A Saliency-Based Object Tracking Method for UAV Application

Jinyu Yang[1], Wenrui Ding[2], Chunlei Liu[2](✉), and Zechen Ha[2]

[1] The Hong Kong University of Science and Technology,
Clear Water Bay, Hong Kong
yangjinyu@buaa.edu.cn
[2] Beihang University, Beijing, China
{ding,liuchunlei,hazechen36}@buaa.edu.cn

Abstract. Visual tracking has been an active and complicated research area in computer vision for recent decades. In the area of unmanned aerial vehicle (UAV) application, establishing a robust tracking model is still a challenge. The kernelized correlation filter (KCF) is one of the state-of-the-art object trackers. However, it could not reasonably handle the severe special situations in UAV application during tracking process, especially when targets undergo significant appearance changes due to camera shaking or deformation. In this paper, we proposed a new compounded feature to track the object by combining saliency feature and color features for the conspicuousness of the objects in the videos captured by UAVs. Considering the speed of real-time application, we use a spectrum-based saliency detection method - quaternion type-II DCT image signatures. In addition, severe drifting can be detected and adjusted by the relocation mechanism. Extensive experiments on the UAV tracking sequences show that the proposed method significantly improves KCF, and achieves better performance than other state-of-the-art trackers.

Keywords: Saliency feature · UAV · Object tracking · KCF

1 Introduction

Unmanned aerial vehicle (UAV) is one of the most important tools for obtaining information, which is implemented by multiple tasks both in civilian and military areas such as normal observation, disaster monitoring, and battlefield detection, etc. Among these applications, UAV object tracking is rather challenging due to various factors like illumination change, occlusion, motion blur, and texture variation [3–6]. To this end, the conventional data association and temporal filters always fail due to the fast motion and changing object or background appearances.

In object tracking filed, generative tracking methods always search for best matched regions in successive frames as results. Most of recent generative methods such as [13–15] focus on building a good representation of the target. Because the generative method focuses on the representation of the target itself and ignores the background information, it is prone to drift when the object changes violently or under occlusion.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11259, pp. 115–125, 2018.
https://doi.org/10.1007/978-3-030-03341-5_10

On the other hand, a popular trend of visual tracking research in recent years is the use of discriminative learning methods, such as [7, 8]. Discriminative method can be more robust than generative method because it distinguishes the background and foreground information significantly. However, these methods face such problems as lack of samples and confusion of boundary demarcation. In addition, many existing methods are slow in operation and expensive in computation, which limits their practical applications [9].

Though many attempts have been made to mitigate these tracking problems for general videos, none of them can be perfectly applied to UAV videos due to the characteristic of small target caused by the long-distance imaging and movement of both target and background caused by the violent motion of the airframe. Inspired by the visual attention mechanism, our works establish a more robust feature representation by combining the saliency feature with the ordinary feature on the basis of the kernelized correlation filter (KCF).

The rest of the paper is organized as follows. Section 2 introduces our method including details about how saliency feature can be efficiently embedded in the KCF framework and the relocation mechanism used to detect and adjust the tracking failure. Section 3 shows the extensive experiments we designed to evaluate the results of our method in comparison of other state-in-the-art methods. Finally, conclusions are put forward in Sect. 4.

2 Methodology

2.1 Tracking by Correlation Framework

Discriminative trackers [9] are thought more robust than generative trackers because the former regard visual tracking as a problem of binary classification. Utilizing the trained classifier, trackers can distinguish foreground and background and estimate target position among vast candidates.

In 2010, correlation filter was introduced into tracking to raise efficiency when facing a large number of samples [1]. Using the circulant structure and ridge regression, Henriques et al. [7] simplify the training and testing process. The purpose of ridge regression training is to find a filter ω which can minimize the squared error over training samples x and their label y. The training problem is formulated as

$$\min_{\omega} ||\phi\omega - y|| + \lambda||\omega|| \tag{1}$$

where ϕ denotes the mapping of all the circular shifts of the template x. If the mapping ϕ is linear, Eq. 1 could be solved directly using the DFT matrix and the target's position could be located when we get its filter ω.

Then the appealing method being able to work solely with diagonal matrices by Fourier diagonalizing circulant matrices can be used simplify the linear regression in Eq. 2.

$$X = Fdiag(\widehat{x}^{*})F^{H} \tag{2}$$

From the Eq. 2, we have

$$\alpha = (K + \lambda I)^{-1} y \tag{3}$$

with the gaussian kernel k, the matrix $K_{i,j} = k(P^{i}x, P^{j}x)$ is circulant. Using the quick calculation of the Eq. 3, the FFT of α is calculated by:

$$\hat{\alpha} = \frac{\hat{y}}{\hat{k}^{xx} + \lambda} \tag{4}$$

where the notation "^" represents discrete Fourier operator, k^{xx} is the first row of the circulant matrix K, and test patch z are evaluated by:

$$\hat{y} = \hat{k}^{zx} \Theta \hat{\alpha} \tag{5}$$

where Θ is the element-wise product. $\hat{y}$ is the output response for all the testing patches in frequency domain. Then we have

$$response = \max(ifft(\hat{y})) \tag{6}$$

The target position is the one with the maximal value among response calculated by Eq. 6. Finally, a new model is trained at the new position, and the obtained values of α and x are linearly interpolated with the ones from the previous frame, to provide the tracker with some memory.

Despite of the huge success of correlation tracking, these trackers could not represent the small target of UAV application robustly since the ground targets always "weak" and "small", which showing low representation. Therefore, how to find a more robust feature and solve the critical drift problem in the UAV application should be a challenging problem.

2.2 Tracking by Compounded Features

Due to the weakly representation of the small target, we can use a saliency method, which is thought to be one of the guiding mechanisms of human visual tracking, to get close to the approximate target position. There have been some works using visual saliency for object tracking. For example, [24] used multiple trackers to modulate the attention distribution and [25] embedded three kinds of attention to help locate the object, while they are both time-consuming. In this paper, we propose a simple and efficient feature model based on the image saliency, combing the saliency feature and the ordinary feature, which can be represented as:

$$C_o(z) = \gamma C_f(z) + (1 - \gamma) C_s(z) \tag{7}$$

in which $C_o(z)$ represents the compounded feature, $C_f(z)$ represents the ordinary feature such as gray and HoG, $C_s(z)$ represents the saliency feature, γ represents the weight coefficient, z means the candidate target. This combination of two parts provides a more robust feature representation. And we will introduce the two parts in detail below respectively.

Ordinary Feature Representation. Low-level features, such as intensity, color, orientation, motion, etc., are widely used. It is because that fast guidance of image key content is usually believed to be driven by bottom-up features. For simplicity and speed, we omit other cues but intensity. Color cue does not exist in gray-scale image sequences. Besides, the computational cost of color or motion is rather expensive. It is not a good trade-off to trade significant amount of computing time for a little performance improvement, especially when our main task is tracking, in which speed is a key performance metric. Therefore, an intensity map $I(z)$ is used to model appearance distribution.

$$I(z) = (r(z) + g(z) + b(z))/3 \tag{8}$$

in which $r(z), g(z), b(z)$ represents the three color channels, respectively.

Saliency Feature Representation. By investigating different saliency models, we note that computational models, especially spectral models, fit the UAV tracking situation better to the real-time application for its speediness. Therefore, to get the saliency feature representation, we cite the image signature method [22], providing an approach to the figure-ground separation problem using a binary, holistic image descriptor.

$$Imagesignature(x) = \mathrm{sign[DCT(x)]} \tag{9}$$

$$\tilde{x} = IDCT[sign(\bar{x})] \tag{10}$$

$$m = g \cdot [\tilde{x} \circ \tilde{x}] \tag{11}$$

where x represents the image, g is a Gaussian kernel, $\cdot$ denotes the convolution operation and $\circ$ denotes the Hadamard production. It has been verified experimentally that this method efficiently suppresses the background and highlight the foreground.

Then, through transferring the single-channel definition of the DCT signature to quaternion DCT(QDCT) signature, we derive a sparsity map using the QDCT method [23], describing the possible target candidates of image around the target in the last frame.

$$m_{QDCT}(I_Q) = g \cdot [\tilde{I}_Q \circ \tilde{I}_Q] \tag{12}$$

$$\tilde{I}_Q = IQDCT^L\{sign[QDCT^L(I_Q)]\} \tag{13}$$

where I_Q is the quaternion image, g is a 10×10 Gaussian kernel with σ=2.5. Using the QDCT method, color information can be totally covered in the saliency detection. The mixture of spectral method and low-level features (color in our method only) improves the performance of original image signature method.

$$QDCT^L(p,q) = \alpha_p^M \alpha_q^N \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} u_Q I_Q(m,n) \beta_{p,m}^M \beta_{q,n}^N \tag{14}$$

$$QDCT^R(p,q) = \alpha_p^M \alpha_q^N \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_Q(m,n) \beta_{p,m}^M \beta_{q,n}^N u_Q \tag{15}$$

$$IQDCT^L(m,n) = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \alpha_p^M \alpha_q^N u_Q C_Q(p,q) \beta_{p,q}^M \beta_{m,n}^N \tag{16}$$

$$IQDCT^R(m,n) = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \alpha_p^M \alpha_q^N C_Q(p,q) \beta_{p,q}^M \beta_{m,n}^N u_Q \tag{17}$$

Where u_Q is a unit (pure) quaternion, $u_Q^2 = -1$ that serves as DCT axis, I_Q is the $M \times N$ quaternion matrix. In accordance with the definition of the traditional type-II DCT, we define α, β and u_Q as follows:

$$\alpha_p^M = \begin{cases} \sqrt{1/M} & for \quad p = 0 \\ \sqrt{2/M} & for \quad p \neq 0 \end{cases} \tag{18}$$

$$\beta_{p,m}^M = \cos[\frac{\pi}{M}(m + \frac{1}{2})p] \tag{19}$$

$$u_Q = -\sqrt{1/3}i - \sqrt{1/3}j - \sqrt{1/3}k \tag{20}$$

Figure 1 shows the process of obtaining saliency feature. In this model, the red, blue and green channels are extracted respectively in the candidate patch, occupying three channels of the QDCT (the residual channel is set as void). Then we get the saliency map of each channel. Through QDCT and IQDCT transform, three channel maps are grouped to the final saliency map, which can be combined with the gray feature.

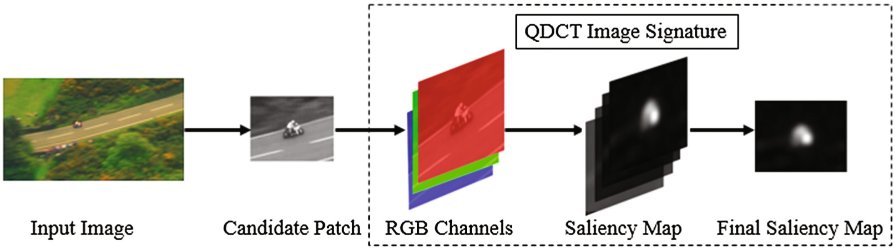

Fig. 1. Process of obtaining saliency feature.

2.3 Tracking by Relocation Mechanism

As shown in the failure detection case in the output constraint transfer theory, the output of the test image is reasonably considered to follow a Gaussian distribution, which is theoretically transferred to be a constraint condition in the Bayesian optimization problem, and successfully used to detect the tracking failure problem. Here, we argue that the Gaussian prior [12] can well help to detect failure by setting $|(y^t_{\max} - \mu^t)/\sigma^t| < T_g$, where μ^t and σ^t are the average and variance of response calculated based on all previous frames in the tracking procedure, $y^t_{\max}$ represents the maximum response, and according to [12], T_g is set as 0.7.

After judging and entering the drift processing mechanism, we get N samples from the target position in the last frame to relocate the target position by a linear stochastic differential equation:

$$p_t^{(n)} = p_{t-1}^{(n)} + A\delta_t^{(n)} \tag{21}$$

where $\delta_t^{(n)}$ is a multivariate Gaussian random variable, t represent t^{th} frame, n represent n^{th} sample and p represents the position, A is the proportionality factor, which can be updated by:

$$A = \frac{P(\delta_t^{(n)}(x))P(\delta_t^{(n)}(y))P(F_t^{(n)})}{1 - P(\delta_t^{(n)}(x))P(\delta_t^{(n)}(y))P(F_t^{(n)})} \tag{22}$$

where x, y represents two directions in a two-dimensional image, $P(\delta_t^{(n)})$ is Gauss probability and $P(F_t^{(n)})$ is the likelihood calculated by Euclidean distance of the feature between the candidate targets and the target. Through Eq. 22, we can refine the candidate location so that obtain a more precise tracking result.

Algorithm 1 Saliency-based Object Tracking Method

```
Initial target bounding box $b_0 = [x_0, y_0, w, h]$
if the frame $n < 20$
repeat
    Crop out the search windows according to $b_{n\text{-}1}$, and extract the saliency-based feature representation.
    Obtain the target position according to KCF algorithm.
    Updating essential parameters of the KCF tracker.
until $n == 20$
end
Compute the mean $\mu$ and variance $\sigma^2$ using all previous frames.
if $n > Num$
repeat
    Crop out the search window and extract the compounded feature representation.
    Obtain the target position according to KCF algorithm.
    if satisfy $|(\mathrm{y}_{\max}^t - \mu^t)/\sigma^t| < \mathrm{T_g}$
    Crop out $N$ samples using Eq. 21.
    Assign different weight to each region using KCF algorithm to obtain the target position.
    Updating essential parameters of KCF tracker.
    Updating $\mu$ and $\sigma^2$
until End of the video sequence.
end
```

3 Experiments

Most visual tracking models have been tested on the commonly used benchmark [10] with many datasets like OTB50 [10], ALOV300++ [21] and so on, which have abundant natural scenes. However, in the field of UAV object tracking, few datasets have been proposed to test the tracking models. In 2016, a new dataset (UAV123) [11] with sequences from an aerial viewpoint was proposed, which contains a total of 123 video sequences and more than 110 K frames, making it the second largest object tracking dataset after ALOV300++. To evaluate the effectiveness and robustness of trackers, we selected 45 challenging sequences from UAV123 dataset, covering typical UAV tracking problems. According to the benchmark [10], each sequence is manually tagged with 11 attributes which represent challenging aspects including illumination variations (IV), scale variations (SV), occlusions (OCC), deformations (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), background clutters (BC), and low resolution (LR). In addition, we compare our tracker with 10 state-of-the–art trackers covering correlation filter-based tracker CT [18], CSK [7], KCF [19], OCT [12], saliency-based tracker SPC [20] and other representative trackers, such as LOT [2], ORIA [17] and DFT [16]. The platform of our experiments is Intel I7 2.7 GZ (4 cores) CPU with 8G RAM.

3.1 Objective Evaluation

Figure 2 shows qualitative results comparing with the other state-of-the-art trackers on challenging sequences. In the first row, our tracker can precisely track the wakeboard, while the conventional KCF tracker fails. The famous OCT tracker could also relocate the target because of the Gaussian prior, although the tracking bounding boxes of the OCT is not as precise as those of ours. While in *uav*, the appearances of the target are weak and small, with severe changing and drifting. Most trackers drift easily due to poor robustness. While our tracker could locate the target accurately because of the robust compounded feature and utilization of relocation method. It is also observed that our proposed tracker works very well in other sequences, e.g., *road* and *truck*. In contrast, all other compared trackers get false or imprecise results in one sequence at least.

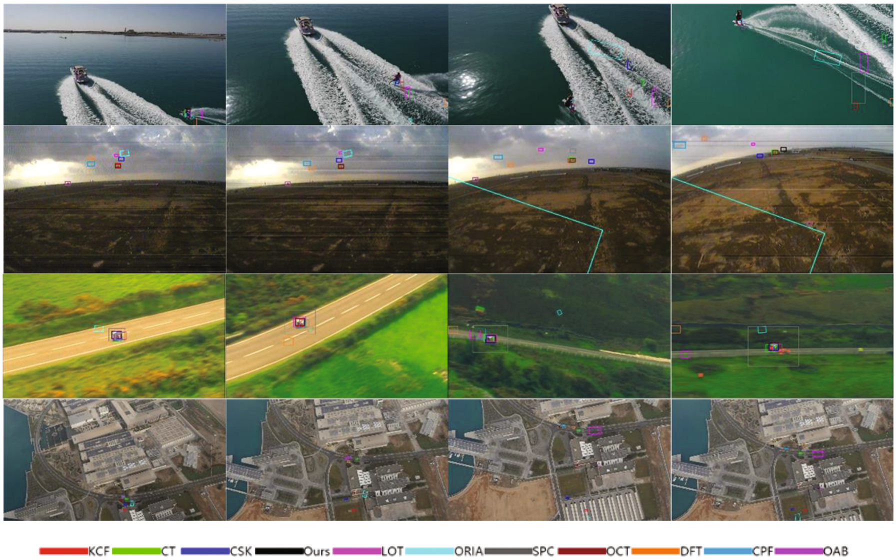

Fig. 2. Illustration of some key frames.

3.2 Subjective Evaluation

In Fig. 3, the overall success and precision plots generated by the benchmark toolbox are reported. These plots report 11 performing trackers in the benchmarks. Our tracker and OCT achieve 90.8% and 84.7% based on the average precision rate when the threshold is set to 20, while the famous KCF and CSK trackers, respectively, achieve 74.1% and 72.4%. In terms of success rate, Ours and OCT, respectively, achieve 56.9% and 55.8%. We also compare with SPC, which presents a saliency prior context model, showing that our tracker achieves a significant performance in terms of precision (18.4% higher) and success rate (10.3% higher). These results confirm that our method performs better than most state-of-the-art trackers.

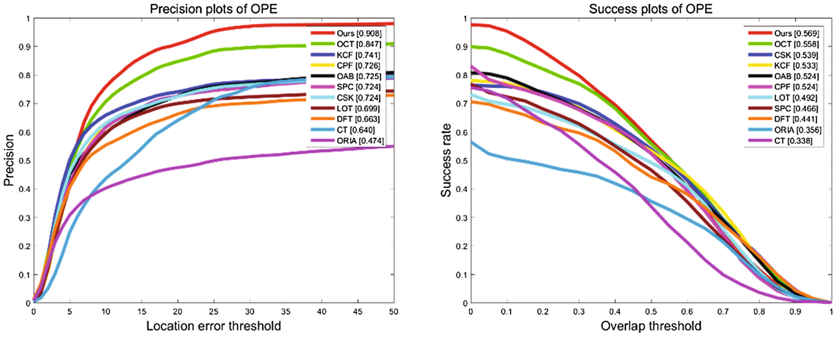

Fig. 3. Success and precision plots.

Results about average success rate for 8 typical attributes are summarized in Table 1. In this table we use bold fonts to indicate the best tracker of different attributes among KCF, OCT and our tracker. Statistic data shows that our method is effective and efficient in front of the challenges including in-plane rotation, out-of-plane rotation, scale variation and fast motion, which occurs frequently in UAV videos. We are pleasantly surprised that our method outperforms the KCF and OCT for low resolution videos, with a success rate of 0.55, meaning our method can perform well although the target is weak and small. Besides, our method hold a tracking speed of 92FPS, so it can realize real-time UAV object tracking.

Table 1. Results for success rate of different attributes

attribute	SV	MB	FM	IPR	OPR	OV	BC	LR
KCF	0.310	**0.689**	0.596	0.336	0.346	0.582	0.585	0.512
OCT	0.378	0.574	0.568	0.396	0.391	**0.587**	**0.621**	0.542
Ours	**0.479**	0.671	**0.614**	**0.468**	**0.482**	0.580	0.591	**0.550**

4 Conclusion

In this paper, we propose a compounded feature and a relocation method to enhance KCF for object tracking in UAV applications. We explore that the incorporation of low-level features and saliency feature could effectively solve the tough problems in UAV tracking, i.e. low resolution and fast motion. Also, we show that the failure judgement and relocation mechanism is appropriate for the weak and small targets in UAV tracking process. In the experimental section, we implement our tracking framework under the compounded feature – the combination of QDCT image signatures and gray feature. It's worth noting that the proposed method could be extended to other features and extensive tracking situations. Extensive experiments of the UAV videos and comparisons on the benchmark show that the proposed method significantly improves the performance of KCF, and achieves a better performance than state-of-the-art tracker.

References

1. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: Computer Vision and Pattern Recognition, pp. 2544–2550 (2010)
2. Avidan, S., Levi, D., Barhillel, A., Oron, S.: Locally orderless tracking. Int. J. Comput. Vision **111**, 213–228 (2015)
3. Yao, R., Shi, Q., Shen, C., Zhang, Y., Hengel, A.V.: Part-based visual tracking with online latent structural learning. In: Computer Vision and Pattern Recognition, pp. 2363–2370, Portland, OR, USA (2013)
4. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: Computer Vision and Pattern Recognition, pp. 798–805, New York, NY, USA (2006)
5. Ross, D.A., Lim, J., Lin, R.-S., Yang, M.-H.: Incremental learning for robust visual tracking. Int. J. Comput. Vis. **77**, 125–141 (2008)
6. Zhuang, B., Lu, H., Xiao, Z., Wang, D.: Visual tracking via discriminative sparse similarity map. IEEE Trans. Image Process. **23**, 1872–1881 (2014)
7. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 702–715. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_50
8. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1409–1422 (2012)
9. Hare, S., Saffari, A., Torr, P.H.: Struck: structured output tracking with kernels. In: International Conference on Computer Vision, pp. 263–270 (2011)
10. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: Computer Vision and Pattern Recognition, pp. 2411–2418 (2013)
11. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 445–461. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_27
12. Zhang, B., Li, Z., Cao, X., Ye, Q., Chen, C., Shen, L., et al.: Output constraint transfer for kernelized correlation filter in tracking. IEEE Trans. Syst. Man Cybernet. Syst. **47**, 693–703 (2017)

13. Arulampalam, M.S., Maskell, S., Gordon, N., et al.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. **50**, 174–188 (2002)
14. X.Q. Zhang, W.M. Hu, S. Maybank, X. Li and M.L. Zhu: Sequential particle swarm optimization for visual tracking. In: Computer Vision and Pattern Recognition, Anchorage, AK, USA, pp. 1–8 (2008)
15. Yang, C., Duraiswami, R., Davis, L.: Efficient mean-shift tracking via a new similarity measure. In: Computer Vision and Pattern Recognition, pp. 176–183 (2005)
16. Learnedmiller, E., Sevillalara, L.: Distribution fields for tracking. In: Computer Vision and Pattern Recognition, pp. 1910–1917 (2012)
17. Wu, Y., Shen, B., Ling, H.: Online robust image alignment via iterative convex optimization. In: Computer Vision and Pattern Recognition, pp. 1808–1814 (2012)
18. Zhang, K., Zhang, L., Yang, M.-H.: Real-time compressive tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 864–877. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_62
19. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **37**, 583–596 (2015)
20. Ma, C., Miao, Z., Zhang, X.P., Li, M.: A saliency prior context model for real-time object tracking. IEEE Trans. Multimedia **19**, 2415–2424 (2017)
21. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: algorithms and benchmark. In: IEEE Trans. Image Process. 1–14 (2015)
22. Hou, X., Harel, J., Koch, C.: Image signature: highlighting sparse salient regions. IEEE Trans. Pattern Anal. Mach. Intell. **34**(1), 194–201 (2012)
23. Schauerte, B., Stiefelhagen, R.: Quaternion-based spectral saliency detection for eye fixation prediction. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 116–129. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_9
24. Choi, J., Chang, H.J., Jeong, J., et al.: Visual tracking using attention-modulated disintegration and integration. In: Computer Vision and Pattern Recognition, pp. 4321–4330 (2016)
25. Wang, Q., Teng, Z., Xing, J., Gao, J., Hu, W., Maybank, S.: Learning attentions: residual attentional siamese network for high performance online visual tracking. In: Computer Vision and Pattern Recognition (2018)