{-W/2:W/2}{W/2})^2}. \end{aligned} \tag{7}$$

As shown in Fig. 3, the spatial window has an aspect ratio that differs from that of the target. The Hann window effectively restricts the backgrounds in the vertical direction; nonetheless some background remains in the horizontal direction. The Gaussian window can retain the features of the target, whereas it restricts

Fig. 3. Examples of Hann (left) and Gaussian (right) windows weighted on the learning sample.

those of the surrounding backgrounds. A higher level i in the context pyramid has a lower θ_i and thus a wider field of interest. In the tracking stage, the corresponding spatial windows $q = \{q_i, i = 1, \ldots, n\}$ are introduced to reduce the backgrounds. However, the target can be situated anywhere, and thusly is the background. To adapt different context levels, we estimate the target distribution and set varying weights t for different levels. Intuitively, the distribution of the target is predicted based on Bayesian inference from views of spatial and value domains,

$$q_i(x) = q_s(x)(t_i q_v(I(x)) + 1 - t_i), \tag{8}$$

where q_s and q_v are the spatial and value weights to denote the probabilities that the target appears in each pixel from views of spatial and value domains, respectively. x denotes any pixel in the sample, and $I(x)$ is the value of x. In addition, t_i is a parameter relative to the level of the pyramid and it ranges in [0,1]. A larger t_i will restrict more backgrounds.

From the spatial view, the center of the sample is located where the target is situated in the last frame. Thus, the target is more likely to remain close to the center. Hence, q_s is defined as

$$q_s(x) = p_{cos}^{\gamma}(x), \tag{9}$$

where $p_{\cos}$ is explained in (6), and γ is a parameter that controls how likely the target is to leave the center of the sample.

With respect to value, a color belongs to the target with more probabilities when more pixels with the given color belongs to the target [31]. Thus, the model can be summarized as

$$q_v(j) = \begin{cases} \frac{\rho_O(j)}{\rho_O(j)+\rho_B(j)}, & if\ j \in O \bigcup B \\ 0.5, & Otherwise \end{cases} \tag{10}$$

where j denotes any color. O and B indicate the target and the background, respectively, and ρ represents the color histogram.

Some weighted tracking samples are demonstrated in Fig. 4. By controlling parameters t_i and γ, the spatial window can be adaptive to different context levels.

3.4. Multi-level context-adaptive correlation tracking

In the context pyramid, each single-level can be used to train a tracker. However, the target can be tracked only in a scene wherein the backgrounds are restricted to a fixed degree. When considering the general tracking problem, we find that the restriction on backgrounds should be more flexible. Different relationship levels can be learned from different context levels. All levels should be simultaneously considered. Hence, the multi-level context-adaptive correlation tracking method is proposed to learn and track using the context pyramid representation in a joint framework.

In the learning stage, we simply allocate weights α, where $\sum_{i=1}^{n}\alpha_i = 1$, to different levels and jointly learn the correlation filters,

$$\min_h ||g - \sum_{i=1}^{n}\alpha_i \sum_{l=1}^{d} h_i^l * r_i^l||_2^2 + \lambda \sum_{i=1}^{n}\sum_{l=1}^{d} ||h_i^l||_2^2. \tag{11}$$

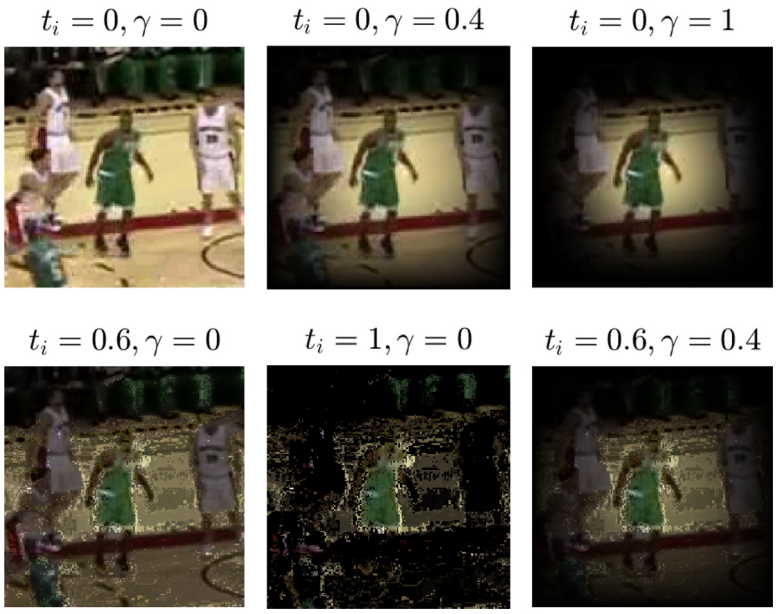

Fig. 4. Backgrounds restricted samples with different t_i and γ. Here, the spatial windows are weighted on the image for visualization. In our method, they are weighted on the features. A greater γ will lead to a narrower field of interest, and the target is considered to scarcely move. A greater t_i contributes to restricting the backgrounds.

The objective function can be solved similar to the approach in (1),

$$H_i^l = \frac{\alpha_i \overline{G} R_i^l}{\sum_{i=1}^{n} \alpha_i^2 \sum_{k=1}^{d} \overline{R_i^k} R_i^k + \lambda}, \begin{cases} l = 1, \ldots, d \\ i = 1, \ldots, n \end{cases} \tag{12}$$

where R denotes the discrete Fourier transforms of the corresponding quantities r. r_i^l and h_i^l denote the i_{th} level and the l_{th} channel.

In the tracking stage, a corresponding correlation response can be obtained by

$$Y = \sum_{i=1}^{n} \alpha_i \sum_{l=1}^{d} H_i^l Z_i^l. \tag{13}$$

However, different levels of the context pyramid have different influences at different moments. Fixed parameter α is unfavorable in the tracking process. Because the background should be more greatly reduced when the target moves smoothly, whereas it should be less restricted when the target rapidly changes or is occluded. In other words, the reliability of each single-level varies in the tracking process. More weights should be placed on more reliable levels in every instant. Accordingly, we propose an adaptive method to rearrange the weights α to β as the specific tracking environment changes.

Firstly, the prediction reliability of each level should be evaluated. We utilize the high-confidence average peak-to-correlation energy (APCE) measure [38].

$$APCE_y = \frac{N|y_{max} - y_{min}|^2}{\sum_x (y(x) - y_{min})}, \tag{14}$$

where N is the number of features in the sample, and $y_{\max}$ and $y_{\min}$ are the respective maximum and minimum value of the response y.

Secondly, we define the prediction loss of each level as

$$L_i = \frac{1}{APCE_{y_i}^2}, \tag{15}$$

where y_i is the inverse discrete Fourier transform of $Y_i = \sum_{l=1}^{d} H_i^l Z_i^l$. A more reliable prediction response with context level i has a greater APCE value and thus a smaller L_i.

Thirdly, the objective of the joint tracking can be denoted as minimizing the loss of all levels,

$$\begin{aligned} &\min_{\beta} \sum_{i=1}^{n} \beta_i L_i + \lambda_L \sum_{i=1}^{n} \frac{\beta_i^2}{\alpha_i^2}, \\ &s.t. \sum_{i=1}^{n} \beta_i = 1, \\ &\beta_i \geq 0, i = 1, \ldots, n, \end{aligned} \tag{16}$$

where λ_L is a regularization parameter. When the λ_L inclines to the infinite, the solution will be $\beta = \alpha$. The weights will not change. When the λ_L is zero, the solution will be all zeros except the one with the smallest loss. Only one level of the pyramid will be selected. With a proper λ_L, the weights will move from low reliable levels to high ones by solving the above optimization problem.

Eq. (16) is a convex quadratic programming, and can be efficiently solved. With the learned β, the response of the current frame can be revised by

$$Y = \sum_{i=1}^{n} \beta_i Y_i. \tag{17}$$

A threshold strategy is used to reduce error update in tracking process. When the APCE value of y in the current frame is greater than its historical average value with a certain ratio, μ, the tracking result is considered reliable. The correlation filter model is updated in the approach similar to (3).

4. Experimental analysis

To evaluate the proposed MCAT method, we first detail the implementation. Then, we analyze the effectiveness of different MCAT components. Finally, many experimental results and analyses are provided.

4.1. Implementation details

The widely-used Colornames [13] and HOG [16] features were implemented. The scale-adaptive strategy involved was the same as SAMF [19], which tracked the target in seven scale samples. The

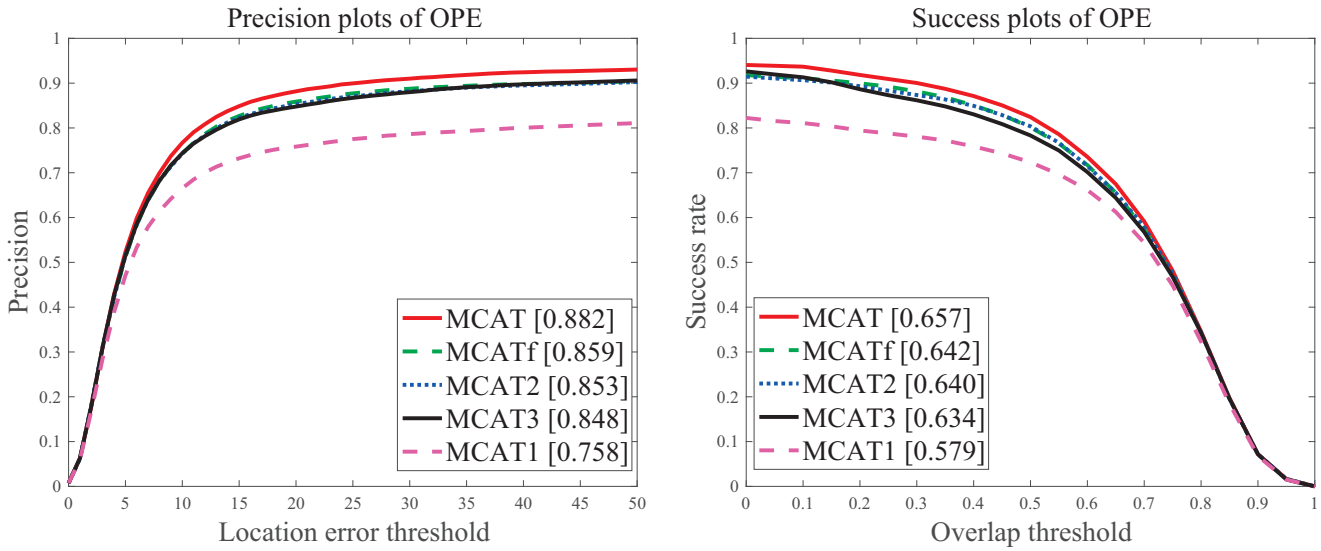

Fig. 5. Comparisons of MCATs with different configurations. The precision plots and success plots are demonstrated. The legends are the DP scores at the threshold of 20 as well as the AUC scores. Different levels show different performances, and the proposed adaptive prediction fusion method is shown to be helpful.

scale step was 1.01. Our tracker ran at an average speed of 22.5 frames per second (FPS) using MATLAB R2016b on a 3.6GHz Intel Core i7 PC with 16 G RAM. The sample window was a square block with an area of 12 times that of the target. We used a three-level structure to represent the pyramid for simplicity and named them levels 1, 2, and 3 from the bottom to the top of the pyramid, respectively. The detailed parameters were $\lambda = 10^{-4}$, $\gamma = 0.4$, $n = 3$, $\theta = [10, 15, 20]$, $t = [0.2, 0.6, 1]$, $\alpha = [0.25, 0.25, 0.5]$, $\lambda_L = 0.0005$, $\eta = 0.009$, and $\mu = 0.2$.

Our method was evaluated on the object tracking benchmark (OTB100) [2] with 100 video sequences. Two criteria for one-pass evaluation were used to assess the performance. The distance precision (DP) score denoted the percentage of successfully tracked frames that the center position error did not exceed the threshold of 20 pixels. The area under the curve (AUC) score was the area under the success rate curves that indicated the average overlap rate at different thresholds. The overlap score of the predicted target region Rt and its ground truth Rg is defined as $S = \frac{Rt \bigcap Rg}{Rt \bigcup Rg}$.

Our method was further evaluated on the 2016 visual object tracking (VOT2016) benchmark [4] with 60 video sequences. The mostly used criteria is expected average overlap (EAO), and a higher value indicates a better tracker.

4.2. Components evaluations

We evaluated the effectiveness of different components, including the context-adaptive spatial window methods, the pyramid representation or multi-level structure, and the adaptive fusion of multi-level prediction responses.

4.2.1. Context-adaptive spatial window evaluations

The context-adaptive methods were used to restrict the backgrounds in learning and tracking. To check their dependent effect, a baseline SAMF [19] with a linear kernel was implemented and analyzed on the OTB100. We tested a single-level spatial window. Firstly, the baseline was tested with an expanded sample region. Secondly, the proposed learning spatial window in (7) was implemented. Thirdly, the method was equipped with the proposed tracking window in (8). The four configurations were those with the same common parameters. We simply append the components, step by step, without further parameter tuning. The experimental results are shown in Table 1.

Table 1
Analysis of our context adaptive spatial windows on OTB100.

	Baseline SAMF	Expanded region	Learning window	Tracking window
AUC	0.562	0.557	0.573	0.583
DP	0.748	0.734	0.77	0.82
FPS	52.6	40	39.7	26.8

Table 2
Comparison of the scale adaptive strategies SAMF [19] and DSST [18] for different levels of MCAT on OTB100.

	Level 1		Level 2		Level 3	
	SAMF	DSST	SAMF	DSST	SAMF	DSST
AUC	0.579	0.549	0.64	0.62	0.634	0.593
DP	0.758	0.742	0.853	0.844	0.848	0.806

The expanded region helped track the target in the case of fast motion. However, the SAMF performance was reduced because of the increased amount of background. Our learning spatial window enabled the sample features to be adaptive to the sizes of the target and sample. The improved performance with the tracking windows confirmed that background restriction in tracking was helpful. The time consumption resulting from the expanded region and background detection was unavoidable but affordable.

Two different scale strategies were tested on different MCAT levels. As mentioned above, they were SAMF [19] and DSST [18]. The experimental results are listed in Table 2. As shown in the table, Level 1 contains a large amount of background, it is easily influenced by the background. Level 2 contains less background and level 3 most notably restricts the background. They both performed well. Level 2 is somewhat better because the surrounding background is helpful for predicting the target location. Level 3 focuses almost only on the target and will fail once the target is occluded or rapidly changes. Different levels have different performances because varying degrees of backgrounds were used. SAMF performed better than DSST with our methods because SAMF estimates the scale and the location simultaneously, whereas DSST estimates the scale after locating the target. Thus, SAMF was finally implemented in our tracker.

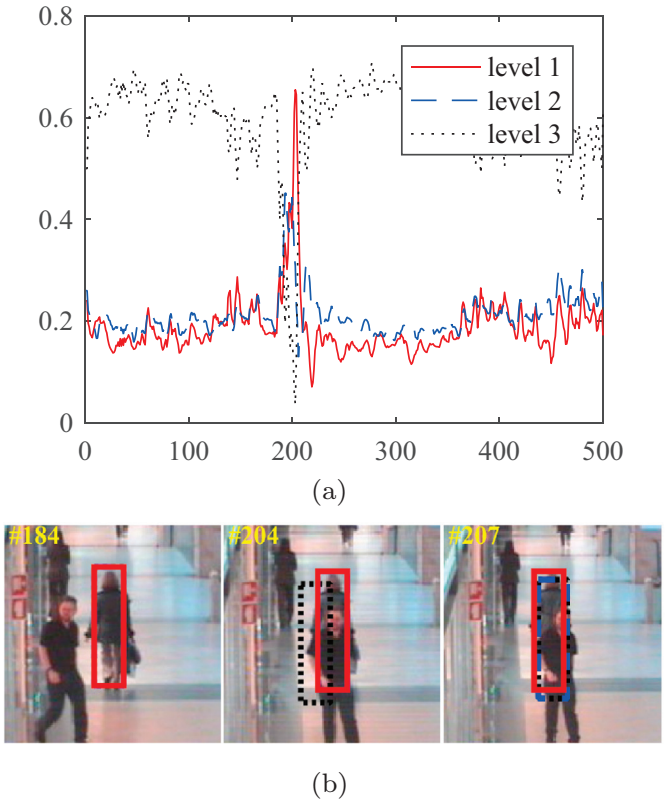

Fig. 6. Real-time fusion parameter plots in the video *Walking2*, and some snapshots. Level 3 causes errors when the target is occluded. Thus, our adaptive fusion method reduces its weight and accurately predicts the target location.

4.2.2. Multi-level context adaptive correlation tracking evaluations

We compared the MCAT with or without the adaptive prediction fusion on OTB100. A complete MCAT tracker was tested. MCAT1, MCAT2, MCAT3 denote 3 different levels of the MCAT tracker, respectively. MCATf was the MCAT tracker with fixed prediction parameter α with (13). The comparisons are shown in Fig. 5. Different levels have different performances. The second level (MCAT2) performed best with an AUC score of 0.64 and a precision score of 0.853. The prediction fusion with fixed parameters (MCATf) slightly improves the performance as the AUC score is increased from 0.64 to 0.642 and the precision score is creased from 0.853 to 0.859. However, with the proposed adaptive prediction fusion method, the tracker is significantly improved as the AUC score is increased raised from 0.64 to 0.657 and the precision score is increased from 0.853 to 0.882.

To further elucidate the multi-level joint tracking method, we obtained the video *Walking2* as an example. Fig. 6 depicts the real-time fusion parameter plots. When the person moves smoothly, the three levels perform well. The Level 3 contains the least background. Thus, it is regarded as the most reliable in general tracking and is set as the greatest weight. However, when the target is occluded, as shown in Fig. 6(b), Level 3 shows a large error in frame 204. The tracker would have turned to track the other person if fixed parameters were used. Nonetheless, the adaptive fusion method allocates more weights on the low levels, as shown in Fig. 6(a), and tracks the target successfully. Therefore, the proposed adaptive prediction fusion method can make different levels complementary.

4.3. State-of-the-art comparison

Using OTB100 [2], the proposed MCAT was compared with nine state-of-the-art trackers based on hand-crafted features, including ECOHC [5], CSRDCF [6], LCT [23], SRDCF [14], SAMF [19], DSST [18], Staple [32], Struck [10], and TLD [9]. We report the overall performance for one-pass evaluation (OPE) in all 100 videos, as shown in Fig. 7. Our method yields the best performance with an AUC score of 65.7% and a DP score of 88.2%. Among the others, ECOHC ranks second with an AUC score of 64.3% and a DP score of 85.6%. Using OTB100, the 100 videos were annotated with 11 attributes representing the challenges of illumination variation (IV), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane rotation (OPR), out of view (OV), background clutter (BC), and low resolution (LR). Each video was labelled with several of the challenges. The success rate plots and AUC scores of the ten trackers in OTB100 are depicted in Fig. 8. MCAT and ECOHC perform better than the other trackers under all challenges. MCAT exhibits the best performance in terms of AUC score in nine attributes: IV (65.9%), SV (62.2%), DEF (61.5%), MB(63.8%), IPR(61.2%), OPR (63.2%), OV (62.1%), and

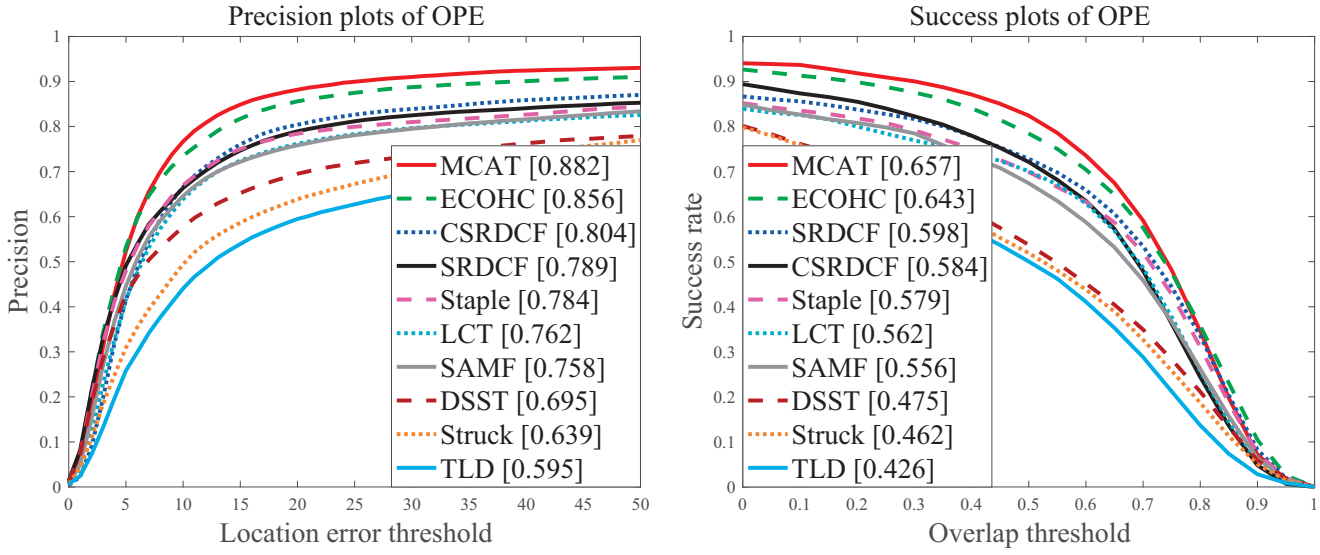

Fig. 7. Comparisons of MCAT with state-of-the-art trackers on OTB100. The DP scores and AUC scores are respectively shown in the legends. Our approach shows the best performance.

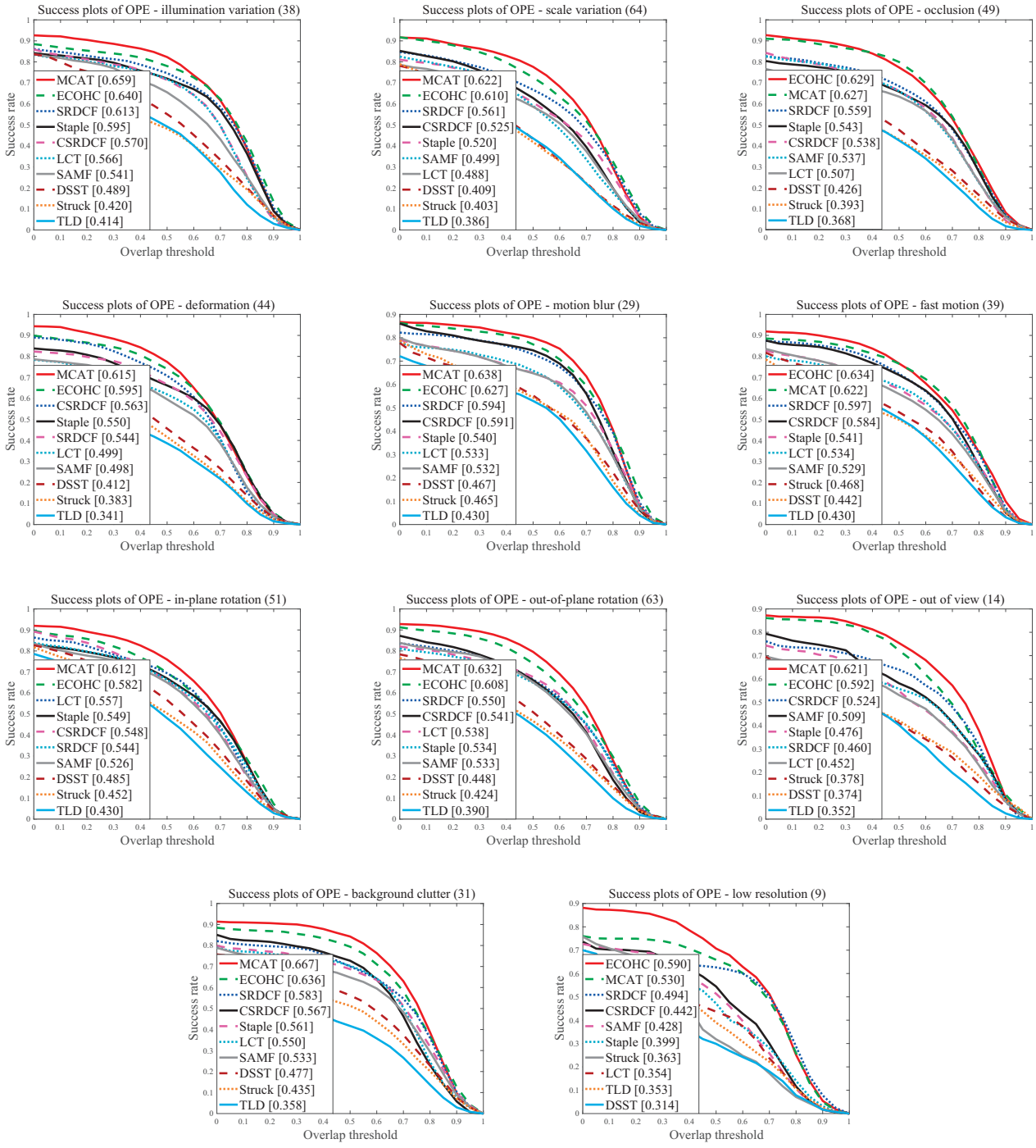

Fig. 8. Success rate plots of different challenges in OTB100. The legends contain the AUC scores.

BC (66.7%). In the remaining two attributes, MCAT also performs similar to ECOHC. Some studies of deep-learning-based methods have also reported the performances, such as CNN-features based trackers: ECO [5], MCPF [36], DeepSRDCF [39], HCF [7], HDT [26], CNN-SVM [40], DNN ones, MDNet [41], CREST [42], DNT [43], PTAV [44], and ADNet [45]. We list their overall performance criteria in Table 3. It is evident that our method MCAT performs better than most deep learning methods and ranks third. Only ECO with CNN features and MDNet perform better than MCAT using hand-crafted features.

With VOT2016, the proposed method was compared with four state-of-the-art trackers including DeepSRDCF [39], SRDCF [14], SAMF [19], and DSST [18]. Fig. 9 shows the EAO, A and R comparisons. MCAT1 is a single-level tracker with an EAO of 0.2538, and is better than the SRDCF with an EAO of 0.2471. MCAT is a 3-level configuration and gets a best EAO of 0.2821, and is better than the deep features based DeepSRDCF with an EAO of 0.2763. VOT is a very challenging benchmark, where background clutters are serious and target rotations are common, so exploiting high-level vision features such as deep features will help. But with hand-crafted features, the multi-level MCAT achieves the best performance. It reveals that our method makes the best of the feature structure.

Table 3
Comparison of MCAT with recent deep-learning-based methods. The top three are labeled in red, green, and blue. MCAT shows a better performance than most deep learning methods and ranks third.

	MCAT	ECO	MCPF	DeepSRDCF	HCF	HDT
AUC	0.657	0.694	0.628	0.635	0.562	0.564
DP	0.882	0.91	0.873	0.851	0.837	0.848
	CNNSVM	MDNet	CREST	DNT	PTAV	ADNet
AUC	0.554	0.678	0.623	0.627	0.635	0.646
DP	0.814	0.909	0.837	0.851	0.849	0.88

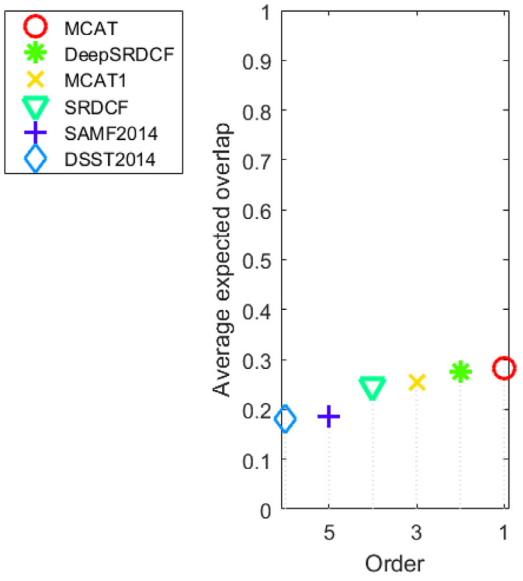

Fig. 9. Experimental results on VOT2016 with respect to expected average overlap (EAO).

5. Conclusion

In this paper, we presented a context pyramid representation for visual tracking. This representation can better describe the relationships between the target and its surrounding backgrounds than traditional representations. We proposed a 3D spatial window to adaptively construct different context levels. The feature extraction must be processed only once and the spatial windows of different levels contribute to different context levels. For each level of the context pyramid, the context adaptive strategy is introduced to restrict the background both in learning and tracking stages. The strategy is adaptive to the size of the target and samples.

In addition, our approach jointly learns and tracks the context pyramid by allocating weights to different levels in the discriminative correlation filters framework. Additionally, the adaptive weights estimation method achieves a complement among different context levels. Furthermore, by means of extensive experimental results, our tracker demonstrates promising performance with extensive experimental results. Using hand-crafted features, our method achieves comparable precision to deep learning methods.

References

[1] Y. Wu, J. Lim, M.-H. Yang, Online object tracking: a benchmark, in: CVPR, 2013, pp. 2411–2418.
[2] Y. Wu, J. Lim, M.-H. Yang, Object tracking benchmark, PAMI 37 (9) (2015) 1834–1848.
[3] M. Kristan, J. Matas, A. Leonardis, The visual object tracking vot2015 challenge results, in: ICCV Workshop, 2015, pp. 564–586.
[4] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. ehovin, T. Vojr, G. Hger, A. Lukei, G. Fernndez, The Visual Object Tracking VOT2016 Challenge Results, Springer International Publishing, 2016.
[5] M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, Eco: Efficient convolution operators for tracking, in: CVPR, 2017, pp. 6931–6939.
[6] A. Lukei, T. Voj, L. ehovin, J. Matas, M. Kristan, Discriminative correlation filter with channel and spatial reliability, in: CVPR, 2017, pp. 4847–4856.
[7] C. Ma, J.-B. Huang, X. Yang, M.-H. Yang, Hierarchical convolutional features for visual tracking, in: ICCV, 2015, pp. 3074–3082.
[8] B. Babenko, M.-H. Yang, S. Belongie, Visual tracking with online multiple instance learning, in: CVPR, 2009, pp. 983–990.
[9] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, PAMI 34 (7) (2012) 1409–1422.
[10] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S.L. Hicks, P.H.S. Torr, Struck: structured output tracking with kernels, PAMI 38 (10) (2016) 2096–2109.
[11] D.S. Bolme, J.R. Beveridge, B.A. Draper, Y.M. Lui, Visual object tracking using adaptive correlation filters, in: CVPR, 2010, pp. 2544–2550.
[12] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, PAMI 37 (3) (2015) 583–596.
[13] M. Danelljan, F. Shahbaz Khan, M. Felsberg, J. Van de Weijer, Adaptive color attributes for real-time visual tracking, in: CVPR, 2014, pp. 1090–1097.
[14] M. Danelljan, G. Hager, F. Shahbaz Khan, M. Felsberg, Learning spatially regularized correlation filters for visual tracking, in: ICCV, 2015, pp. 4310–4318.
[15] H. Kiani Galoogahi, T. Sim, S. Lucey, Correlation filters with limited boundaries, in: CVPR, 2015, pp. 4630–4638.
[16] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, 2005, pp. 886–893.
[17] C. Rui, P. Martins, J. Batista, Exploiting the circulant structure of tracking-by-detection with kernels, in: ECCV, 2012, pp. 702–715.
[18] M. Danelljan, G. Hager, F.S. Khan, M. Felsberg, Discriminative scale space tracking, PAMI 39 (8) (2017) 1561–1575.
[19] Y. Li, J. Zhu, A scale adaptive kernel correlation filter tracker with feature integration, in: ECCV, 2014, pp. 254–265.
[20] Y. Li, J. Zhu, S.C.H. Hoi, Reliable patch trackers: robust visual tracking by exploiting reliable patches, in: CVPR, 2015, pp. 353–361.
[21] T. Liu, G. Wang, Q. Yang, Real-time part-based visual tracking via adaptive correlation filters, in: CVPR, 2015, pp. 4902–4912.
[22] S. Liu, T. Zhang, X. Cao, C. Xu, Structural correlation filter for robust visual tracking, in: CVPR, 2016, pp. 4312–4320.
[23] C. Ma, X. Yang, C. Zhang, M.-H. Yang, Long-term correlation tracking, in: CVPR, 2015, pp. 5388–5396.
[24] A. Bibi, M. Mueller, B. Ghanem, Target response adaptation for correlation filter tracking, in: ECCV, 2016, pp. 419–433.
[25] M. Danelljan, G. Hger, F.S. Khan, M. Felsberg, Adaptive decontamination of the training set: a unified formulation for discriminative visual tracking, in: CVPR, 2016, pp. 1430–1438.
[26] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, M.-H. Yang, Hedged deep tracking, in: CVPR, 2016, pp. 4303–4311.
[27] H.K. Galoogahi, A. Fagg, S. Lucey, Learning background-aware correlation filters for visual tracking, in: ICCV, 2017, pp. 1144–1152.
[28] M. Danelljan, A. Robinson, F.S. Khan, M. Felsberg, Beyond correlation filters: learning continuous convolution operators for visual tracking, in: ECCV, 2016, pp. 472–488.
[29] M. Yang, Y. Wu, G. Hua, Context-aware visual tracking, PAMI 31 (7) (2009) 1195–1209.
[30] M. Mueller, N. Smith, B. Ghanem, Context-aware correlation filter tracking, in: CVPR, 2017, pp. 1387–1395.
[31] H. Possegger, T. Mauthner, H. Bischof, In defense of color-based model-free tracking, in: CVPR, 2015, pp. 2113–2120.
[32] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, P.H.S. Torr, Staple: complementary learners for real-time tracking, in: CVPR, 2016, pp. 1401–1409.
[33] E. Gundogdu, A.A. Alatan, Spatial windowing for correlation filter based visual tracking, in: ICIP, 2016, pp. 1684–1688.
[34] M. Tang, J. Feng, Multi-kernel correlation filter for visual tracking, in: ICCV, 2016, pp. 3038–3046.
[35] A. Bibi, B. Ghanem, Multi-template scale-adaptive kernelized correlation filters, in: ICCV Workshop, 2015, pp. 613–620.
[36] T. Zhang, C. Xu, M.-H. Yang, Multi-task correlation particle filter for robust object tracking, in: CVPR, 2017, pp. 4819–4827.
[37] L. Zhang, P.N. Suganthan, Robust visual tracking via co-trained kernelized correlation filters, Pattern Recognit. (2017).
[38] M. Wang, Y. Liu, Z. Huang, Large margin object tracking with circulant feature maps, in: CVPR, 2017, pp. 4800–4808.
[39] M. Danelljan, G. Hager, F.S. Khan, M. Felsberg, Convolutional features for correlation filter based visual tracking, in: ICCV Workshop, 2015, pp. 621–629.
[40] S. Hong, T. You, S. Kwak, B. Han, Online tracking by learning discriminative saliency map with convolutional neural network, in: ICML, 2015, pp. 597–606.
[41] H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, in: CVPR, 2016, pp. 4293–4302.
[42] Y. Song, C. Ma, L. Gong, J. Zhang, R. Lau, M.H. Yang, Crest: convolutional residual learning for visual tracking, in: ICCV, 2017, pp. 2574–2583.
[43] Z. Chi, H. Li, H. Lu, M.H. Yang, Dual deep network for visual tracking, TIP 26 (4) (2017) 2005–2015.

[44] H. Fan, H. Ling, Parallel tracking and verifying: a framework for real-time and high accuracy visual tracking, in: ICCV, 2017, pp. 5487–5495.
[45] S. Yun, J. Choi, Y. Yoo, K. Yun, Y.C. Jin, Action-decision networks for visual tracking with deep reinforcement learning, in: CVPR, 2017, pp. 1349–1358.

Peng Liu is an associate professor at the College of Computer Science and Technology, Harbin Institute of Technology. He received his Ph.D. degree of microelectronics and solid state electronics from Harbin Institute of Technology in 2007. His research interest covers image processing, computer vision, and pattern recognition.

Chang Liu is a Ph.D. candidate at the College of Computer Science and Technology, Harbin Institute of Technology. He received his bachelor's degree of Computer Science and Technology from Harbin Institute of Technology in 2014. His research interest covers computer vision and pattern recognition.

Wei Zhao is an associate professor at the College of Computer Science and Technology, Harbin Institute of Technology. She received her Ph.D. degree of computer application technology from Harbin Institute of Technology in 2006. Her research interest covers computer vision and pattern recognition. Corresponding author of this paper.

Xianglong Tang is a professor at the College of Computer Science and Technology, Harbin Institute of Technology. He received his Ph.D. degree of computer application technology from Harbin Institute of Technology in 1995. His research interest covers pattern recognition.