EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Recurrently exploiting co-saliency of target for part-based visual tracking

Song-Chen Han[1,2], Zhao-Huan Zhan[1], Wei Li[1,2*] and Xin-Yan Zhang[3]

Abstract

Visual tracking in condition of occlusion has been a challenging task over years. Recently, part-based algorithms have made great progress in handling occlusion. However, the existing part-based methods neglect different importance between central parts and marginal parts. Besides, scale variation remains a difficulty for part-based tracking. In this paper, we propose a novel part-based tracker to solve the above problems. Specifically, we introduce a visual attention mechanism recurrently exploiting co-saliency of target to guide the sampling of parts, which aims to highlight the importance of salient parts and guarantee the semantic integrity so as to improve the robustness handling occlusion. Considering the drift of prediction caused by mutual influence of parts, we implement the non-maximum suppression operation to reduce the high overlaps between parts, and introduce an effective correlation filter as base tracker. To balance the global distribution and local partiality of parts, appropriate update strategy including scale estimation method inspired by particle filters and correlation filters, Hough-voting scheme for target's center prediction, and principles of part resampling are also fused into the algorithm. The experimental results on VOT 2017 and OTB-50 benchmarks showed that the proposed method is in comparison to the state-of-the-art trackers and good at dealing with occlusion situations particularly.

Keywords: Visual tracking, Part-based, Correlation filter, Occlusion, Attention mechanism

1 Introduction

Visual tracking, which locates a target in a video sequence, is one of the most challenging tasks in computer vision with numerous applications, such as behavior analysis, scene understanding, and video surveillance. Tracking algorithms can be simply categorized into generative methods (e.g., [1–3]) and discriminative methods (e.g., [4, 5, 6]). Generally, tracking ability of trackers can be enhanced by cooperating with effective algorithms used in pattern recognition like rotation invariants [7–9]. With the continuous advancement in machine learning, correlation filters [10] have become the mainstream in target tracking.

Most proposed tracking methods rely on global appearance features. However, occlusion or deformation may dramatically deteriorate the performance of such methods. Recently, the part-based tracker has attracted researchers' attention. The part-based tracking methods divide the target into multiple parts and then use the generative tracker or discriminative tracker to track parts of the target. Most part-based trackers retain the structural layout of the target. For example, Yao et al. [11] propose a part-based appearance model that utilizes the spatial structure of parts to predict new position of the target by minimizing the appearance and deformation costs. Liu et al. [12] propose an oversaturated part-based tracking algorithm based on spatio-temporal context learning, which includes a structural layout constraint and a model updating strategy. Gao et al. [13] propose an end-to-end deep regression model utilizing the advantages of convolutional neural networks, which fully exploits the context information of the parts to preserve the spatial layout structure of the target, and learns the reliability of the parts to emphasize the importance of the parts. Liu et al. [14] propose a structural correlation filter SCF that preserves the overall structure of the target through maintaining similar circular shifts for all parts. With comprehensive progress of tracking algorithms, many part-based tracking methods absorb the ideas from different methods. Bhargava et al. [15] propose a multi-parts and multi-feature tracking method, exploiting strong

* Correspondence: li.wei@scu.edu.cn
[1]School of Aeronautics and Astronautics, Sichuan University, Chengdu, China
[2]Key Laboratory of Air Traffic Control Automation System, Sichuan University, Chengdu, China
Full list of author information is available at the end of the article

features to measure the confidence and reducing the impact of weak features. Niu et al. [16] introduce a background tracker to typical part-based framework, which aims to determine the occurrence of occlusion, so as to adjust the update strategy of the target tracker. Additionally, Wang et al. [17] creatively introduce the imagery ranking-based method into the field of part-based tracking. Li et al. [18] combine the idea of particle filter with correlation filter. Liu et al. [19] take advantage of correlation filter and Bayesian inference and propose a tracker for real-time component tracking. Recently, Johnander et al. [20] propose a deformable filter which is represented as a linear combination of sub-filters, and establish a unified formulation to learn a deformable convolution filter.

Although many novel algorithms are proposed with the advancement of tracking algorithms, most of them are similar in terms of part sampling strategy and base tracker. Firstly, the most popular part sampling strategies are fixed sampling strategies [15] and Gaussian distribution [12, 18] strategies. The former sample the parts by recurrently dividing the target into several fixed-size parts, while the latter is similar to particles that sample parts based on Gaussian distribution. For the base tracker, most part-based tracking algorithms choose the KCF filter [11, 14, 18, 19] because of its high speed. However, the KCF filter cannot adjust the scale variation of the parts, and its features extraction capability is limited.

In this paper, we propose a novel part-based tracking algorithm that recurrently exploiting co-saliency of the target (REC tracker) for visual tracking. Our contributions can be mainly summarized in the following aspects: (1) we propose a part sampling method guided by co-saliency of target which not only highlight the importance of salient parts during tracking process, but also guarantee the semantic integrity. (2) We design an appropriate part update strategy to balance the global distribution and local partiality of the parts. (3) We introduce the handcrafted feature version of efficient convolution operator (ECO-hc) to improve the accuracy of part tracking.

The rest of the paper is organized as follows. Section 2 explains the three main components in our proposed tracker: the part sampling method, the base tracker, and the scale estimation method in detail. And we also provide the process of the proposed tracker. Section 3 shows the experimental results on tracking benchmarks and ablation study of the proposed method. Finally, the conclusions of our research are summarized in Section 4.

2 Methods

In this section, we first illustrate the attention mechanism and apply the target's saliency to the part sampling. We then introduce the ECOhc tracker, a correlation filter with powerful ability in feature extraction, into the part-based tracking framework. At last, we analyze the scale estimation of the existing algorithms and propose a new scale estimation method.

2.1 Part sampling method

In order to address the problems of semantic fragmentation and background interference in the current part-based tracking algorithms, we use the most attractive part of the target to guide the part sampling and therefore improve the representation of the parts in this study. Specifically, we follow the principle of visual stimulation in a single picture, which can identify the most salient objects from the background. As a pre-processing operation before part sampling, the saliency detection should meet requirements of efficient computation. We employ a cluster-based saliency detection method [21], which uses the contrast and spatial cues of a single image to construct a saliency map of the target area of each frame, and we then sample the part based on the saliency map.

2.1.1 Saliency map

We apply the k-means clustering method to divide the pixels of the image $\mathbf{I}$ into K clusters and then compute the contrast cues and spatial cues of each cluster. Then, the two cues are merged into a saliency map.

As a traditional method for measuring the uniqueness of visual features, contrast cue is widely used in the saliency detection of single image. In our study, contrast cue measures the difference between features and can be expressed as follows:

$$w^c(k) = \sum_{i=1, i\neq k}^{K} \left(\frac{n^i}{N} \left\| u^k - u^i \right\|_{l_2} \right) \tag{1}$$

where the superscript "c" refers to contrast cue, and $w^c(k)$ denotes the salient score of contrast cue on cluster C^k, u^k represents the center of cluster C^k, the l_2-norm is used to calculate the distance between features in the feature space, n^i denotes the number of pixels of cluster C^i, and N is the total number of pixels of the image. It can be seen from Eq. (1) that the larger the cluster C^i, the greater influence it plays to contrast cue of cluster C^k.

Contrast cue assigns higher salient scores to the minor clusters. However, it is ineffective for handling the complex background. Thus, we introduce spatial cue. Similar to the cosine window operation in the correlation filter, the spatial cue suppresses the saliency of the region far from the center of image based on the assumption that the central region of image is more significant than other regions. Spatial cue can be described as follows:

$$w^s(k) = \frac{1}{n^k}\sum_{i=1}^{N}\left\{G\left(\left\|t^i - o\right\|_{l_2}^2 | 0, \sigma^2\right) \cdot \delta\left[b\left(x^i\right) - C^k\right]\right\} \tag{2}$$

where the superscript "s" refers to spatial cue, $w^s(k)$ is the salient score of spatial cue on cluster C^k, and t^i denotes the normalized location of the pixel x^i in the image $\mathbf{I}$. The normalization coefficient n^k represents the number of pixels of cluster C^k, and the Gaussian kernel $G(\cdot)$ is used to calculate the Euclidean distance between the normalized pixel t^i and the image center o, and the variance σ^2 is the normalized radius of the image. $\delta(\cdot)$ is the Kronecker delta function, and function b associates the pixel x^i with the cluster index b(x^i).

The two cues are merged by element-wise multiplication operations. The saliency probability of each cluster is calculated using the following formula to obtain the cluster-level significance value:

$$p\left(C^k\right) = w^c(k) \circ w^s(k) \tag{3}$$

In order to get the pixel-level significance value, the relation between pixels and their clusters should be established. For each pixel in the cluster, its salient likelihood follows a Gaussian distribution:

$$p\left(x|C^k\right) = G\left(\left\|v_x, u^k\right\|_{l_2} | 0, \sigma_k^2\right) \tag{4}$$

where v_x denotes the feature vector of pixel x and σ_k represents the variance of the clusterC^k. The pixel-level feature map can be described as the sum of all cluster saliency values:

$$p(x) = \sum_{k=1}^{K} p\left(x|C^k\right) p\left(C^k\right) \tag{5}$$

According to (5), the saliency feature map that reflects the saliency distribution of the target area is obtained. The cluster-based saliency detection method not only calculates the pixel-level saliency value, but also performs the segmentation operation of different regions of the target; it is plausible to use saliency map to guide the part sampling.

2.1.2 Part sampling

With a saliency feature map, the part sampling can be easily performed on the target. Similar to most of the part-based algorithms, the parts mentioned in this paper are in the form of rectangular bounding boxes. The specific size of the parts are 0.7 times of target size. Here, the ith rectangular bounding box can be described as $r^i(x_m^i, w^i, h^i, s^i)$, where x_m^i denotes the pixel in the rectangular bounding box, $m \in \{1, 2, \ldots, w^i \times h^i\}$ is the pixel index, and w^i and h^i represent the width and height of the rectangular bounding box, respectively. s^i represents the saliency score of the rectangular bounding box, which is defined as follows:

$$s^i = \sum_{m=1}^{w^i \times h^i} p\left({x_m}^i\right) \tag{6}$$

Rectangular bounding boxes with low saliency scores are eliminated by implementing the non-maximum suppression (NMS) operation: we first sort rectangular bounding boxes according to saliency scores, then calculate the intersection between rectangular bounding boxes according to the sorting result. The rectangular bounding boxes with the intersection ratio greater than a certain threshold θ are eliminated. The calculation of the intersection ratio θ can be expressed by the following formula:

$$\theta = \frac{r^i \cap r^j}{r^i \cup r^j} \tag{7}$$

Although exhaustive computation is implemented in sampling, the computational cost is still limited because it is only carried out in the target area. Since the NMS operation eliminates redundant rectangular bounding boxes, the rectangular bounding boxes are widely distributed over the salient areas and cover the entire target area, which balances the emphasis and uniformity. We use an example to illustrate it as shown in the Fig. 1. Figure 1a shows the input image, and Fig. 1b shows the co-saliency map which we sample parts based on. Apparently, the sampling results of the Gaussian sampling method (Fig. 1c) contain too much background information and ruin the semantic integrity of the target. However, by the proposed part sampling method (Fig. 1d), the areas which attract visual attention, for instance the shirt, are sampled as a whole.

2.2 ECOhc tracker

In this paper, we choose ECOhc tracker, a pioneering correlation filter, as base tracker. The principle of correlation filter is to obtain correlation peak in target's center while depressing the response of the background. Compared with other correlation filters, the ECOhc has compact structure, a representative samples fusion model, and an adaptive update strategy, which ensures the high accuracy in tracking. During the tracking, the detection scores of the target can be described as follows:

$$\mathbf{S} = \mathbf{Pf} * \mathbf{J} \tag{8}$$

where $\mathbf{S}$ is the detection scores, $\mathbf{J}$ represents the interpolated feature vector, and $\mathbf{f}$ is a set of filters. $\mathbf{P}$ is a matrix learned in the first frame, with the aim of reducing feature dimensions. The operator "$*$" denotes the convolution

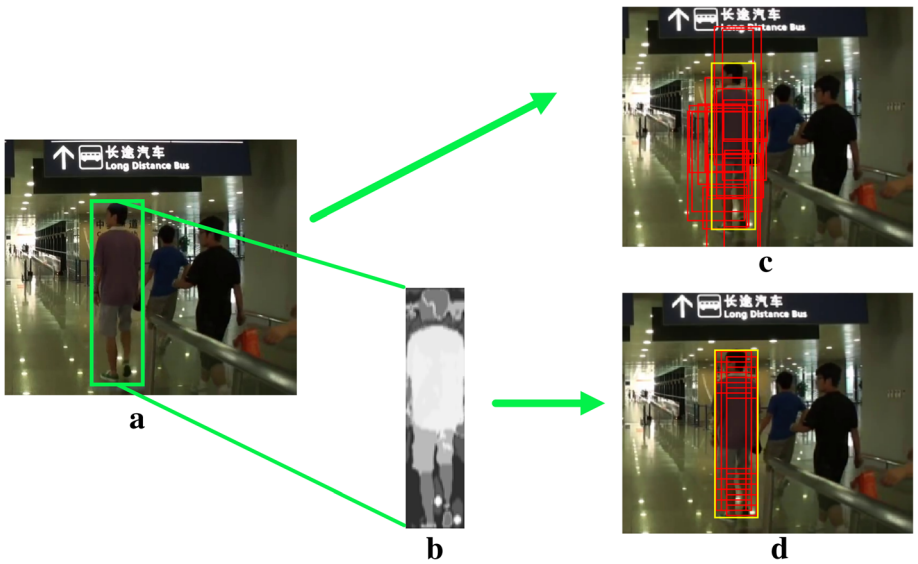

Fig. 1 A comparison of different sampling strategy. **a** Input image. **b** Co-saliency map. **c** Gaussian sampling method. **d** The proposed method. The sampling results of the Gaussian sampling method contain too much background information and ruin the semantic integrity of the target, while the proposed part sampling method samples the salient areas which attract visual attention

operation, and more details are referred to [4]. The filter is learned by minimizing the following objective in the Fourier domain:

$$\arg\min \mathbf{f} = \left\| \hat{\mathbf{z}}^{\mathrm{T}} \mathbf{P}\, \hat{\mathbf{f}} - \hat{\mathbf{y}} \right\|_{l_2}^2 + \sum_{q=1}^{Q} \left\| \hat{\kappa} * \hat{f}^{q} \right\|_{l_2}^2 + \beta \|\mathbf{P}\|_F^2 \tag{9}$$

where $\mathbf{z}^{\wedge}$ is interpolated feature map, $\mathbf{y}^{\wedge}$ is the expected detection score, $\|\cdot\|_{l_2}^2$ means the square of l_2-norm, $\|\cdot\|_F^2$ represents the Frobenius norm, and $\kappa^{\wedge}$ represents a spatial penalty which mitigates the drawbacks of the periodic assumption. q is the index of the filter and Q denotes the number of filters. In order to limit $\mathbf{P}$, a weight parameter β is added in the equation. Equation (9) is regarded as a nonlinear least squares problem, which can be solved by employing the Gauss-Newton and the Conjugate Gradient method.

2.3 Scale estimation

The scale estimation of the target is one of the difficulties in the part-based visual tracking algorithm. On the one hand, since scale variation of parts is almost unrelated to the change of the target's overall scale, it is hard to determine the optimal size of the target by scale variation of base trackers. On the other hand, due to the small number of parts, the scale estimation methods, which are based on dense sampling, are unable to accurately determine the scale variation of the target. In this paper, we design a new mechanism of scale estimation for part-based tracking framework, which is inspired by the scale variation methods of particle filter and correlation filter. The scale variation of particle filter reflects the relationship between target size and motion of parts. And the method of correlation filter constrains the drastic change of the scale which usually happen among particle filters. To the best of our knowledge, it is the first time that we introduce the concept of overlap between the parts and the target into the scale estimation, and we fully consider and introduce the overlaps between the parts and target into the scale estimation. Specifically, 33 scale gradients are set, and the displacement of each part between adjacent frames is calculated. Then we compute the Euclidean distance of the displacement and the scale gradient is computed, which aims to select the scale with the shortest distance as candidate scale of target.

To illustrate the method, we design an objective function for scale estimation. The center coordinates of the target in the tth frame is μ_c^t, the center coordinates of the ith part is μ_i^t, and the number of parts is represented by l. The objective function of scale estimation can be described as follows:

$$\mathrm{L}(\mathrm{v}) = \min\left(\frac{1}{l} \sum_{i=1}^{l} \left(\frac{\mu_c^t - \mu_i^t}{\mu_c^{t-1} - \mu_i^{t-1}} \right)^2 - \lambda^{\nu} \right)^2 \tag{10}$$

where the λ means the gradient base number, which is set to 1.02 after a series of experiments, and $\nu \in \{-16, -15, \ldots, 16\}$ represents the index of λ. The aim of the objective

function is to obtain the candidate scale λ^{v_a}. In order to reduce the influence of the part tracking error in the process of scale estimation, a judgment coefficient $\eta \in \{0, 1\}$ is introduced as follows:

$$\begin{cases} \eta = 1, \phi \geq 0.3 \\ \eta = 0, \phi < 0.3 \end{cases} \tag{11}$$

if the intersection ratio ϕ of parts' total area and target area is less than 0.3, $\eta = 0$, otherwise $\eta = 1$. Let S^t denotes the scale of target in tth frame. Thus, the final scale estimation is as follows:

$$S^t = S^{t-1} \cdot \lambda^{v_a \cdot \eta} \tag{12}$$

2.4 REC algorithm

In this section, we propose the process of the proposed algorithm. The tracker model $\mathbf{M}$ is composed of multiple part tracking models: $\mathbf{M}^t = \{H_1^t, H_2^t, ..., H_l^t\}$. Similar to the RPT tracker [18], we use the Hough voting scheme to predict the center coordinates of target. However, we do not classify parts as positive parts or negative parts, because the algorithm does not use the Monte Carlo framework to distribute parts based on Gaussian distribution. Thus, in our proposed algorithm, all the parts can be considered as positive samples. The center coordinates of target can be predicted by using the Hough voting scheme:

$$\hat{p}^t = \sum_{i=1}^{l} \omega_i^t H_i^t \tag{13}$$

where the weight ω is determined by the confidence of each part. Specifically, the normalized peak-to-sidelobe ratio (PSR) of each part is used as the weight. It can be seen from Eq. (13) that the parts with higher response value has a greater influence on the target state evaluation, which corresponds to the visual principle that different parts of the target tracking have different importance.

After predicting the target center, the target's scale S^t can be estimated by Eq. (12), and finally, the state of the target can be obtained as follows:

$$\mathbf{T}_{\text{target}}^t = \left(\hat{p}^t, S^t\right) \tag{14}$$

With the iteration of part tracking, the tracking error is also accumulated. In order to reduce the error, it is necessary to resample parts. Traditionally, the algorithms resample all of parts, which causes expensive cost of computation [22]. In our algorithm, we only resample certain parts based on the saliency of target. In addition, we have to avoid the overlaps between parts, which may lead to the drift of the target center. In this paper, the criteria for judging the parts that need to be resampled can be summarized as the following:

(a) Low confidence. As mentioned before, parts confidence is measured by evaluating their PSR values. A low confidence means that the part's current tracking result is unreliable, and there is a high probability of tracking errors, so the part should be resampled. We should discard parts with low confidence and add with high confidence new parts.
(b) Far center. During the tracking, some parts may gradually move away from the target center, which may seriously affect the prediction of the target center. Therefore, we identify a part with the center 1.5 times of the target size away from the target center as the far center, which should be resampled.
(c) High overlap. Our resampling strategy is also based on the saliency of the target. However, during the resampling, if an added new part highly overlaps with the existing parts, it affects the accuracy of the tracking through Hough Voting scheme. Thus, when the intersection ratio between a new part and an existing part by more than 50%, the new part will not be added. In summary, the process of the entire algorithm in this paper is shown in Algorithm 1.

Algorithm 1 REC:

Require:

The model $\boldsymbol{M}^{t-1}$ and new arrived image $\boldsymbol{I}^t$

Ensure:

The updated Model $\boldsymbol{M}$ for tracked target;

The new target state, $\boldsymbol{T}_{\text{target}}^t$.

1: **for** every H^{t-1} in $\boldsymbol{M}^{t-1}$ **do**

2: Track H_i^{t-1} with the base tracker f_i^{t-1} in $\boldsymbol{I}^t$.

3: **end for**

4: Vote target's position $\hat{p}^t$ according to Equation 13

5: Estimate target's scale S^t according to Equation 12

6: Reset parts according to (a), (b), (c)

7: Get target state $\boldsymbol{T}_{\text{target}}^t$ according to Equation 14

8: **return** updated $\boldsymbol{M}^t$ and $\boldsymbol{T}_{\text{target}}^t$

3 Results and discussion

In this section, we perform several experiments to testify the effectiveness of our proposed algorithm. We first evaluate the REC tracker on the VOT 2017 and OTB-50 benchmarks. Then we select some typical sequences

from public datasets to demonstrate the tracker's ability to counter occlusion. In addition, we analyze the design of proposed algorithm.

3.1 Implementation details

The experiments are run in the Matlab R2017b, and the hardware environment includes an Intel i5-8400 2.80 GHz CPU, 16 GB RAM. We apply the HOG and color names (CN) as feature representations. All the experiments are carried out using the following fixed parameters, which are set after several experiments. More details about the setting of parameters are described in Section 3.5. For part sampling, the size of part is 0.7 times of the target size, the intersection threshold of the part sampling θ is set to 0.4. As to saliency detection, the input image $\mathbf{I}$ is resized to 150×150, and the cluster number K equals to 8. For scale estimation, the gradient base number λ is set to 1.02 after several experiments, while the intersection threshold ϕ is set to 0.3. For the ECOhc tracker, we use default parameter configuration. In these experiments, the proposed algorithm runs around 14 frames per second.

3.2 VOT 2017 results

3.2.1 Datasets and evaluation metrics

The VOT 2017 dataset consists of 60 video sequences, and each sequence is per-frame annotated. The overall tracking performance is evaluated in terms of expected average overlap (EAO) which takes into account both accuracy and robustness. The larger the value of EAO is, the better the performance of the tracking algorithm will be. The details of calculation of EAO are referred to [23].

3.2.2 Baseline methods

Similar to [24], we choose representative baselines as experimental comparisons from the following perspectives:

1. The recent state-of-the-art tracking baselines: DSST (BMVC 2014) [25], KCF (PAMI 2015) [26], ANT (WACV 2016) [27], ECOhc (CVPR 2017) [4], MEEM (ECCV 2014) [28], Staple (CVPR 2016) [29], ASMS (PRL 2014) [1], SRDCF (ICCV 2015) [30], and Struck (ICCV 2011) [5]
2. The recent state-of-the-art part-based baselines: DPRF [31], DPT (TCYB 2018) [24], CMT (CVPR 2015) [32], LGT (PAMI 2013) [33], FoT (2014) [34], and CGS [31]
3. Deep feature-based baselines: SiamFC (ECCV 2016) [35], GMD [31], GMDnetN [31], and FSTC [31]

3.2.3 Comparison with state-of-the-art

Table 1 shows the performance evaluation results of different state-of-the-art trackers. The best result is highlighted with italic style. Compared with the part-based trackers, the proposed REC method performs well against the DPT [24] (by 27%), DPRF [31] (by 76%), CMT [32] (by 105%), LGT [33] (by 40%), FoT [34] (by 55%), and CGS [31] (by 44%). It is obvious that the REC has excellent competitiveness compared to the existing part-based trackers. The proposed method is also better than algorithms based on deep features such as SiamFC [35] (by 7%), GMD [31] (by 55%), GMDnetN [31] (by 27%), and FSTC [31] (by 7%), which prove that the tracking frameworks based on hand-crafted feature still have great potential. The overall evaluation of 20 trackers is shown in Fig. 2a, b. With the powerful strategy and modeling ability well-designed for global object tracking, the ECOhc tracker gets the best result in the overall performance. Noticed that the proposed method underperforms the ECOhc tracker in the overall performance, but outperforms the ECOhc tracker under occlusion situation, which is further explained in Section 3.4. It can be seen from the experimental results that the REC algorithm not only achieves the state-of-the-art effect among the part-based trackers, but also is superior to most global tracking algorithms.

3.3 OTB-50 results

3.3.1 Datasets and evaluation metrics

The OTB-50 benchmark [36] consists of 50 video sequences, and 25% of sequences of OTB benchmark are gray sequences. The overall tracking performance is evaluated in terms of precision plots and success plots. Specifically, the precision plots measure the Euclidean distance between the predicted center and the ground truth center, while the success plots measure the overlap between the predicted bounding box and the ground truth bounding box. The larger the area under the precision and success curves are, the better the performance of the tracking algorithm will be. In this paper, we perform one-pass evaluation (OPE), which means that only an initial state of target will be given at the first frame.

3.3.2 Baseline methods

Similar to [36], we evaluate our method by comparing with 29 trackers whose original source codes are publicly available.

3.3.3 Comparison with state-of-the-art

The overall performance is shown in Fig. 3a, b. Our method is ranked top both in precision plots and success plots. Note that the proposed method exceeds the performance of the second-best tracker by over 19% in precision plots at representative location error threshold of 20 pixels. When the overlap threshold is from 0 to 0.5, the proposed REC method achieves better than the

Table 1 Comparison with state-of-the-art on the VOT 2017 dataset based on EAO

Method	DPT	KCF	SRDCF	SiamFC	ECOhc	DSST	DPRF	ANT	CMT	LGT
EAO	0.158	0.135	0.119	0.188	*0.238*	0.079	0.114	0.168	0.098	0.144
Method	Struck	MEEM	Staple	ASMS	FSTC	GMDnetN	GMD	FoT	CGS	REC
EAO	0.097	0.193	0.169	0.169	0.188	0.158	0.130	0.130	0.140	*0.201*

second-best tracker in success plots of OPE. In summary, the evaluation illustrate that our method is dependable in target's center prediction and target's scale adaptation.

3.4 Performance under occlusion

3.4.1 Datasets and evaluation metrics

The VOT 2017 challenge organizers annotate per frame with attributes and construct a subset of frames that target is partially or completely occluded. The OTB-50 benchmark also constructs a subset containing 29 sequences with attributes "occlusion." These subsets can be used to analyze the performance of trackers to handle occlusion. In addition, we choose typical occlusion sequences from Temple Color [37], OTB 100, and VOT datasets for experiments to evaluate the proposed method. The EAO scores (VOT 2017 metrics), precision plots, and success plots (OTB-50 metrics) are used as criteria since the evaluation of overall performance and performance under occlusion is proceeded simultaneously. In additional comparison, distance precision (DP) at a threshold of 20 pixels, which equals to the precision value at the location error threshold of 20 pixels in

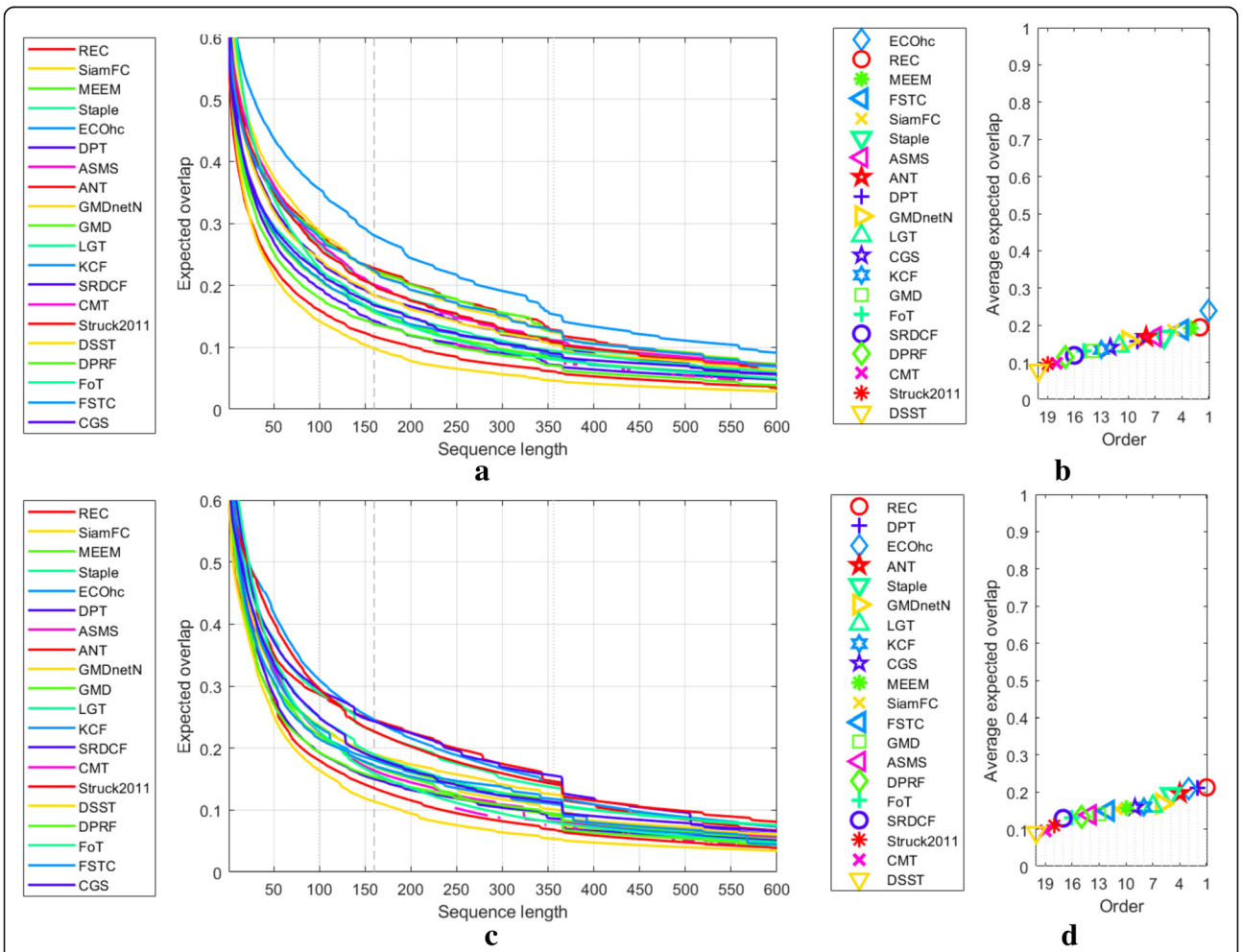

Fig. 2 Evaluation on the VOT2017 dataset. **a** The overall EAO curves. **b** The overall EAO orders. **c** The EAO curves for occlusion attribute. **d** The EAO orders for occlusion attribute

the Precision plots, is used to measure the tracker's performance. The DP value can be seen as a representative precision score [36, 38].

3.4.2 Baseline methods

Because the evaluation of the overall performance and performance under occlusion on the VOT 2017 and OTB-50 benchmark is proceeded at the same time, the baselines used are the same as Section 3.2 and Section 3.3, respectively. In additional comparison, we compare the proposed method with ECOhc and RPT tracker. The reasons that we choose these two trackers as strong baselines are that the proposed method is most like to these two trackers: we apply ECOhc as our base tracker and design the algorithm based on the core concept of the RPT tracker—the fusion of particle filters and correlation filters. However, the proposed method is different to the ECOhc and RPT tracker in many aspects such as part sampling strategy, model update scheme, and scale estimation.

3.4.3 Comparison with state-of-the-art

The evaluation of performance under occlusion on the VOT 2017 benchmark is shown in Fig. 2c, d. It is obvious that the REC provides the best results with the occlusion situation, and another part-based tracker DPT takes the second place; both of them outperform the ECOhc tracker which obtains the highest score in overall performance. The order reflects the advantage of part-based tracker dealing with occlusion situation. As a matter of fact, the part-based framework was proposed to cope with the occlusion situation, so the superior capability dealing with occlusion is the most significant goal throughout the design of part-based algorithms.

The evaluation of performance under occlusion on the OTB-50 benchmark is shown in Fig. 3c, d. It is clear that the proposed method outperforms other 29 trackers. Considering that the proposed method also takes the first place in overall performance evaluation on the OTB-50 benchmark, the experimental results demonstrate the excellent performance of our algorithm.

Additionally, a comparison is made between the proposed algorithm and state-of-the-art trackers, including ECOhc and RPT. The experimental results are listed in Table 2. The best result of each sequence is highligted with italic style. The proposed method outperforms all the competitors. Specifically, the average

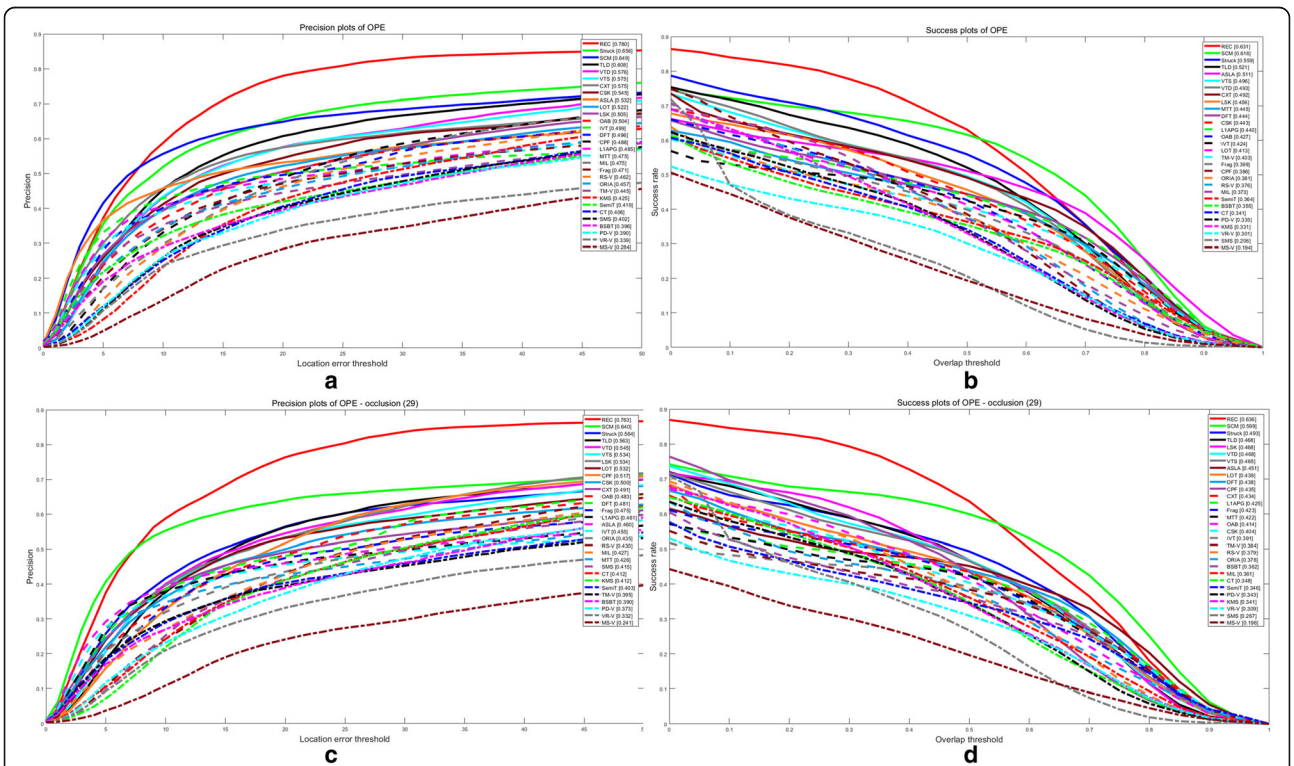

Fig. 3 Precision plots and success plots on the OTB-50 benchmark. **a** The overall precision plots. **b** The overall success plots. **c** The precision plots under occlusion situations. **d** The success plots under occlusion situations

Table 2 A comparison on occlusion sequences in distance precision at a threshold of 20 pixels

Sequence	REC	ECOhc	RPT
Tiger	*0.93*	0.88	0.81
Basketball	*0.99*	0.98	0.92
Shaking	0.97	0.96	*0.99*
Bolt1	*1.00*	0.01	0.02
Godfather	0.71	0.49	*1.00*
Bolt2	*1.00*	0.01	*1.00*
Handball1	*1.00*	0.47	0.64
Road	*1.00*	0.71	0.99
Sheep	*0.90*	0.35	0.62
Marching	*0.98*	0.89	0.90
Airport_ce	0.45	*0.88*	0.39
Birds1	*0.98*	0.62	0.01
Average	*0.91*	0.60	0.69

precision of REC is 52% higher than ECOhc and 32% higher than RPT. Apparently, part-based methods are good at dealing with occlusion problem, while the REC is better than RPT tracker generally. For qualitative analysis, we visualize the comparison, and the results are shown in Fig. 4.

3.5 Ablation study

In this section, we demonstrate the impact of the design of the proposed tracker by progressively integrating our contributions. The performance is evaluated on the public datasets with related evaluation methodology.

3.5.1 Size of part

The size of part, however, significantly influences the performance of tracking. On the one hand, if the size is too small, the base trackers cannot learn enough information for tracking. On the other hand, if the size is as big as the target, the base trackers may confuse the foreground and background. Many part-based trackers set the part size range from 1/3 to 2/3 of the target size. To demonstrate the impact of part size setting, we evaluate the proposed algorithm with different part size on the VOT 2017 datasets. As shown in Fig. 5a, it can be easily seen that the EAO scores gradually rise with the increase of the part size until it is 0.7 times of target size.

3.5.2 Sampling strategy

The most important contribution of our method is the sampling strategy based on target's saliency. To demonstrate the effect of our sampling strategy, we analyze the impact of different sampling strategies. Specifically, the following strategies are respectively evaluated on the VOT 2017 benchmark: sampling based on Gaussian distribution, sampling based on ITTI saliency model [39], and sampling based on co-saliency. For objective comparison, other parameters of the REC tracker are not modified. The results are shown in Table 3, and the best results are highlighted with italic style. The sampling strategy based on co-saliency outperforms other strategies. It is clear that the saliency detection benefits sampling of parts for part-based tracker, and advanced saliency detection method dramatically improves the overall performance of the tracker.

We also investigate the effect of the NMS operation in part-based tracking. The operation is introduced into the proposed algorithm for reducing the repeatedly

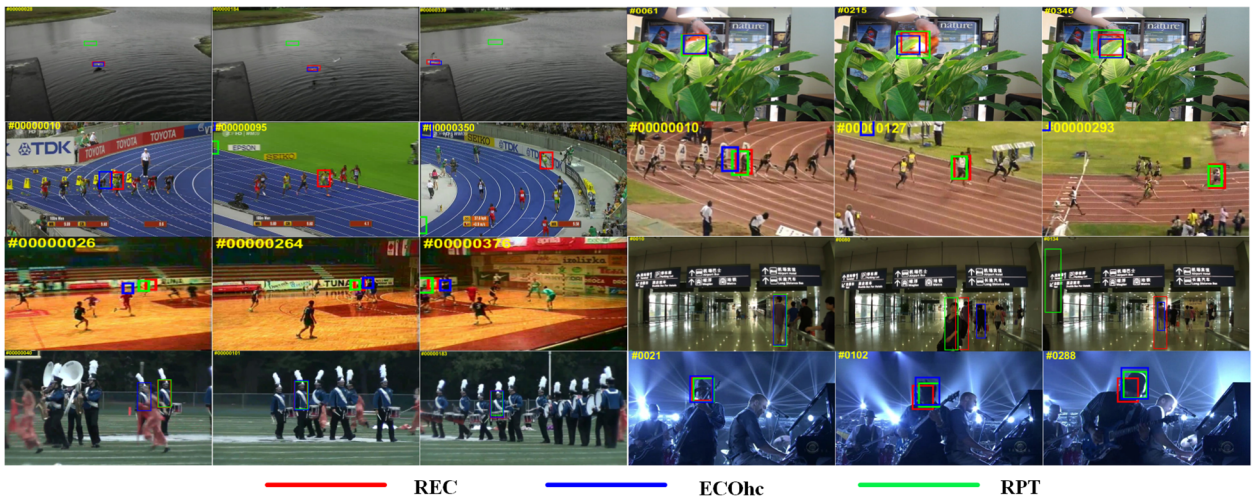

Fig. 4 Tracking results of the trackers in our evaluation on challenging sequences. From left to right and top to down are birds1, tiger1, bolt1, bolt2, handball1, airport, marching, shaking, respectively

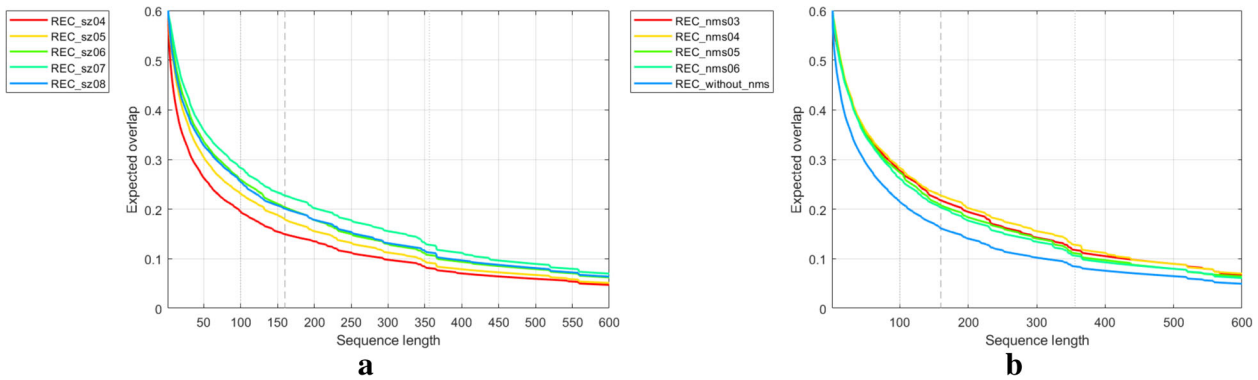

Fig. 5 Performance of REC variants in terms of EAO curves on the VOT 2017 dataset. The number denotes the part size or value of the NMS threshold θ. For example, "REC_sz04" means the REC tracker with part size which is equal to 0.4 times of the target size, and "REC_nms04" means the value of θ is 0.4. **a** Performance of different part sizes. **b** Performance of different θ

sampling on same parts of the target. The results of comparison are displayed in Fig. 5b. Among the REC variants that with different NMS threshold θ, the best result is achieved when the θ is 0.4, which EAO value is 5% higher than the second-best variant and 44% higher than the REC without NMS operation. It is obvious that the NMS operation significantly enhances the performance of the tracker; however, the value of θ should be carefully filtrated since both the low overlaps and high overlaps between parts could deteriorate the performance of tracker.

3.5.3 Scale estimation

To demonstrate the effect of scale estimation method used in this paper, we perform several experiments on OTB-50 benchmark. We investigate the impact of λ, the gradient base number of scale estimation, which decides the degree of size change. The higher the value of λ is, the greater the target size changes. The results of comparison are displayed in Fig. 6. Noticed that when the λ equals 1, it also represents the tracker without scale estimation. We first evaluate two variants of REC: one with scale estimation and the other one without scale estimation. The results prove that the scale estimation method improves the performance of REC. The REC with scale estimation has better performance than the REC without scale estimation. Furthermore, it is interesting that 1.02 is the most suitable value for λ, which is the same as [25] by coincidence. When the value of λ is 1.02, the scale estimation method not only appropriately adapts the scale change but also restricts the divergence of the size change. Hence, we choose 1.02 as the gradient base number of scale estimation.

Table 3 Analysis of sampling strategy on the VOT 2017 dataset

Strategy	EAO	fps
REC_Co-saliency	*0.201*	*14.1*
REC_ITTI	0.186	16.8
REC_Gaussian	0.151	16.3

4 Conclusions

In this paper, we propose a novel part-based tracker named REC for robust visual tracking. The proposed tracker recurrently exploits the co-saliency distribution of the target area to guide part sampling and employs efficient correlation filter called ECOhc to track those parts. During the sampling, we also consider the overlaps between the parts and suppress the inference to avoid the drift of prediction. To guarantee the part reliability, we propose an appropriate updating strategy. Additionally, we also combine the idea of scale variation of particle filter and correlation filter and propose a new scale estimation method. The REC tracker outperforms most of the state-of-the-art part-based trackers as well as state-of-the-art trackers based on global information of target. The REC tracker is evaluated on highly challenging benchmarks by comparing with 19 trackers on VOT 2017 benchmark, 29 trackers on OTB-50 benchmark, and 2 representative trackers on the subset of occlusion sequences. As a result, the experimental results on the public datasets show that the proposed tracker not only reaches the state-of-the-art level in the overall performance, but also outperforms all competitors under occlusion situation. We also investigate factors related to the performance of part-based tracker such

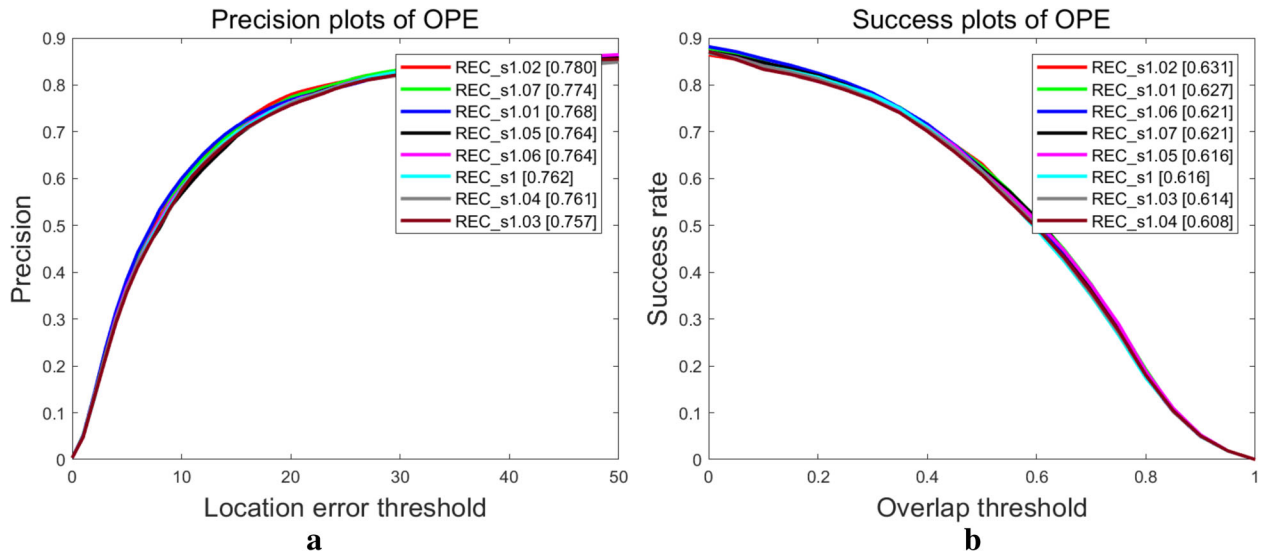

Fig. 6 a, **b** Performance of REC variants on the OTB-50 benchmark. The number at the end of the tracker's name in the legend denotes the value of the gradient base number of scale estimation λ

as size of part, part sampling strategy, and scale estimation, which could be useful in optimizing hyper-parameters of part-based tracking algorithms. In the future, we will improve the tracker by taking the advantage of deep features and advanced attention mechanism.

Abbreviations

EAO: Expected average overlap; NMS: Non-maximum suppression; PSR: Peak-to-sidelobe ratio

Acknowledgements

The authors would like to thank the National Key R&D Program of China No. 2018YFC0809500 and National Natural Science Foundation of China No. 61403065 for the financial assistance, as well as laboratory equipment provided by the school of Aeronautics and Astronautics, Sichuan University.

Funding

This work was supported by the National Key R&D Program of China No. 2018YFC0809500, the National Natural Science Foundation of China under Grants No. 61403065, and the Fundamental Research Funds for the Central Universities No. YJ201450.

Availability of data and materials

The visual tracking data is obtained from the OTB-50 benchmark, VOT 2017 dataset, and Temple Color dataset provided in [31, 36], respectively.

Authors' contributions

SCH conceived of the study and supervised the work and helped to draft the manuscript. ZHZ made the main contributions to the conception and tracking algorithm's design, as well as drafting the article. WL provided significant revising for important intellectual content and gave the final approval of the current version to be submitted. XYZ provided the technical advices and checked the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]School of Aeronautics and Astronautics, Sichuan University, Chengdu, China. [2]Key Laboratory of Air Traffic Control Automation System, Sichuan University, Chengdu, China. [3]Key Laboratory of Fundamental Synthetic Vision Graphics & Image Science for National Defense, Sichuan University, Chengdu, China.

Received: 22 August 2018 Accepted: 27 January 2019
Published online: 20 February 2019

References

1. T. Vojir, J. Noskova, J. Matas, Robust scale-adaptive mean-shift for tracking. Pattern Recogn. Lett. **49**, 250–258 (2014) https://doi.org/10.1016/j.patrec.2014.03.025
2. D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5), 564–577 (2003) https://doi.org/10.1109/TPAMI.2003.1195991
3. K. Nummiaro, E. Koller-Meier, L. Van Gool, An adaptive color-based particle filter. Image Vis. Comput. **21**(1), 99–110 (2003) https://doi.org/10.1016/S0262-8856(02)00129-4
4. M Danelljan, G Bhat, FS Khan, M Felsberg, in IEEE CVPR. ECO: efficient convolution operators for tracking, 1(2), 3 (2017) doi: https://doi.org/10.1109/CVPR.2017.733
5. S. Hare, A. Saffari, P.H. Torr, in *IEEE IeCCV*. Struck: structured output tracking with kernels (2011), pp. 263–270 https://doi.org/10.1109/TPAMI.2015.2509974
6. Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1409 (2012) https://doi.org/10.1109/TPAMI.2011.239
7. M. El Mallahi, A. Mesbah, H. Karmouni, A. El Affar, A. Tahiri, H. Qjidaa, in *ICMCS (ICMCS - International Conference on Multimedia Computing and Systems, IEEE, Marrakech, Morocco, 29 Sept.-1 Oct. 2016)*. Radial Charlier Moment Invariants for 2D Object/Image Recognition (2016), pp. 41–45. https://doi.org/10.1109/ICMCS.2016.7905531.
8. M. El Mallahi, A. Zouhri, H. Qjidaa, Radial Meixner moment invariants for 2D and 3D image recognition. Pattern Recogn. Image Anal. **28**(2), 207–216 (2018) https://doi.org/10.1134/S1054661818020128
9. M.E.I. Mallahi, A. Zouhri, A.E.I. Affar, A. Tahiri, H. Qjidaa, Radial Hahn moment invariants for 2D and 3D image recognition. Int. J. Autom. Comput. **15**(3), 227–289 (2017) https://doi.org/10.1007/s11633-017-1071-1

10. D.S. Bolme, J.R. Beveridge, B.A. Draper, Y.M. Lui, in *IEEE CVPR*. Visual Object Tracking Using Adaptive Correlation Filters (2010), pp. 2544–2550 https://doi.org/10.1109/CVPR.2010.5539960
11. R. Yao, S. Xia, F. Shen, Y. Zhou, Q. Niu, Exploiting spatial structure from parts for adaptive kernelized correlation filter tracker. IEEE Signal Process. Lett. **23**(5), 658–662 (2016) https://doi.org/10.1109/LSP.2016.2545705
12. W. Liu, J. Li, Z. Shi, X. Chen, X. Chen, Oversaturated part-based visual tracking via spatio-temporal context learning. Appl. Opt. **55**(25), 6960–6968 (2016) https://doi.org/10.1364/AO.55.006960
13. J. Gao, T. Zhang, X. Yang, C. Xu, P2T: part-to-target tracking via deep regression learning. IEEE Trans. Image Process. **27**(6), 3074–3086 (2018) https://doi.org/10.1109/TIP.2018.2813166
14. S. Liu, T. Zhang, X. Cao, X. C, in *IEEE CVPR*. Structural correlation filter for robust visual tracking (2016), pp. 4312–4320 https://doi.org/10.1109/CVPR.2016.467
15. N. Bhargava, S. Chaudhuri, in *ICVGIP (ICVGIP - Indian Conference on Computer Vision, Graphics and Image Processing, IIT Guwahati, December 18–22, 2016)*. MPMF: Multi-Part Multi-Feature Based Object Tracking, vol 17 (2016) https://doi.org/10.1145/3009977.3010057
16. X. Niu, X. Fang, Y. Qiao, in *IEEE ASCC (ASCC - Asian Control Conference, Gold Coast, December 17–20, 2017)*. Robust Visual Tracking Via Occlusion Detection Based on Staple Algorithm (2017), pp. 1051–1056 https://doi.org/10.1109/ASCC.2017.8287316
17. J. Wang, C. Fei, L. Zhuang, N. Yu, in *IEEE ICIP (ICIP - International Conference on Image Processing, Phoenix, September 25–28, 2016)*. Part-based multi-graph ranking for visual tracking (2016), pp. 1714–1718. https://doi.org/10.1109/ICIP.2016.7532651
18. Y. Li, J. Zhu, S.C.H. Hoi, in *IEEE CVPR*. Reliable patch trackers: robust visual tracking by exploiting reliable patches (2015), pp. 353–361 https://doi.org/10.1109/CVPR.2015.7298632
19. T. Liu, G. Wang, Q. Yang, in *IEEE CVPR*. Real-time part-based visual tracking via adaptive correlation filters (2015), pp. 4902–4912 https://doi.org/10.1109/CVPR.2015.7299124
20. J. Johnander, M. Danelljan, F.S. Khan, M. Felsberg, in *CAIP (CAIP - International Conference on Computer Analysis of Images and Patterns, Ystad, August 22–24, 2017)*. DCCO: towards deformable continuous convolution operators for visual tracking (2017), pp. 55–67 https://doi.org/10.1007/978-3-319-64689-3_5
21. H. Fu, X. Cao, Z. Tu, Cluster-based co-saliency detection. IEEE Trans. Image Process. **22**(10), 3766–3778 (2013) https://doi.org/10.1109/TIP.2013.2260166
22. P. Sakar, Sequential Monte Carlo methods in practice. Technometrics **45**(1), 106–106 (2003) https://doi.org/10.1198/tech.2003.s23
23. M. Kristan, J. Matas, A. Leonardis, et al., A novel performance evaluation methodology for single-target trackers. IEEE Trans. Pattern Anal. Mach. Intell. **38**(11), 2137–2155 (2016) https://doi.org/10.1109/TPAMI.2016.2516982
24. A. Lukezic, L.C. Zajc, M. Kristan, Deformable parts correlation filters for robust visual tracking. IEEE Trans. Cybern. **48**(6), 1849–1861 (2018) https://doi.org/10.1109/TCYB.2017.2716101
25. M. Danelljan, G. Häger, F. Khan, M. Felsberg, in *BMVC (BMVC - British Machine Vision Conference, Nottingham, September 1–5, 2014)*. Accurate Scale Estimation for Robust Visual Tracking (BMVA Press, 2014) https://doi.org/10.5244/C.28.65
26. J.F. Henriques, C. Rui, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 583–596 (2015) https://doi.org/10.1109/TPAMI.2014.2345390
27. L. Čehovin, A. Leonardis, M. Kristan, in *IEEE WACV*. Robust visual tracking using template anchors (2016), pp. 1–8 https://doi.org/10.1109/WACV.2016.7477570
28. J. Zhang, S. Ma, S. Sclaroff, in *ECCV*. MEEM: robust tracking via multiple experts using entropy minimization (2014), pp. 188–203 https://doi.org/10.1007/978-3-319-10599-4_13
29. L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, P.H. Torr, in *CVPR*. Staple: complementary learners for real-time tracking (2016), pp. 1401–1409 https://doi.org/10.1109/CVPR.2016.156
30. M. Danelljan, G. Hager, F. Shahbaz Khan, M. Felsberg, in *ICCV*. Learning spatially regularized correlation filters for visual tracking (2015), pp. 4310–4318 https://doi.org/10.1109/ICCV.2015.490
31. The VOT 2017 Dataset. http://www.votchallenge.net/vot2017/dataset.html. Accessed 17 Aug 2018.
32. G. Nebehay, R. Pflugfelder, in *CVPR*. Clustering of static-adaptive correspondences for deformable object tracking (2015), pp. 2784–2791 https://doi.org/10.1109/CVPR.2015.7298895
33. L. Cehovin, M. Kristan, A. Leonardis, Robust visual tracking using an adaptive coupled-layer visual model. IEEE Trans. Pattern Anal. Mach. Intell. **35**(4), 941–953 (2013) https://doi.org/10.1109/TPAMI.2012.145
34. T. Vojíř, J. Matas, *In the enhanced flock of trackers. registration and recognition in images and videos* (Springer, Berlin, Heidelberg, 2014), pp. 113–136
35. L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H. Torr, in *ECCV*. Fully-convolutional Siamese networks for object tracking (2016), pp. 850–865 https://doi.org/10.1007/978-3-319-48881-3_56
36. W. Y, J. Lim, M.H. Yang, in *IEEE CVPR*. Online object tracking: a benchmark (2013), pp. 2411–2418 https://doi.org/10.1109/CVPR.2013.312
37. The Temple Color Dataset. http://www.dabi.temple.edu/~hbling/data/TColor-128/TColor-128.html. Accessed 17 Aug 2018
38. B. Babenko, M.H. Yang, S. Belongie, Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **33**(7), 1619–1632 (2011) https://doi.org/10.1109/TPAMI.2010.226
39. L. Itti, C. Koch, E. Niebur, Model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998) https://doi.org/10.1109/34.730558