Aggregation Signature for Small Object Tracking

Chunlei Liu, Wenrui Ding, Jinyu Yang, Vittorio Murino, *Senior Member, IEEE,* Baochang Zhang, *Member IEEE,* Jungong Han, *Member IEEE,* Guodong Guo, *Senior Member, IEEE,*

***Abstract*—Small object tracking becomes an increasingly important task, which however has been largely unexplored in computer vision. The great challenges stem from the facts that: 1) small objects show extreme vague and variable appearances, and 2) they tend to be lost easier as compared to normal-sized ones due to the shaking of lens. In this paper, we propose a novel aggregation signature suitable for small object tracking, especially aiming for the challenge of sudden and large drift. We make three-fold contributions in this work. First, technically, we propose a new descriptor, named aggregation signature, based on saliency, able to represent highly distinctive features for small objects. Second, theoretically, we prove that the proposed signature matches the foreground object more accurately with a high probability. Third, experimentally, the aggregation signature achieves a high performance on multiple datasets, outperforming the state-of-the-art methods by large margins. Moreover, we contribute with two newly collected benchmark datasets, i.e., small90 and small112, for visually small object tracking. The datasets will be available in https://github.com/bczhangbczhang/.**

***Index Terms*—aggregation signature, small object tracking, saliency, image signature**

I. Introduction

WHILE several tracking methods have been developed over the past decade [1], [2], [3], [4], [5], [6], [7] and have been proven to be successful in many applications, such as robotics or video surveillance, tracking small objects in videos still remains a challenging problem, in particular when the complex scenarios and real time constraints are to be considered. In this paper, small objects mean that the targets in images have sizes of less than 1% of the whole image. The challenge of small object tracking mainly roots in two main facts: first, the visual features of small objects are extremely fickle, thus making feature representation difficult; second, sudden and large drift always occurs to small objects in tracking because of the shaking of the lens, compared to the normal-sized objects. The so-called sudden and large drift is that the target distance between two adjacent frames in the image coordinate system is two times larger than the target size.

For a long time, researchers only reported tracking results on common benchmarks using reasonably sized targets, but paid less attention to the small-object tracking problem. Just few existing algorithms related to the small object tracking, while that were designed to enhance the visual features of such a type of targets, with the hope that tracked objects would no longer be lost if robust features were exploited. For instance, the method in [8], [9] integrates both spatial and frequency domain features in order to localize the targets more accurately. Alternatively, the method in [10] tends to enhance the robustness of a tracker by strengthening the feature representations (e.g., target attributes) for the small targets. Recently, Rozumnyi et al. [11] have proposed to deal with fast moving and motion blur problems of the objects, but the performance is unsatisfactory due to low resolution and complex background clutters. Regarding now deep learning methods [12][13] are developed, we think the high-level features seem not to be effective for small objects. Moreover, we doubt that a continuous tracking of small-sized objects can be guaranteed even if robust visual features are exploited, considering the fact that small targets can easily be confused with the noise and clutters in real scenes. In other words, it might be more realistic to allow small objects to get lost during tracking, while investigating a better solution to re-detect them.

The intuition here is about "how human beings recognize the small target when it is lost due to clutter background?" Most likely, humans first look at the salient objects/regions popping up in the scene, and further verify whether one of the salient objects is the target of interest [14]. A few works mimic human being's behavior and involve the saliency information in object tracking. For example, the method in [15] integrates saliency for the representation of context, while [16], [17], [18] incorporate saliency into appearance models in various ways in order to improve the robustness of the tracker. However, as they mostly focus on the target appearances in the image domain, performance is not satisfactory since the appearance is implicitly weak for small objects. Therefore, they might only be reliably applied for tracking normal-sized objects. In this paper, we propose a new saliency online learning framework, termed aggregation signature, and focus on small object tracking. To the best of our knowledge, no saliency-based methods have utilized all context information, including intensity, saturation, saliency and motion information, for small object tracking yet.

Unlike handcrafted image signatures, which are simple yet powerful tools to spatially match the sparse foreground

C. Liu is with School of Electrical and Information Engineering, Beihang University, Beijing, China. E-mail: liuchunlei@buaa.edu.cn

W. Ding is with Unmanned System Research Institute, Beihang University, Beijing, China. E-mail: ding@buaa.edu.cn

J. Yang is with School of Computer Science in the University of Birmingham, British. E-mail: yangjinyu@buaa.edu.cn

V. Murino is with University of Verona, Genoa, Italy, and is also with Pattern Analysis and Computer Vision department at the Istituto Italiano di Tecnologia, Genoa, Italy. E-mail: vittorio.murino@iit.it

B. Zhang is with School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. Baochang Zhang is also with Shenzhen Academy of Aerospace Technology, Shenzhen, China. E-mail: bczhang@buaa.edu.cn

J. Han is with the WMG Data Science Group, University of Warwick, Coventry CV4 7AL, U.K. (e-mail: jungonghan77@gmail.com)

G. Guo is with Institute of Deep Learning, Baidu Research and National Engineering Laboratory for Deep Learning Technology and Application. E-mail: guoguodong01@baidu.com

Wenrui Ding and Baochang Zhang are the corresponding authors.

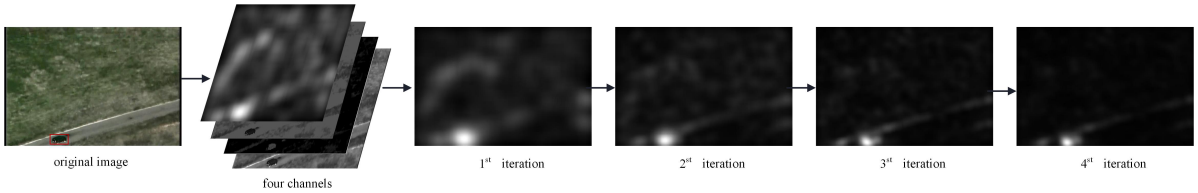

Fig. 1: The aggregation signature results are shown at different iterations, reflecting that the tracked target becomes more salient in the learning process. Our aggregation signature constitutes the first attempt to incorporate the tracked target information into the quaternion discrete cosine transform (QDCT) image signature, whose aggregation capacity is proved in theoretical terms.

objects in an image [19], [20], the explicit advantage of our aggregation signature lies in a learning mechanism exploited to build an adaptive target signature. The result is that it can quickly detect the salient objects even though they are very small, which can further improve the (re-)localization performance of the trackers. We open up a new direction to track small objects by mimicking the human attention mechanism. In particular, the theoretical evidence proves that it is more effective, and that the resulting foreground saliency map from our aggregation signature becomes more consistent with the target appearance along iterations, as shown in Fig. 1. Moreover, the aggregation signature is so generic that it can be integrated into other trackers. In summary, the contributions of this paper include:

(i) The proposed aggregation signature is proved, in the theoretical terms, to be more efficient for sparse foreground detection, makings the tracked target more salient as compared to the background.

(ii) The aggregation signature improves the capacity of accumulating information for the target based on a learning mechanism, whereas the conventional image signatures are handcrafted and more likely prone to fail to adapt to the target.

(iii) New challenging datasets – small90 and small112 – are collected for small object tracking evaluation. The datasets are publicly available for further research development.

II. Aggregation Signature

Image signature is a simple yet powerful tool to spatially match the sparse foreground of an image [19]. By using the sign function of DCT, the resulting handcrafted descriptor can approximately detect salient image regions efficiently. Rather than separating a color image into three channel images and computing image signatures respectively, QDCT [20] can discriminate the relative importance of four components by introducing a quaternion component. In general, both DCT and QDCT based image signatures are handcrafted methods with no involvement of a learning process. Differently, the proposed aggregation signature improves the discriminative capability of QDCT signature via *learning* multi-cue information, in particular the target prior information.

A. Definition of Aggregation Signature

We begin by considering an image x which exhibits the following structure:

$$\mathbf{x} = \mathbf{f} + \mathbf{b}, \quad \mathbf{x}, \mathbf{f}, \mathbf{b} \in \mathbb{R}^N, \tag{1}$$

where $\mathbf{f}$ represents the foreground and $\mathbf{b}$ represents the background. Please refer to Table I for the definitions used throughout the rest of this section. Formally, the aggregation signature (AS) is defined as:

$$\mathrm{AS}\left(\mathbf{x}_Q^i\right) = \mathrm{sign}\left(\mathrm{QDCT}\left(\mathbf{x}_Q^i\right)\right), \tag{2}$$

where $\mathrm{sign}(\cdot)$ is the entrywise sign operator, i represents the i^{th} iteration and Q represents the 4 channels in use. Then, the reconstructed image can be defined as:

$$\bar{\mathbf{x}}_Q^i = \mathrm{IQDCT}\left(\mathrm{AS}\left(\mathbf{x}_Q^i\right)\right), \tag{3}$$

and

$$\mathbf{x}_Q^i = \boldsymbol{\pi} \circ \mathbf{S}^i + \mathbf{I_1} j + \mathbf{I_2} k + \mathbf{I_3} h, \tag{4}$$

where $\mathbf{S}^i = \bar{\mathbf{x}}_Q^{i-1} \circ \widetilde{\bar{\mathbf{x}}}_Q^{i-1}$ [1], $\bar{\mathbf{x}}_Q^i$ represents the reconstructed result in the i^{th} iteration with $\widetilde{\bar{\mathbf{x}}}_Q^i$ as its conjugate form, and $\circ$ represents the element wise product. $\mathbf{I_1}$, $\mathbf{I_2}$, $\mathbf{I_3}$ represent three different channels such as any one channel of RGB, image intensity and image saturation (or motion in tracking). $\boldsymbol{\pi}$ is a two-dimensional prior related to the tracked target, which will be elaborated in Section IV.

B. Foreground Aggregation Signature Properties

In this section, we provide evidence that, for an image which adheres to a certain mathematical structure, the background can be suppressed by the aggregation signature.

Proposition: The image reconstructed from the aggregation signature matches the foreground object more accurately in the learning process with a high probability as follows:

$$\begin{aligned} &\mathcal{P}\left(E\left(\frac{<\mathbf{f}, \mathbf{S}^{i+1}>}{\|\mathbf{f}\| \cdot \|\mathbf{S}^{i+1}\|}\right) - E\left(\frac{<\mathbf{f}, \mathbf{S}^{i}>}{\|\mathbf{f}\| \cdot \|\mathbf{S}^{i}\|}\right) > 0\right) \\ &> \left(1 - 2\varepsilon\left(1-\varepsilon\right)\right)^N, \quad \text{for} \quad \forall i \geq 1, \end{aligned} \tag{5}$$

where $\mathcal{P}$ stands for probability, ε is a small positive value, N represents total image pixel number, $\|\cdot\|$ represents the l^2 norm, $<\cdot,\cdot>$ denotes the inner-product. $E(\cdot)$ denotes expectation, which reveals about the similarity between the foreground and the object saliency information obtained by aggregation signature.

[1] If $\mathbf{S}^i$ is the image signature based on DCT, we have $\mathbf{S}^i = \bar{\mathbf{x}}^{i-1}$.

TABLE I: Notation and terms used in this paper.

Terms	Notation
$\widehat{\mathbf{x}}$	DCT($\mathbf{x}$).
$\mathrm{sign}(\mathbf{x})$	The entrywise sign operator.
$\widetilde{\mathbf{x}}$	The conjugate form of $\mathbf{x}$.
$\bar{\mathbf{x}}$	IDCT [sign $(\widehat{\mathbf{x}})$], the reconstructed image of DCT.
$\bar{\mathbf{x}}_{\mathrm{Q}}$	IQDCT [sign (QDCT $(\mathbf{x})$)],the reconstructed image of QDCT.
$E(X)$	The expectation of random variable X.
$\|\mathbf{x}\|_p$	The l^p norm of vector $\mathbf{x}$. (p=2 if omitted).
$<\mathbf{x},\mathbf{y}>$	The inner-product of $\mathbf{x}$ and $\mathbf{y}$.
$\circ$	The Hadamard (entrywise) product operator.
$\Omega_{\mathbf{x}}$	Support set of $\widehat{\mathbf{x}}$.

Proof: We know the transform between QDCT and DCT is

$$\begin{aligned}\mathrm{QDCT}\left(\mathbf{x}_{\mathrm{Q}}^{i+1}\right) &= \sqrt{1/3}\mathrm{DCT}\left(-\mathbf{I_1}-\mathbf{I_2}-\mathbf{I_3}\right)\\ &+\sqrt{1/3}\mathrm{DCT}\left(\mathbf{I_3}+\boldsymbol{\pi}\circ\mathbf{S}^i-\mathbf{I_2}\right)j\\ &+\sqrt{1/3}\mathrm{DCT}\left(\mathbf{I_1}+\boldsymbol{\pi}\circ\mathbf{S}^i-\mathbf{I_3}\right)k\\ &+\sqrt{1/3}\mathrm{DCT}\left(\mathbf{I_2}+\boldsymbol{\pi}\circ\mathbf{S}^i-\mathbf{I_1}\right)h.\end{aligned} \tag{6}$$

For ease of explanation, we only focus on one channel, that is to say $\mathbf{S}^i=\bar{\mathbf{x}}^{i-1}$ and the result can be easily generalized for the quaternion case in a straightforward way, then we have

$$\begin{aligned}\|\bar{\mathbf{x}}^{i+1}\|^2 &= \mathrm{IDCT}<\left(\mathrm{sign}\left(\widehat{\mathbf{x}}^{i+1}\right)\right),\left(\mathrm{sign}\left(\widehat{\mathbf{x}}^{i+1}\right)\right)>\\ &=\sum_p\left(\mathrm{sign}\left(\widehat{x}^{i+1}(p)\right)\right)^2,\end{aligned} \tag{7}$$

where $\widehat{\mathbf{x}}=\mathrm{DCT}(\mathbf{x})$ and p represents the points of the corresponding support set. We note that the proof is applicable to channels j,k,h in Equ. (6), so we take the channel j for example. Then, we have

$$\begin{aligned}\mathrm{sign}\left(\widehat{\mathbf{x}}^{i+1}\right) &= \mathrm{sign}\left(\mathrm{DCT}\left(\mathbf{I_3}+\boldsymbol{\pi}\circ\bar{\mathbf{x}}^i-\mathbf{I_2}\right)\right)\\ &=\mathrm{sign}\left(\mathrm{DCT}\left(\mathbf{I_3}-\mathbf{I_2}\right)+\boldsymbol{\pi}\circ\mathrm{sign}\left(\widehat{\mathbf{x}}^i\right)\right).\end{aligned} \tag{8}$$

Since the results obtained by DCT are independent of each other, we assume

$$\begin{aligned}\mathcal{P}\left(\widehat{x}(p)=\pi(p)\right) &= \mathcal{P}\left(\widehat{x}(p)=0\right)\\ &=\mathcal{P}\left(\widehat{x}(p)=-\pi(p)\right)=\varepsilon,\end{aligned} \tag{9}$$

where ε is very small, since the probability that the DCT output is equal to a certain value is very small. Then we have the following statement:

$$\mathcal{P}\left(\|\bar{\mathbf{x}}^i\|^2=\|\bar{\mathbf{x}}^{i+1}\|^2\right)>\left(1-2\varepsilon(1-\varepsilon)\right)^N, \tag{10}$$

which means that in a high probability we have $\|\bar{\mathbf{x}}^i\|=\|\bar{\mathbf{x}}^{i+1}\|$, considering that ε is very small.

Similarly, we have

$$\begin{aligned}&<\bar{\mathbf{f}},\bar{\mathbf{x}}^{i+1}>=<\mathrm{sign}\left(\widehat{\mathbf{f}}\right),\mathrm{sign}\left(\widehat{\mathbf{x}}^{i+1}\right)>\\ &=<\mathrm{sign}\left(\widehat{\mathbf{f}}\right),\mathrm{sign}\left(\mathrm{DCT}\left(\mathbf{I_3}-\mathbf{I_2}\right)+\boldsymbol{\pi}\circ\mathrm{sign}\left(\widehat{\mathbf{x}}^i\right)\right)>.\end{aligned} \tag{11}$$

Since $\pi\geq 0$, if $|\mathrm{DCT}\left(I_3(p)-I_2(p)\right)|>\pi(p)$, then we have

$$\begin{aligned}&\mathrm{sign}\left(\mathrm{DCT}\left(I_3(p)-I_2(p)\right)\right)-\pi(p)\,\mathrm{sign}\left(\widehat{x}^i(p)\right)\\ &<\mathrm{sign}\left(\mathrm{DCT}\left(I_3(p)-I_2(p)\right)+\pi(p)\,\mathrm{sign}\left(\widehat{x}^i(p)\right)\right)\\ &<\mathrm{sign}\left(\mathrm{DCT}\left(I_3(p)-I_2(p)\right)\right)+\pi(p)\,\mathrm{sign}\left(\widehat{x}^i(p)\right).\end{aligned} \tag{12}$$

Combining (11) and (12), we have

$$\begin{aligned}&<\bar{\mathbf{f}},\bar{\mathbf{x}}^{i+1}>=<\mathrm{sign}\left(\widehat{\mathbf{f}}\right),\mathrm{sign}\left(\widehat{\mathbf{x}}^i+\mathrm{DCT}\left(\mathbf{I_3}-\mathbf{I_2}\right)\right)>\\ &\geq<\mathrm{sign}\left(\widehat{\mathbf{f}}\right),\boldsymbol{\pi}\circ\mathrm{sign}\left(\widehat{\mathbf{x}}^i\right)-\mathrm{sign}\left(\mathrm{DCT}\left(\mathbf{I_3}-\mathbf{I_2}\right)\right)>\\ &=<\bar{\mathbf{f}},\bar{\mathbf{x}}^i>+<\mathrm{sign}\left(\widehat{\mathbf{f}}\right),\mathrm{sign}\left(\mathrm{DCT}\left(\mathbf{I_2}-\mathbf{I_3}\right)\right)>.\end{aligned} \tag{13}$$

Based on the image signature proposed by Hou [19], we have

$$\begin{aligned}&<\mathrm{sign}\left(\widehat{\mathbf{f}}\right),\mathrm{sign}\left(\mathrm{DCT}\left(\mathbf{I_2}-\mathbf{I_3}\right)\right)>\\ &\geq|\Omega_{\mathbf{f}}|-|\Omega_{\mathbf{I_2}-\mathbf{I_3}-\mathbf{f}}|,\end{aligned} \tag{14}$$

where $\Omega_{\mathbf{x}}$ represents the support set of $\widehat{\mathbf{x}}$. Given the bound $|\Omega_{\mathbf{f}}|\geq\frac{2}{3}N$ [19], we have

$$\mathrm{E}\left(<\mathrm{sign}\left(\widehat{\mathbf{f}}\right),\mathrm{sign}\left(\mathrm{DCT}\left(\mathbf{I_2}-\mathbf{I_3}\right)\right)>\right)\geq\frac{1}{3}N. \tag{15}$$

And then it becomes

$$\mathrm{E}\left(<\bar{\mathbf{f}},\bar{\mathbf{x}}^{i+1}>-<\bar{\mathbf{f}},\bar{\mathbf{x}}^i>\right)\geq\frac{1}{3}N. \tag{16}$$

For a spatially sparse foreground, we have the following statement:

$$\mathrm{E}\left(<\mathbf{f},\bar{\mathbf{x}}^{i+1}>-<\mathbf{f},\bar{\mathbf{x}}^i>\right)>0. \tag{17}$$

Together with Equ. (10), we have

$$\begin{aligned}&\mathcal{P}\left(\mathrm{E}\left(\frac{<\mathbf{f},\bar{\mathbf{x}}^{i+1}>}{\|\mathbf{f}\|\cdot\|\bar{\mathbf{x}}^{i+1}\|}\right)-\mathrm{E}\left(\frac{<\mathbf{f},\bar{\mathbf{x}}^i>}{\|\mathbf{f}\|\cdot\|\bar{\mathbf{x}}^i\|}\right)>0\right)\\ &>\left(1-2\varepsilon(1-\varepsilon)\right)^N,\end{aligned} \tag{18}$$

which proves the proposition.

Remark: Here, ε is very small as in Equ. (9), e.g., $\varepsilon=0.0001$, and the probability mentioned above is 81% when $N=1024$. In other words, background is suppressed more during learning aggregation signature with high probability. We also did a statistic analysis on ε in Equ. (9) based on the MSRA-B dataset [21] , which indicates that ε is very small less than $\varepsilon=1.5e-9$.

III. Aggregation Signature Tracker

We exploit the aggregation signature to enhance the re-detection process for small object tacking, which is called aggregation signature tracker (AST). More specifically, when a target is found drifting by a thresholding method, a saliency detection with the tracked target as prior will be triggered, which enables the online aggregation signature to suppress the background data. Together with the context information indicated in different channels, we re-detect the objects to relocate the tracked target. The whole tracking procedure is

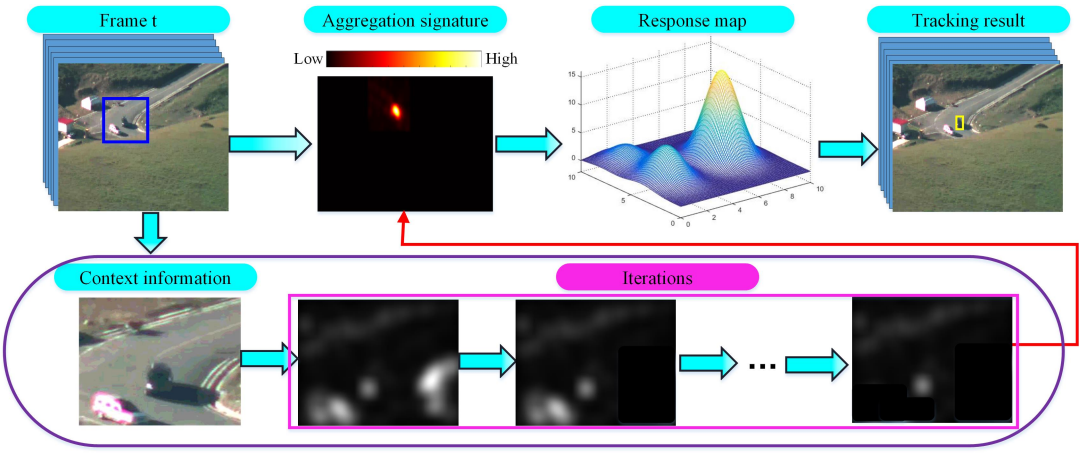

Fig. 2: Scheme of the aggregation signature tracker, which includes the base tracker and re-detection stages, particularly for small objects. The part of aggregation signature calculation illustrates the saliency map calculation in the re-detection procedure. Once a drifting is detected, we choose the search region around the center of the previous target location to calculate the saliency map via aggregation signature. The blue box is the search region. In the learning process, the target prior ($\boldsymbol{\pi}$) and the context information in the blue box are used to learn the saliency map that helps to find a new initial position, where the base tracker will be performed again for re-detection.

illustrated in Fig. 2(a) and Algorithm 1, and we elaborate each key component in the following.

Drifting detection: As evident on output constraint transfer tracking method (OCT) [22], a simple distribution is necessary and significant to achieve high efficiency. OCT builds upon a reasonable assumption that the response to the target image follows a Gaussian distribution, so we trigger the re-detection process based on a thresholding method as:

$$|r - \mu| > T_g, \tag{19}$$

where μ represents the mean response using all previous frames, r represents the maximum response of the current frame, and T_g is the threshold. The target is supposed to be lost if the response of the current frame is far from the average response. Once the target is occluded or out of view, this mechanism helps us search continuously in the following frames.

Saliency map calculation: The aggregation signature is used to obtain the saliency map and to further coarsely re-localize the target. Through R iterations, we gradually smooth the aggregation signature by a Gaussian kernel [19] to obtain the saliency map. The salient regions are regarded as the coarse candidate positions of the target, on which a re-detection process is performed still based on the selected base trackers. It should be mentioned that involving the targeted object, as a prior in saliency detection, does not occur in the conventional methods. Two key components are elaborated as follows:

1) Channels design: We denote the input image captured at frame t as $\mathbf{F}^t$, where $\mathbf{R}^t$, $\mathbf{G}^t$, and $\mathbf{B}^t$ are the red, green and blue channels of $\mathbf{F}^t$. Then, we obtain three channels used in our aggregation signature representing as: intensity $\mathbf{I}_1^t = (\mathbf{R}^t + \mathbf{G}^t + \mathbf{B}^t)/3$, saturation $\mathbf{I}_2^t = \max(\mathbf{R}^t, \mathbf{G}^t, \mathbf{B}^t)$ and movement $\mathbf{I}_3^t = |\mathbf{I}_1^t - \mathbf{I}_1^{t-\tau}|/3$, respectively, where τ is a constant. We deploy image signature [19] to calculate the initial saliency map as the first channel $\boldsymbol{\pi} \circ \mathbf{S}^i$.

2) Target Prior: As shown in Fig. 2 (b), we select M salient regions similar to the target in the last frame in size. Next, we assign each candidate a weight indicating the similarity to a target prior information, which is measured simply by the Euclidean distance as:

$$\pi_n^t = \frac{1}{\sqrt{2\pi}\xi} e^{-\frac{1-d_n^t}{2\xi^2}}, \tag{20}$$

where π_n^t denotes the weight of the n^{th} region for the candidate saliency map at the t^{th} frame, ξ is a constant. $d_n^t = \sum_{i=1}^{255} \sqrt{\left(\mathbf{H}^t(i) - \mathbf{y}_n^t(i)\right)^2}$, where $\mathbf{y}_n^t$ represents the histogram of the candidate saliency map, while $\mathbf{H}^t$ denotes the target histogram for the t^{th} frame calculated by

$$\mathbf{H}^t = \sigma \mathbf{H}^t + (1 - \sigma)\mathbf{H}^l, \tag{21}$$

where σ is 0.5 in this paper. We note that the weights are set to 1 for the regions outside the selected salient areas.

IV. Experiments

In this section, we evaluate the aggregation signature based on our small90 dataset and a visual saliency benchmark MSRA-B [21]. We further test the performance of our aggregation signature based tracker on the small90, small112, UAV123_10fps [23] and UAV20L [23] according to the object tracking benchmark [24]. The test platforms are Intel I7 2.7 GZ (4 cores) CPU with 8G RAM, and GPU with NVIDIA GeForce GTX 1070.

A. Datasets

Few datasets are available for small object tracking task. We establish a comprehensive database, termed small90 benchmark, consisting of 90 annotated small-sized object sequences, where several additional challenges, such as target drifting and low resolution, have been encompassed. We add 22 more challenging sequences into small90, and obtain another new

Fig. 4: The first frames of selected sequences from small90. The red bounding box indicates the ground truth.

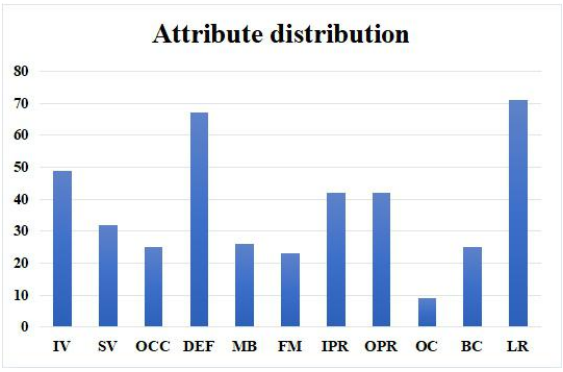

Fig. 3: Attribute distribution across small90.

dataset termed as small112. Each sequence is categorized with 11 attributes - 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), for better analysis of the tracking approaches. The attribute distribution in our dataset is plotted in Fig. 3, which shows that some attributes occur more frequently, e.g., LR, than the others. We note that one sequence is often annotated with multiple attributes. The examples of first frames from our datasets are illustrated in Fig. 4.

B. Aggregation Signature on Image

We first evaluate how aggregation signature can enhance the performance of saliency detection, based on the commonly used metrics including location-based metrics normalized scanpath saliency (NSS) [25], mean absolute error (MAE) [26] and distribution-based metric similarity (SIM) [27]. The comparative DCT image signature (IS) and QDCT image signature (QIS) are computed to extensively validate the effectiveness of our aggregation signature (AS) method, particularly on both MSRA-B [21] and small90 databases. There are 5000 images in MSRA-B, which is a large scale image database for quantitative evaluation of visual attention algorithms. From the results in Table II, we observe that our method achieves overall better performance quantitatively than IS and QIS in terms of the MAE, NSS, SIM measures, and thus leading to a better estimation of the visual distance between the predicted saliency map and the ground truth. Fig. 5 provides the saliency maps of different methods, and the ground truth on images from small90, which shows that the background is more suppressed in aggregation signature with resepct to the others methods. In terms of running speed, the aggregation signature module achieves 32 frames per second (FPS) in our experiments.

TABLE II: Experiments on metrics (MAE, NSS, SIM) among IS, QIS and AS on MSRA-B and small90. Bold fonts highlight the best performance.

	MSRA-B			small90		
	MAE	NSS	SIM	MAE	NSS	SIM
IS	0.2659	0.9710	0.3574	0.1513	3.6968	0.0163
QIS	0.2607	0.9808	0.3630	0.1260	4.4293	0.0288
AS	**0.2559**	**0.9844**	**0.3695**	**0.0660**	**7.2658**	**0.0611**

C. Aggregation Signature on Tracking

We empirically set the iteration number R equal to 4, the saliency patches M as 6. For other parameters, we follow the previous work [22] and set $\xi = 1$, $\tau = 3$, $T_g = 1.6$ in all experiments for fair comparisons.

We then test the performance of aggregation signature in tracking (AST) by comparing with DCT image signature,

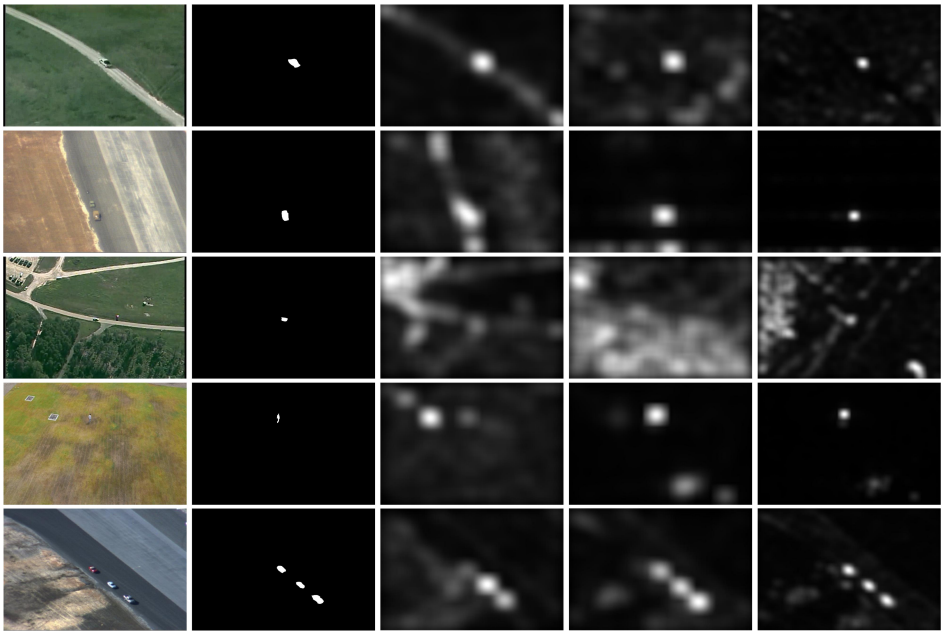

Fig. 5: Representative results of different signature methods. For each group from left to right, they are the original image, ground truth, IS, QIS and AS, respectively. Comparing with the other methods, AS yields the best background suppression performance.

QDCT image signature, which are incorporated with KCF, based on small90. The results in Fig. 6 reveal that the aggregation signature clearly outperforms other signatures in small object tracking. Also, we use one-pass evaluation (OPE) [24] to evaluate our results in the whole experiments section. Furthermore, we compare KCF_AST with other saliency-based trackers, including saliency prior context model (SPC) [15] and structuralist cognitive tracker (SCT) [28] in the same figure. KCF_AST (76.6%) is about 22% higher than SPC (54.9%), and 9% higher than SCT (67.7%) in terms of the precision, while KCF_AST (46.6%) is about 16% higher than SPC (30.9%) and 5% higher than SCT (42.1%) based on the average success rate.

We also compare our trackers with OCT, which also exploits the similar failure detection scheme to improve KCF. One can note that the performance of KCF_AST is higher than OCT by 13.2% and 7.8% in terms of precision and success rate, respectively.

The small90 benchmark: In Fig. 7, we further show the precision and success plots of 30 state-of-the-art trackers including SiamRPN [29] [30], LDES [31], SAT [32], TLD [3], LCT [33], OCT [22], CSK [34], CT [35], STC [36], KCF [37], ECO [38], MDNet [39], LCCF [40], SRDCF [41] and CPF [42], generated by the benchmark toolbox. While several baseline algorithms, e.g., LDES, DaSiamRPN, ECO, have shown promising potential in tracking small objects, our AST still helps achieve the precision rates of 84.9% (LDES_AST), 83.1% (DaSiamRPN_AST), 83.2% (ECO_AST) which improve its counterpart base trackers by 1.6%, 0.9%, 1.7% respectively. Meanwhile, the above three trackers with our AST on achieves a success rate of 68.6%, 69.7%, 64.3%, outperforming the base trackers by 1.7%, 0.4%, 0.9% respectively. Besides, our MDNet_AST outperforms by 7.1%

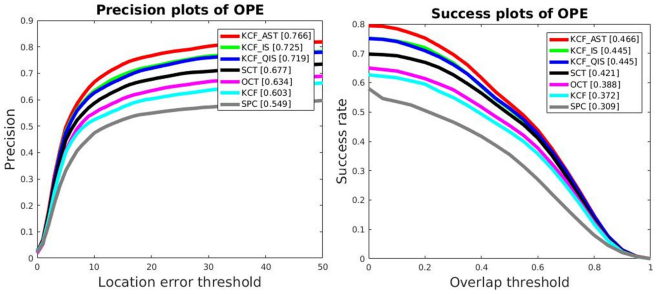

Fig. 6: Precision and success rate plots for AST performance on small90, in which SPC and SCT are the other two saliency-based trackers, while KCF_QDCT and KCF_DCT are employed to compare the aggregation signature performance in tracking.

and 4.0% respectively to achieve a precision rate of 86.6% and a success rate of 65.9% compared to MDNet. This again confirms that our aggregation signature can consistently improve the performance of base trackers. Likewise, LCCF_AST also shows a significant incremental performance, compared with the base tracker LCCF. Besides, when compared with the state-of-the-art re-detection trackers, our LCCF_AST (54.8%) significantly outperforms its base tracker LCCF (46.4%), and also TLD (52.7%), LCT(46.7%) and OCT (54.2%) by 2.1%, 8.3% and 0.7% in terms of the success rate on small90, respectively. The superior tracking performance confirms that our method is more effective than the state-of-the-art re-detection trackers such as TLD, LCT and OCT.

We illustrate some examples for KCF_AST in Fig. 8 to show how our aggregation signature helps to improve the tracking performance. In the sequences selected from smal-

Algorithm 1 - Aggregation signature tracker

Initial target bounding box $\mathbf{b}^1 = [p_x, p_y, w, h]$
Initial $\xi = 1$, $R = 4$, $M = 6$
if the frame $t <= 3$ **then**
  repeat
    Crop out the search windows according to $\mathbf{b}^t$, and extract feature
    Compute the maximal response according to base tracker
    The position is obtained according to the maximal response r
    Updating essential parameters of the base tracker
    $t = t + 1$
  until t==3
end if
Compute the mean μ of response using all previous frames
if the frame $t > 3$ **then**
  repeat
    Crop out the search windows according to $\mathbf{b}^t$, and extract feature
    Compute the maximal response according to base tracker
    if $|r - \mu| > T_g$ **then**
      Crop out the target search regions
      Obtain channels $\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{S}$ according to channels design in section 4
      Calculate the aggregation signature saliency map as illustrated in section 3.1
      Obtain the coarse target location based on the saliency map
      Compute new target location according to base tracker
    end if
    $t = t + 1$
  until the end of the video
  Updating parameters of the base tracker
  Updating μ
end if
end

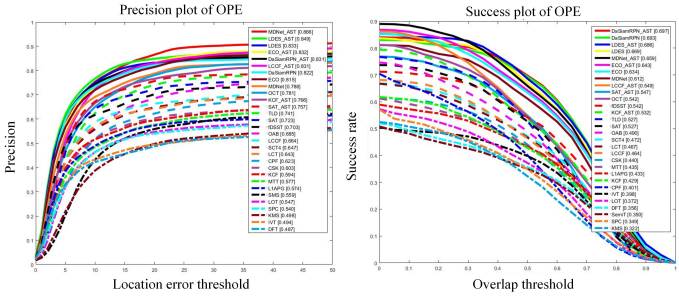

Fig. 7: Precision and success rate plots on small90.

l90, the tracked objects are subject to severe image quality deterioration during the tracking process. In particular: 1) the background of the scene presents clutters while many objects are similar to the target in appearance and 2) severe drifting or

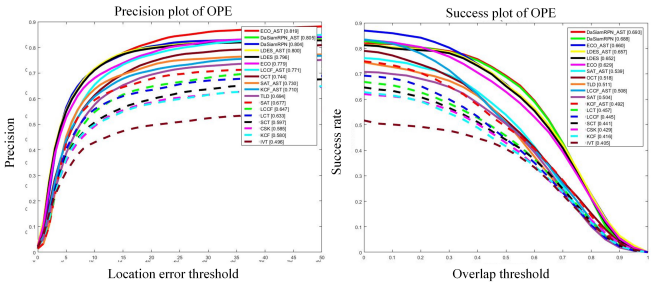

Fig. 9: Precision and success rate plots on small112.

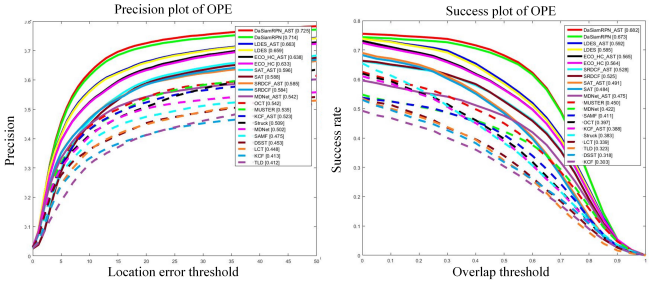

Fig. 10: Precision and success rate plots on UAV123_10fps.

long-time out of view results in directly drift of the target in the far range. In addition, we adopt the MDNet, LCCF (deep feature) and the KCF as base trackers in our frameworks for comparision of visual tracking experiments. Results are shown in Fig.11; our main goal here is to show how our method helps to drastically reduce the tracking failure.

Observed from the results on Fig. 8 and Fig. 11, we can conclude that the aggregation signature can effectively improve the performance of base trackers, especially for small object tracking, and both saliency detection and tracking are enhanced by incorporating our image signature. As a final consideration, we acknowledge that the proposed method has the ability to relocate the target when drifting, and performs very well on the small target sequences.

The small112 benchmark: We further collect a new benchmark dataset with 112 fully annotated sequences to facilitate the performance evaluation. On the basis of small90, the added 22 sequences are more difficult sequences. As shown in Fig. 9, KCF_AST, LCCF_AST, ECO_AST improve the performance of KCF, LCCF, ECO from 58.0%, 64.7%, 77.9% to 71.0%, 77.1%, 81.9% on precision rate and 41.6%, 44.5%, 62.9% to 49.2%, 50.8%, 66.0% on success rate, which demonstrates that AST improves these base trackers significantly on complex small object tracking sequences. Though the baseline trackers, such as SiamRPN, LDES, perform very well, still 0.1% and 0.4% improvements on precision and 0.5% and 0.5% improvements on success rate have been obtained by AST, which validates the effectiveness of AST. Observed from the experimental results, all the trackers endowed with the aggregation signature module perform consistently better than the base trackers, which further validates the effectiveness of the proposed approach. Also, the results show that better base trackers gain less performance improvements. The reason might be that aggregation signature is less useful if the drifting

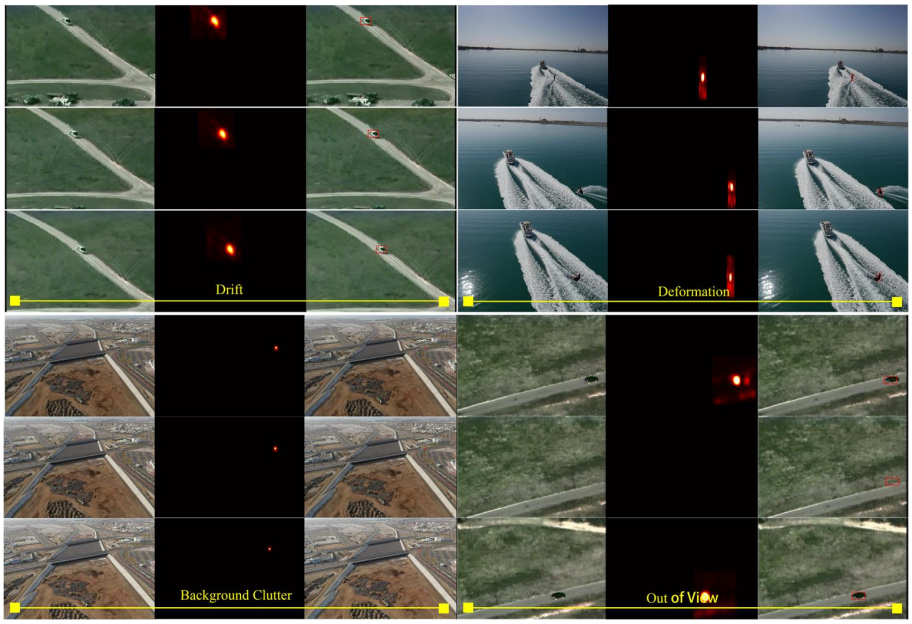

Fig. 8: Representative tracking results on four challenging sequences (fastcar, wakeboard, truck and blackcar). For each subfigure, the current frame, the saliency maps obtained by Aggregation Signature (AS) and the corresponding tracking results are shown in the first, middle and right column, respectively. We can see that our AST tracker could tackle the drifting, deformation, background clutter and out-of-view challenge due to the usage of aggregation signature.

is not obvious, which is the case of using a better tracker.

The UAV123_10fps benchmark: We test ASTs on UAV123_10fps [23] as shown in Fig. 10, which contains 123 sequences posing many challenges. Compared to the base tracker MDNet, we can see the aggregation signature (MDNet_AST) significantly improves the performance of MDNet from 50.2% to 54.2% in precision rate and 42.2% to 47.5% in success rate, which further validates the effectiveness of the proposed method. While KCF_AST is about 6% higher than KCF based on the precision, and is about 8% higher based on success rate. As for these more recent state-of-the-art trackers such as LDES, DaSiamRPN, ECO, their corresponding ASTs still achieve better results than these base trackers.

The UAV20L benchmark: We also test ASTs on the well-known benchmark UAV20L [23] as shown in Fig. 11, where some of the tracked objects are very small. The state-of-the-art SRDCF is chosen as the base tracker, leading to our SRDCF_AST. Apparently, SRDCF_AST obtains better performances with respect to the state-of-the-art. As compared to the base tracker SRDCF, we can see the aggregation signature (SRDCF_AST) significantly improves the performance of SRDCF from 50.7% to 53.1% in precision rate, which further validates the effectiveness of the proposed method. LCCF_AST is about 7% higher than LCCF, while KCF_AST is about 3% higher than KCF based on the precision. In addition, LCCF_AST and KCF_AST, though showing no outstanding performance in terms of success rate, still achieved better results than their base trackers, respectively. Furthermore, as for the more state-of-the-art trackers LDES and DaSiamRPN, we also show that LDES_AST and DaSiamRPN_AST improve their base trackers by a clear margin.

Quantitative Attribution Evaluation of Benchmarks:

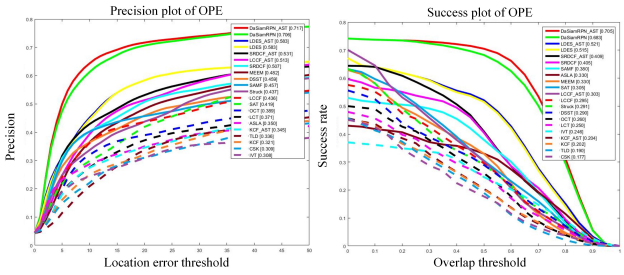

Fig. 11: Precision and success rate plots on UAV20L.

The full set of plots generated by the benchmark toolbox for small90 are also shown in Table III. From the results, we can conclude that AST trackers achieve a much better performance in most cases for small-sized objects, especially for motion blur and fast motion, in which we can see all AST trackers improve dramatically, since saliency-based AST trackers can be more robust than base trackers to the variations mentioned previously. To conclude, AST can consistently improve the results of base trackers in most cases, and AST-trackers achieve new state-of-the-art results.

Speed analysis: In terms of tracking speed on small90, KCF_AST has a processing rate of 120.88 frames per second (FPS), while LCCF_AST based on deep features has 16.52 FPS, which show that our proposed trackers not only achieve the state-of-the-art results, but also performs in real time. Although the frame rate of the proposed tracking framework has a drop, as compared to the original base tracker, the tracking performance is significantly improved on small90, e.g., 8.2% improvement on LCCF in terms of success rate.

TABLE III: Precision and Success rate for the 11 attributes in Small90. Bold fonts highlight the best performance

Precision	SPC	fDSST	OCT	SCT	KCF	KCF_AST	LCCF	LCCF_AST	MDNet	MDNet_AST	ECO	ECO_AST
IV	0.396	0.619	0.715	0.551	0.491	0.707	0.538	0.765	0.707	**0.798**	0.719	0.747
SV	0.495	0.710	0.723	0.618	0.574	0.805	0.706	0.805	0.775	0.794	0.768	**0.809**
OCC	0.619	0.678	0.751	0.726	0.673	0.772	0.692	0.732	0.799	**0.803**	0.757	0.758
DEF	0.542	0.706	0.767	0.676	0.599	0.757	0.671	0.805	0.807	**0.844**	0.777	0.793
MB	0.303	0.491	0.631	0.421	0.353	0.582	0.390	0.684	0.516	0.717	0.696	**0.726**
FM	0.353	0.573	0.746	0.500	0.412	0.645	0.452	0.789	0.573	**0.809**	0.770	0.803
IPR	0.438	0.672	0.811	0.604	0.522	0.752	0.623	0.844	0.787	**0.877**	0.779	0.805
OPR	0.464	0.704	0.838	0.625	0.551	0.782	0.650	0.869	0.831	**0.921**	0.808	0.833
OC	0.237	0.374	0.880	0.327	0.293	0.611	0.431	0.795	0.721	**0.855**	0.494	0.664
BC	0.533	0.696	0.770	0.655	0.599	0.769	0.653	0.815	0.786	**0.855**	0.789	0.804
LR	0.578	0.717	0.816	0.666	0.625	0.783	0.697	0.858	0.805	**0.900**	0.845	0.863
Success rate	SPC	fDSST	OCT	SCT	KCF	KCF_AST	LCCF	LCCF_AST	MDNet	MDNet_AST	ECO	ECO_AST
IV	0.209	0.379	0.423	0.328	0.291	0.422	0.322	0.445	0.430	**0.487**	0.451	0.464
SV	0.264	0.459	0.396	0.361	0.324	0.416	0.393	0.439	0.511	0.519	0.504	**0.524**
OCC	0.343	0.435	0.465	0.460	0.439	0.469	0.446	0.461	0.502	**0.507**	0.480	0.479
DEF	0.305	0.454	0.456	0.425	0.378	0.460	0.411	0.477	0.524	**0.540**	0.508	0.514
MB	0.150	0.299	0.396	0.260	0.201	0.368	0.230	0.402	0.324	0.464	0.453	**0.474**
FM	0.185	0.367	0.473	0.317	0.246	0.421	0.282	0.481	0.374	0.537	0.514	**0.538**
IPR	0.251	0.412	0.470	0.367	0.316	0.446	0.374	0.483	0.486	**0.541**	0.480	0.491
OPR	0.262	0.433	0.481	0.376	0.329	0.460	0.386	0.495	0.514	**0.570**	0.500	0.510
OC	0.150	0.263	0.408	0.209	0.181	0.382	0.242	0.436	0.425	**0.512**	0.329	0.427
BC	0.305	0.451	0.471	0.416	0.376	0.476	0.407	0.493	0.511	**0.552**	0.526	0.532
LR	0.334	0.469	0.499	0.414	0.382	0.475	0.417	0.507	0.527	**0.587**	0.561	0.571

V. Conclusions

A new aggregation signature has been proposed to improve the small target tracking performance. The aggregation signature uses the target as a prior to adaptively locate the salient object, which is deployed to re-detect the tracked objects when drifting. It is generic and can be used in conjunction with other trackers. We evaluated our tracking framework with KCF, SRDCF, LCCF, ECO, SAT, LDES, DaSiamRPN and MDNet. To validate the resulting aggregation signature tracker, we have also collected new video datasets named small90 and small112, which contain fully annotated video sequences for small target tracking. The experimental results have clearly demonstrated how our methods improve the performance for the challenging situations, such as severe drifting, deformation and out of view. Furthermore, our approach will be extended to different applications in the future, such as large-scale retrieval [43][44] and classification [45].

VI. Acknowledgment

The work was supported by the National Key Research and Development Program of China (Grant No. 2016YFB0502602) and National Natural Science Foundation of China under Grant 61672079, in part by Supported by Shenzhen Science and Technology Program (No.KQTD2016112515134654).

References

[1] S. Stalder, H. Grabner, L. V. Gool, Cascaded confidence filtering for improved tracking-by-detection, in: European Conference on Computer Vision, 2010, pp. 369–382.

[2] M. Heber, M. Godec, M. Rther, P. M. Roth, H. Bischof, Segmentation-based tracking by support fusion, Computer Vision and Image Understanding 117 (6) (2013) 573–586.

[3] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (7) (2012) 1409–1422.

[4] M. Danelljan, G. Hager, F. S. Khan, M. Felsberg, Discriminative scale space tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (8) (2017) 1561–1575.

[5] S. Hare, A. Saffari, P. H. S. Torr, Struck: Structured output tracking with kernels, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (10) (2016) 2096–2109.

[6] B. Zhang, A. Perina, Z. Li, V. Murino, J. Liu, R. Ji, Bounding multiple gaussians uncertainty with application to object tracking, International Journal of Computer Vision 118 (3) (2016) 364–379.

[7] C. Deng, Y. Han, B. Zhao, High performance visual tracking with extreme learning machine framework, IEEE Trans. on Cyber. URL DOI:10.1109/TCYB.2018.2886580

[8] K. Ahmadi, E. Salari, Small dim object tracking using frequency and spatial domain information, Pattern Recognition 58 (2016) 227–234.

[9] Y. Han, C. Deng, B. Zhao, B. Zhao, Spatial-temporal context-aware tracking, IEEE Signal Process. Lett. 26 (3) (2019) 500–504.

[10] K. Ahmadi, E. Salari, Small dim object tracking using a multi objective particle swarm optimisation technique, IET Image Processing 9 (9) (2015) 820–826.

[11] D. Rozumnyi, J. Kotera, F. Sroubek, L. Novotny, J. Matas, The world of fast moving objects, in: Computer Vision and Pattern Recognition, 2017, pp. 4838–4846.

[12] C. Liu, W. Ding, X. Xia, B. Zhang, J. Gu, J. Liu, R. Ji, D. Doermann, Circulant binary convolutional networks: Enhancing the performance of 1-bit dcnns with circulant back propagation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2691–2699.

[13] C. Liu, W. Ding, X. Xia, , Y. Hu, B. Zhang, J. Liu, B. Zhuang, G. Guo, Rbcn: Rectified binary convolutional networks for enhancing the performance of 1-bit dcnns, 2019.

[14] J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, Y. C. Jin, Attentional correlation filter network for adaptive visual tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4828–4837.

[15] C. Ma, Z. Miao, X. P. Zhang, M. Li, A saliency prior context model for real-time object tracking, IEEE Transactions on Multimedia PP (99) (2017) 1–1.

[16] K. M. Yi, H. Jeong, B. Heo, H. J. Chang, Y. C. Jin, Initialization-insensitive visual tracking through voting with salient local features, in: IEEE International Conference on Computer Vision, 2014, pp. 2912–2919.

[17] J. Fan, Y. Wu, S. Dai, Discriminative spatial attention for robust tracking, in: European Conference on Computer Vision, 2010, pp. 480–493.

[18] Y. Luo, J. Yuan, Salient object detection in videos by optimal spatio-temporal path discovery, in: Acm International Conference on Multimedia, 2013, pp. 509–512.

[19] X. Hou, J. Harel, C. Koch, Image signature: Highlighting sparse salient regions, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (1) (2012) 194.

[20] B. Schauerte, R. Stiefelhagen, Quaternion-based spectral saliency detec-

tion for eye fixation prediction, in: European Conference on Computer Vision, 2012, pp. 116–129.
[21] T. Liu, J. Sun, N. N. Zheng, X. Tang, H. Y. Shum, Learning to detect a salient object, in: Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[22] B. Zhang, Z. Li, X. Cao, Q. Ye, C. Chen, L. Shen, A. Perina, R. Jill, Output constraint transfer for kernelized correlation filter in tracking, IEEE Transactions on Systems Man and Cybernetics Systems 47 (4) (2017) 693–703.
[23] M. Mueller, N. Smith, B. Ghanem, A benchmark and simulator for uav tracking, in: European Conference on Computer Vision, 2016, pp. 445–461.
[24] Y. Wu, J. Lim, M. H. Yang, Object tracking benchmark, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (9) (2015) 1834.
[25] A. Borji, D. N. Sihite, L. Itti, Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study, IEEE Transactions on Image Processing 22 (1) (2013) 55.
[26] C. J. Willmott, K. Matsuura, Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, Climate Research 30 (1) (2005) 79.
[27] J. Li, C. Xia, Y. Song, S. Fang, X. Chen, A data-driven metric for comprehensive evaluation of saliency models, in: IEEE International Conference on Computer Vision, 2015, pp. 190–198.
[28] J. Choi, H. J. Chang, J. Jeong, Y. Demiris, Y. C. Jin, Visual tracking using attention-modulated disintegration and integration, in: Computer Vision and Pattern Recognition, 2016, pp. 4321–4330.
[29] Z. Zhu, Q. Wang, L. Bo, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: European Conference on Computer Vision, 2018.
[30] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[31] Y. Li, J. Zhu, W. Song, Z. Wang, Robust estimation of similarity transformation for visual object tracking, Association for the Advance of Artificial Intelligence.
[32] Y. Han, C. Deng, B. Zhao, D. Tao, State-aware anti-drift object tracking, IEEE Transactions on Image Processing 28 (8) (2019) 4075–4086.
[33] C. Ma, X. Yang, C. Zhang, M. H. Yang, Long-term correlation tracking, in: Computer Vision and Pattern Recognition, 2015, pp. 5388–5396.
[34] C. Rui, P. Martins, J. Batista, Exploiting the circulant structure of tracking-by-detection with kernels, in: European Conference on Computer Vision, 2012, pp. 702–715.
[35] K. Zhang, L. Zhang, M. H. Yang, Real-time compressive tracking, in: European Conference on Computer Vision, 2012, pp. 864–877.
[36] K. Zhang, L. Zhang, Q. Liu, D. Zhang, M. H. Yang, Fast visual tracking via dense spatio-temporal context learning, in: European Conference on Computer Vision, 2014, pp. 127–141.
[37] J. F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (3) (2015) 583–596.
[38] M. Danelljan, G. Bhat, F. Shahbaz Khan, M. Felsberg, Eco: Efficient convolution operators for tracking, in: Computer Vision and Pattern Recognition, 2017, pp. 6931–6939.
[39] H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, in: Computer Vision and Pattern Recognition, 2016, pp. 4293–4302.
[40] B. Zhang, S. Luan, C. Chen, J. Han, W. Wang, A. Perina, L. Shao, Latent constrained correlation filter, IEEE Transactions on Image Processing PP (99) (2018) 1–1.
[41] M. Danelljan, G. Hager, F. S. Khan, M. Felsberg, Learning spatially regularized correlation filters for visual tracking, in: IEEE International Conference on Computer Vision, 2015, pp. 4310–4318.
[42] P. Perez, C. Hue, J. Vermaak, M. Gangnet, Color-based probabilistic tracking, European Conference on Computer Vision I (2002) 661–675.
[43] G. Wu, J. Han, Y. Guo, L. Liu, L. Shao, Unsupervised deep video hashing via balanced code for large-scale video retrieval, IEEE Transactions on Image Processing 28 (4) (2018) 1993–2007.
[44] G. Wu, J. Han, Z. Lin, G. Ding, B. Zhang, Q. Ni, Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning, IEEE Transactions on Industrial Electronics 66 (12) (2019) 9868–9877.
[45] G. Ding, Y. Guo, K. Chen, C. Chu, J. Han, Q. Dai, Decode: Deep confidence network for robust image classification, IEEE Transactions on Image Processing 28 (8) (2019) 3752–3765.

Chunlei Liu received her B.S. degree in Information Engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016. She is now pursuing her phd degree at the Department of Electrical and information Engineering in Beihang University. Her research interests include computer vision and machine learning.

Wenrui Ding received the doctorate degree in electrical and information engineering from Beihang University. She is currently in charge of information transmission and processing data link in the Unmanned System Research Institute in Beihang University. Her research interests include computer vision, the command and control of UAV, image processing, and pattern recognition.

Jinyu Yang received her B.S. and M.Sc. degrees in Technology and Apparatus of Measuring and Control from Beihang University, Beijing, China and in Electronic Engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2018 and 2019, respectively. She is now pursuing her Ph.D. degree in Computer Science in the University of Birmingham. Her research interests includes computer vision and machine learning.

Vittorio Murino is full professor at the University of Verona, Italy, and director of PAVIS (Pattern Analysis and Computer Vision) department at the Istituto Italiano di Tecnologia. He took the Laurea degree in Electronic Engineering in 1989 and the Ph.D. in Electronic Engineering and Computer Science in 1993 at the University of Genova, Italy. From 1995 to 1998, he was assistant professor at the Dept. of Mathematics and Computer Science of the University of Udine, Italy, and since 1998 he works at the University of Verona. He was chairman of the Department of Computer Science from 2001, year of foundation, to 2007, and coordinator of the Ph.D. program in Computer Science in the same university from 1999 to 2003. Prof. Murino is scientific responsible of several national and European projects, and evaluator of EU project proposals related to several frameworks and programs. Since 2009, he is working at the Istituto Italiano di Tecnologia in Genova, Italy, leading the PAVIS department involved in computer vision, machine learning, pattern recognition and image analysis. His main research interests include computer vision, pattern recognition and machine learning, more specifically, statistical and probabilistic techniques for image and video processing, with applications on (human) behavior analysis and related applications such as video surveillance, biomedical imaging, and bioinformatics. Prof. Murino is co-author of more than 400 papers published in refereed journals and international conferences, member of the technical committees of important conferences (CVPR, ICCV, ECCV, ICPR, ICIP, etc.), and guest co-editor of special issues in relevant scientific journals. He is also member of the editorial board of Computer Vision and Image Understanding, Machine Vision and Applications, and Pattern Analysis and Applications journals. Finally, prof. Murino is IEEE Senior Member and IAPR Fellow.

Baochang Zhang received the B.S., M.S. and Ph.D. degrees in Computer Science from Harbin Institue of the Technology, Harbin, China, in 1999, 2001, and 2006, respectively. From 2006 to 2008, he was a research fellow with the Chinese University of Hong Kong, Hong Kong, with Istituto Italiano di Tecnologia, Italy, and with Griffith University, Brisban, Australia. Currently, he is a tenured associate professor with Beihang University, Beijing, China. His current research interests include pattern recognition, machine learning, face recognition, and wavelets.

Jungong Han is currently a tenured Data Science Associate Professor with the University of Warwick, U.K. He has published more than 80 papers in the top venues including IEEE/ACM Transactions and A* conference papers. His research interests include computer vision, artificial intelligence, and machine learning.

Guodong Guo (M'07-SM'07) received the B.E. degree in automation from Tsinghua University, Beijing, China, the Ph.D. degree in computer science from University of Wisconsin, Madison, WI, USA. He is currently the Deputy Head of the Institute of Deep Learning, Baidu Research, and also an Associate Professor with the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), USA. In the past, he visited and worked in several places, including INRIA, Sophia Antipolis, France; Ritsumeikan University, Kyoto, Japan; and Microsoft Research, Beijing, China; He authored a book, *Face, Expression, and Iris Recognition Using Learning-based Approaches* (2008), co-edited two books, *Support Vector Machines Applications* (2014) and *Mobile Biometrics* (2017), and published over 100 technical papers. His research interests include computer vision, biometrics, machine learning, and multimedia. He received the North Carolina State Award for Excellence in Innovation in 2008, Outstanding Researcher (2017-2018, 2013-2014) at CEMR, WVU, and New Researcher of the Year (2010-2011) at CEMR, WVU. He was selected the "People's Hero of the Week" by BSJB under Minority Media and Telecommunications Council (MMTC) in 2013. Two of his papers were selected as "The Best of FG'13" and "The Best of FG'15", respectively.