Joint Scale-Spatial Correlation Tracking with Adaptive Rotation Estimation

Mengdan Zhang, Junliang Xing, Jin Gao, Xinchu Shi, Qiang Wang, Weiming Hu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
No. 95, Zhongguancun East Road, Beijing 100190, P. R. China
{mengdan.zhang, jlxing, jgao, xcshi, qiang.wang, wmhu}@nlpr.ia.ac.cn

Abstract

Boosted by large and standardized benchmark datasets, visual object tracking has made great progress in recent years and brought about many new trackers. Among these trackers, correlation filter based tracking schema exhibits impressive robustness and accuracy. In this work, we present a fully functional correlation filter based tracking algorithm which is able to simultaneously model target appearance changes from spatial displacements, scale variations, and rotation transformations. The proposed tracker first represents the exhaustive template search in the joint scale and spatial space by a block-circulant matrix. Then, by transferring the target template from the Cartesian coordinate system to the Log-Polar coordinate system, the circulant structure is well preserved for the target even after whole orientation rotation. With these novel representation and transformation, object tracking is efficiently and effectively performed in the joint space with fast Fourier Transform. Experimental results on the VOT2015 benchmark dataset demonstrate its superior performance over state-of-the-art tracking algorithms.

1. Introduction

Visual object tracking has long been an active research topic in computer vision, with numerous practical applications in visual surveillance, human computer interaction, intelligent transportation system, and robot navigation [27]. In the past years, due to the development and popularization of several large standardized benchmark datasets [14, 15, 26, 24, 16], the research on visual object tracking has witnessed great progress, resulting many new algorithms with very promising tracking performance [26, 24, 15].

For the class of online single arbitrary object tracking, a general setting is to manually initialize a bounding box around the target and then estimate its state in the following frames. Although much progress has been made for this specific tracking task, it remains a very challenging problem, especially when the target undergoes serious variations in appearance, scale, pose, shape, motion, etc.

Recently, the correlation filter (CF) based tracking scheme [3, 11, 12] has demonstrated very promising performance with good computational efficiency. The core idea of this tracking scheme comes from the Convolution Theorem in signal processing which states that the correlation operation in time domain corresponds to an element-wise multiplication in the Fourier domain. Since the exhaustive spatial template matching in the image plane can be viewed as a correlation operation in the time domain, template matching based tracking can be efficiently performed in the frequency domain. What is more, by employing the response of correlation filter as a similarity measure between two image signals, a very reliable distance metric for the observation model of the target can be obtained.

The early CF based trackers only [3, 11] perform correlation operation with only one target template of fixed size, thus can only model the positional changes and is likely to fail when target changes its size or rotates within the image plane. To deal with these problems, we propose a fully functional correlation filter based tracking algorithm. Compared to other CF based trackers, our tracking algorithm is able to simultaneously model target appearance changes from spatial displacements, scale variations, and rotation transformations.

Although several CF based trackers with scale adaptation have been proposed recently [5, 17], our tracking algorithm performs scale-spatial correlation jointly using a novel block-circulant structure for the object template, which provides a principled solution for the object scale estimation problem. Previous scale adaptation methods, provide only an approximation [5] or ignores the correlations between the translation and scale changes of the target [17].

To model target rotations, our algorithm proposes to transform the object templates from the Cartesian coordinate system to the Log-Polar coordinate system. With this transformation, the circulant structure of the object template undergoing rotation changes got preserved, thus enabling the tracker to model the object rotations in the same framework as spatial displacements and scale changes.

After the transformation of the coordinate system, our algorithm extracts dense cyclical training samples in the whole orientation space. The trained model with these samples, therefore, can deal with all possible rotations of the target, which also provide a complete solution to the object rotation estimation problem.

To model the object appearance in the CF tracking framework, we employ a set of robust image features, including the histogram of orientation gradients (HOG) and the 10 dimensional color names (CN). With an early fusion strategy, the obtained observation model of the target can model the appearance variations of the target robustly .

By integrating all these innovations together, we have obtained a very robust CF based tracker. Experimental evaluation on the VOT2015 benchmark has demonstrated that our tracker not only performs significantly better than previous CF based trackers, but also exhibits obvious superiority over many state-of-the-art tracking algorithms.

The main contributions of this work can be summarized as follows: (1) we extend the correlation tracking to the joint scale-spatial space with the observation of the block-circulant structure in the dense template matching samples, which leads to a new efficient scale adaptation scheme; (2) we propose the correlation analysis in the whole orientation space, which further enriches the correlation tracking framework and gives a new perspective on rotation estimation; (3) the competitive results of VOT2015 prove our tracker's robustness and adaptability.

2. Related Work

For online single arbitrary object tracking, appearance model and tracking strategy are two key components, on which extensive studies have been performed. Early generative methods build object templates as the appearance model and perform tracking by exhaustively searching for the best candidate locations in the next frame [23, 4, 25, 1]. This kind of methods are very straightforward to implement, but are often very time-consuming and may fail when an object changes its appearance in size, scale, pose, etc.

To overcome the limitations of template matching based tracking methods, discriminate learning methods [2, 10, 9] formulate object tracking as a classification problem and employ online learning methods to learn a discriminative classifier between the target and the backgrounds, which improves the tracking robustness to appearance changes of the target. To avoid evaluating all possible samples from the object and backgrounds, this kind of methods usually selects only a small subset of the samples for training and updating the classifier, as well as testing the candidate states. The number and quantity of the selected samples, therefore, will impact a lot on the tracking results.

Recently, the template matching based tracking methods have regained attention due to the introduction of correlation filters into the visual tracking problem [3, 11]. The main advantage of the correlation filter based tracking scheme is that it can perform effective template matching exhaustively while it keeps high computational efficiency. This is achieved by transforming the spatial correlation operation into the dot multiplication in the Fourier domain by means of the fast Discrete Fourier Transform. By employing the kernel trick on the correlation filter, the KCF tracker [12] obtains very promising results on two recent tracking benchmarks [24, 15].

Despite its great success, the original KCF method also has some serious limitations, e.g., it can not deal with the scale changes and rotations of the target. To surmount these limitations, many extensions over the KCF algorithm have been proposed [5, 17, 6, 20, 13, 18, 19, 22], which are designed to adapt KCF algorithm to scale changes [5, 17], perform part based KCF tracking [18, 19], or improve the feature representations [6, 20, 13, 22]. In this work, the proposed tracking algorithm is also inspired by the KCF algorithm, but we provide a fully functional correlation KCF tracker which can simultaneously deal with object appearance changes from spatial shifting, scale variation, and rotation transformation.

3. Overall Approach

In this section, we first briefly review the correlation tracking scheme to make our paper self-contained. Then we introduce our tracker which performs the correlation operation in the joint scale and spatial space. After that, we will present our rotation adaptation scheme in the tracker. Lastly, we will describe our feature fusion strategy.

3.1. Correlation tracking

The correlation tracking [3, 11] provides an elegant framework for efficient and effective object tracking. The KCF tracker [11], kernelized version of the correlation tracking, demonstrates impressive robustness and efficiency. The key for its high robustness is that the KCF tracker uses dense cyclic sampling to dig out the target's structural information and models this special structure with the circulant matrix. Instead of exhaustively matching dense samples to the target template in the spatial space, the tracker performs efficient element-wise multiplication in the frequency domain through Fast Fourier Transform (FFT). The fast computation benefits from the circulant theorem [7] and the convolution theorem [21].

In KCF, the template matching is trained on a single $M \times N$ image patch x centered around the target. According to the circular convolution theorem, to avoid spectrum aliasing, the patch is usually larger than twice the size of the target. When we sample continuously around the target, without considering the boundary effect, the translation of

the search window can be approximately considered as the cyclic shift of the base sample x. Thus, all cyclic shifts $\{x_{m,n}\}_{m\in\{1,\dots,M-1\},n\in\{1,\dots,N-1\}}$, are considered as the training samples for the template estimation. The matching score is modeled as a Gaussian response y, so that $y(m,n)$ is the matching score for the training sample $x_{m,n}$. The solution w is obtained by minimizing the ridge regression error:

$$\min_{w} \sum_{m,n} |\langle \phi(x_{m,n}), w\rangle - y(m,n)|^2 + \lambda \|w\|^2, \quad (1)$$

where ϕ is the mapping to the Hilbert space induced by the kernel κ, defining the inner product as $\langle \phi(x), \phi(\tilde{x})\rangle = \kappa(x,\tilde{x})$. The constant $\lambda \geq 0$ is the regularization parameter controlling the model simplicity.

After the nonlinear transform, the solution w can be expressed as $w = \sum_{m,n} \alpha(m,n)\phi(x_{m,n})$. We denote the Discrete Fourier Transform (DFT) of a vector with a hat. So the dual solution can be:

$$\hat{\alpha} = \frac{\hat{y}}{\hat{k}^{xx} + \lambda}, \quad (2)$$

where $k^{xx} = \kappa(x_{m,n}, x)$. Usually, we use the Gaussian kernel to compute the kernel correlation k^{xx} with efficient element-wise products in the Fourier domain. For an image patch with C feature channels, the base sample is the concatenation $x = [x_1, x_2, \dots, x_C]$. Thus, we have:

$$k^{xx'} = \exp\left(-\frac{1}{\sigma^2}\left(\|x\|^2 + \|x'\|^2 - 2\mathcal{F}^{-1}\left(\sum_{c=1}^{C} \hat{x}^c \odot (\hat{x'}^c)^*\right)\right)\right), \quad (3)$$

where $\odot$ represents element-wise products and c is the index of feature channels.

In the tracking step, a patch z with the same size as x is cropped out in the new frame. The matching scores for all the cyclic patches of z can be calculated via

$$y = \mathcal{F}^{-1}(\hat{k}^{x^\star z} \odot \hat{\alpha}), \quad (4)$$

where $x^\star$ denotes the learned target appearance. The target's position in the new frame is then estimated by finding the translation maximizing the matching score.

3.2. Joint scale-spatial correlation tracking

The KCF tracker is not qualified to deal with large scale variance, because a single base sample only contains target information from one scale level. Thus, to incorporate scale estimation into visual tracking, the JSSC tracker [28] extracts image patches continuously from the joint scale-spatial space. For each scale level, there is a base sample centered around the target. Consequently, the block-circulant structure is dug out and similarly transform the dense template matching problem into the Fourier domain. Since the JSSC tracker estimates the target's position and size simultaneously, it is less likely to accumulate tracking error and cause drift, which contributes to the improvement of the performance of the correlation filter based trackers. This is very different to the previous scale approximation methods for KCF tracker [5, 17].

For simplicity, assume a $1D$ image and a single-channel feature. The JSSC tracker is trained using S base samples of size $1 \times N$ obtained from the recent scale level and neighboring levels. Taking advantages of the cyclic property and appropriate padding, it considers all cyclic shifts $\{x_s(n)\}, s \in 1, 2, \dots, S, n \in \{0, 1, \dots, N-1\}$ as the training samples for the target template estimation. The matching scores y obey a multivariate Gaussian distribution in the joint scale-spatial space. To minimize the squared error over sample response and the defined matching scores, it uses the regularized Ridge Regression with the kernel trick:

$$\min_{w} \sum_{n,s} |\langle \phi(x_s(n)), w\rangle - y_s(n)|^2 + \lambda \|w\|^2, \quad (5)$$

where $y_s(n)$ is the matching score of the sample $x_s(n)$. Furthermore, the closed-form solution in the dual space for the Kernelized Ridge Regression is obtained:

$$\alpha = (K + \lambda I_{SN})^{-1} y. \quad (6)$$

The $S \times S$ block matrix K is a collection of the kernel matrices generated between different scale levels and reveals their correlation. Concisely, a block $K_{ij}(i, j = 1, 2, \dots, S)$ denotes the $N \times N$ kernel matrix calculated from the scale levels x_i and x_j. Additionally, the output of the kernel function for each pair of samples from the two scale layers can be given by:

$$K_{ij}(q,l) = \kappa(x_i(q), x_j(l)), (q, l = 0, 1, \dots, N-1). \quad (7)$$

It is time-consuming to calculate the inverse of a large non-sparse matrix in (6). However, it can be quite simple when the block matrix K implies block-circulant structure. According to the KCF tracker [12], each block of the matrix K can be shown circulant. Then, select elements from the same place of each block of K and store them in an $S \times S$ matrix. Finally, an $N \times N$ block-circulant matrix $\bar{K}$ is obtained, which can be diagonalized by the DFT matrix. The first row of the block-circulant matrix $\bar{K}$ is considered as the base block sequence, denoted $[\Psi_1, \Psi_2, \cdots, \Psi_N]$. As in [7], the block-circulant matrix is diagonalized as:

$$\bar{K} = W \,\mathrm{diag}(g(u_0), g(u_1), \cdots, g(u_{N-1}))W^H, \quad (8)$$

$$g(x) = \Psi_1 + \Psi_2 x + \cdots + \Psi_N x^{N-1}, \quad (9)$$

$$W = F \otimes I_S, \quad (10)$$

$$u_k = \exp(-j\frac{2\pi k}{N}), \tag{11}$$

where $g(x)$ calculates the DFT of the base block sequence and F is the DFT matrix.

Since the JSSC tracker estimate the target's size continuously, the value S, the block size of the block diagonal matrix, can be small. As a result, it is convenient to calculate the inverse of these small blocks. The JSSC solution in the Fourier domain is extended as

$$\hat{\bar{\alpha}}^* = (\text{diag}(g(u_0), g(u_1), \cdots, g(u_{N-1})) + \lambda I_{SN})^{-1} \hat{\bar{y}}^*, \tag{12}$$

$$g(u_c) = \begin{bmatrix} \hat{k}_c^{x_1 x_1} & \hat{k}_c^{x_1 x_2} & \dots & \hat{k}_c^{x_1 x_S} \\ \hat{k}_c^{x_2 x_1} & \hat{k}_c^{x_2 x_2} & \dots & \hat{k}_c^{x_2 x_S} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{k}_c^{x_S x_1} & \hat{k}_c^{x_S x_2} & \dots & \hat{k}_c^{x_S x_S} \end{bmatrix}, \tag{13}$$

where $k_c^{x_i x_j}$ is the $c-$th element of the base sample of the Gaussian kernel matrix K_{ij}, the horizontal bars represent the rearrangement.

In the tracking section, the candidates Z to be matched with the target template are extracted in the same way from the joint scale-spatial space. The matching scores can be evaluated via

$$f(Z) = K^{ZX}\alpha, \tag{14}$$

where X is the learned target appearance in the joint space. The block matrix K^{ZX} shows the kernel correlation between all candidate patches and the target templates. Considering the block-circulant matrix properties, the full tracking response is given by

$$\hat{\bar{f}}(Z) = \text{diag}(h^*(u_0), h^*(u_1), \cdots, h^*(u_{N-1}))\hat{\bar{\alpha}}, \tag{15}$$

$$h(u_c) = \begin{bmatrix} \hat{k}_c^{z_1 x_1} & \hat{k}_c^{z_1 x_2} & \dots & \hat{k}_c^{z_1 x_S} \\ \hat{k}_c^{z_2 x_1} & \hat{k}_c^{z_2 x_2} & \dots & \hat{k}_c^{z_2 x_S} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{k}_c^{z_S x_1} & \hat{k}_c^{z_S x_2} & \dots & \hat{k}_c^{z_S x_S} \end{bmatrix}, \tag{16}$$

Moreover, linear interpolation is adopted according to the tracking response to ensure the continuity of the scale and position estimation.

3.3. Adaptive rotation estimation

During tracking, an object may be viewed at different orientations due to the object or camera motion. The changes in object orientation may degenerate the tracking performance of general trackers whose tracking bounding box is aligned with the Cartesian coordinates. This is mainly due to the fact that the appearance model may not be appropriate at the moment and much more noise may be introduced. This observation suggests that a good tracker should model the object rotation transformations together with the translation changes and scale variations of the object. We propose to perform rotation estimation using a unified correlation tracking framework by taking the Log-Polar transformation.

The Log-Polar image geometry is motivated because scaling and rotation in Cartesian domain corresponds to pure translation in Log-Polar domain. Moreover, Log-Polar transform of an image patch has high resolution at the center compared to the periphery, which attaches more attention to the target than the surrounding background. Log-polar coordinates in the plane consist of a pair of real numbers (ρ, θ), where ρ is the logarithm of the distance between a given point and the origin and θ is the angle between the reference line (the x axis) and the line through the origin and the point. The formulas for the transformation from Cartesian coordinates to Log-Polar coordinates are given by

$$\rho = \log\sqrt{x^2 + y^2}, \tag{17}$$

$$\theta = \arctan\frac{y}{x}. \tag{18}$$

In the training section, the target patch is extracted according to the estimated target position, scale and orientation. The base rotation sample x_r can be obtained by transforming the target patch to the Log-Polar domain. Considering the properties of the Log-Polar transformation, the rotation template can be trained on all the cyclic shift versions of x_r, denoted by $x_r(\theta), \theta \in \{1, 2, \dots, R\}$. The sample interval is $\Delta = \frac{2\pi}{R}$. Each sample is also assigned with a score generated by a Gaussian function y_r. Similarly, by minimizing the regression error, we get the solution via

$$\hat{\alpha}_r = \frac{\hat{y}_r}{\hat{k}^{x_r x_r} + \lambda}, \tag{19}$$

where $k^{x_r x_r}$ is a vector whose ith element is $\kappa(x_r(i), x_r)$. Since we just care about the shift along the angular coordinate, we consider the logarithmic coordinate as multi-feature channels. In visual tracking scenarios, the base candidate patch is extracted according to the orientation in the last frame. The template matching scores are calculated as

$$y_r = \mathcal{F}^{-1}(\hat{k}^{x_r^\star z_r} \odot \hat{\alpha}_r), \tag{20}$$

where $x_r^\star$ denotes the learned target appearance in the Log-Polar domain. The linear interpolation based on the Gaussian response is also used to compensate for angular sampling errors.

4. Experiments

To evaluate the performance of proposed tracking algorithm, we conduct three different sets of experiments. In the following, we will first introduce some experimental details and settings about our tracker and the visual object tracking (VOT) challenge. After that we will present the first set of

Table 1. Comparisons of accuracy and robustness among correlation filter-based trackers in the baseline experiment.

	accuracy (overlap ratio)					robustness (failure times)				
	KCF	KCF14	SAMF	DSST	RAJSSC	KCF	KCF14	SAMF	DSST	RAJSSC
ball	0.702	0.758	0.775	0.568	0.767	1	1	1	1	0
basketball	0.574	0.645	0.751	0.638	0.621	2	0	0	1	0
bicycle	0.454	0.630	0.618	0.583	0.712	1	0	0	0	0
bolt	0.522	0.490	0.562	0.562	0.705	3	3	2	1	0
car	0.421	0.713	0.512	0.742	0.734	0	0	0	0	0
david	0.746	0.822	0.818	0.807	0.796	0	0	0	0	0
diving	0.233	0.255	0.246	0.442	0.288	5	4	4	1	5
drunk	0.434	0.536	0.569	0.551	0.576	0	0	0	0	0
fernando	0.402	0.411	0.395	0.340	0.468	1	1	1	1	1
fish1	0.438	0.419	0.496	0.321	0.436	3	3	3	1	5
fish2	0.299	0.266	0.299	0.353	0.432	4	6	5	4	2
gymnastics	0.528	0.537	0.538	0.632	0.582	3	1	2	5	1
hand1	0.389	0.563	0.547	0.215	0.596	6	3	3	2	2
hand2	0.438	0.498	0.465	0.528	0.550	8	6	5	6	2
jogging	0.760	0.799	0.822	0.790	0.534	1	1	1	1	1
motocross	0.372	0.366	0.402	0.421	0.661	5	2	4	4	1
polarbear	0.662	0.780	0.709	0.635	0.712	0	0	0	0	0
skating	0.488	0.677	0.452	0.586	0.645	0	1	0	0	0
sphere	0.713	0.90	0.880	0.927	0.738	0	0	0	0	0
sunshade	0.761	0.763	0.759	0.783	0.773	0	0	0	0	0
surfing	0.797	0.805	0.804	0.906	0.821	0	0	0	0	0
torus	0.757	0.857	0.841	0.811	0.791	0	0	0	0	0
trellis	0.546	0.798	0.825	0.808	0.817	0	0	0	0	0
tunnel	0.318	0.687	0.553	0.812	0.718	0	0	0	0	0
woman	0.755	0.744	0.761	0.790	0.653	2	1	1	1	1
MEAN	0.540	0.629	0.616	0.622	0.645	1.80	1.32	1.28	1.16	0.84

experimental results to evaluate the proposed tracker with several competing trackers based on the correlation tracking framework. And then we will present the second set of experimental results to compare our algorithm with other state-of-the-art tracking algorithms. Lastly, we will present the full evaluation results on the VOT2015 challenge.

4.1. Experimental details and settings

We implement the proposed tracker by native Matlab without optimization. All the experiments are conducted on an Intel i7-4770 CPU (3.40 GHz) PC with 8 GB memory. Let J_t denote the scale coefficient of the last frame and $S = 5$ be the number of scale layers. We resize the current image with scale factors $J_t a^l (l \in \{\lfloor -\frac{S-1}{2} \rfloor, \ldots, \lfloor \frac{S-1}{2} \rfloor\})$. Here, $a = 1.02$ restricts the sampling granularity in the scale space. The scale variance of the multivariate Gaussian distribution is $\sigma_s^2 = \frac{\sigma_x^2}{0.021}$. We extract 36 samples from the orientation space and the sample interval is $\Delta = \frac{2\pi}{36}$. The orientation variance of the Gaussian distribution is same as the spatial displacement variance σ_x^2. The learning rates for the appearance model update are changed adaptively to the steepness of the Gaussian response. Other parameters are similar to those employed in KCF [12].

In all the experiments, we use the Pascal VOC overlap ratio (VOR) as the evaluation criteria for accuracy [8]. It is defined as $VOR = \frac{Area(B_T \bigcap B_G)}{Area(B_T \bigcup B_G)}$, where B_T is the tracking bounding box, and B_G is the ground truth bounding box. The larger value means the more accurate result. The robustness is the number of times the tracker failed. A re-initialization is triggered when the overlap drops to zero. Finally, the per-visual attribute normalized AR-rank plot [15] is obtained by ranking trackers with respect to each attribute and averaging the ranking lists.

Table 2. Comparisons of accuracy and robustness among correlation filter-based trackers for the scale change attribute.

	Accuracy (overlap ratio)				Robustness (failure times)			
	KCF14	SAMF	DSST	RAJSSC	KCF14	SAMF	DSST	RAJSSC
baseline	0.580	0.565	0.528	0.592	20	18	15	13
perturbation	0.524	0.518	0.510	0.536	21.80	21.13	17.93	17.87

4.2. Compare with other correlation trackers

To evaluate the performance gain of our fully functional correlation filter based tracker, we compare it with four variants of correlation filter based trackers on the VOT2014 benchmark including KCF [12], KCF14, an enhanced KCF with scale estimation [15], SAMF [17], and DSST [5]. In the baseline experiment, we run trackers on all sequences by initializing them on the ground truth bounding boxes in the first frame. In the experiment with bounding box perturbation, we perform visual tracking with noisy bounding boxes by drawing perturbations uniformly from the $\pm 10\%$ interval of the ground truth bounding box size

Table 3. VOT2014 competition report. The top, second and third lowest average ranks are shown in red, blue and green respectively.

	baseline		region_noise				
	Acc. Rank	Rob. Rank	Acc. Rank	Rob. Rank	Acc. Rank	Rob. Rank	Rank
RAJSSC	4.96	11.49	6.17	11.41	5.56	11.45	8.51
DSST	5.99	12.52	5.96	12.87	5.98	12.69	9.33
SAMF	5.76	14.34	5.65	12.78	5.70	13.56	9.63
KCF	5.51	15.41	5.58	13.05	5.54	14.23	9.89
PLT_14	14.72	6.44	13.74	5.01	14.23	5.73	9.98
DGT	11.67	9.55	9.15	10.11	10.41	9.83	10.12
PLT_13	18.34	3.83	17.38	4.83	17.86	4.33	11.10
eASMS	14.22	13.87	11.42	14.34	12.82	14.11	13.46
HMMTxD	10.28	20.77	9.81	19.57	10.05	20.17	15.11
MCT	16.88	14.08	17.58	13.00	17.23	13.54	15.38
ACAT	13.84	15.23	17.81	14.96	15.82	15.10	15.46
MatFlow	22.00	8.88	19.14	14.64	20.57	11.76	16.17
ABS	20.68	18.62	15.44	15.29	18.06	16.95	17.51
ACT	20.85	16.63	22.27	15.22	21.56	15.92	18.74
qwsEDFT	17.58	19.45	18.64	21.06	18.11	20.25	19.18
LGTv1	29.23	11.78	26.45	9.39	27.84	10.59	19.21
VTDMG	21.48	18.40	20.65	16.98	21.06	17.69	19.38
BDF	23.17	17.93	21.74	18.15	22.45	18.04	20.25
Struck	20.87	21.12	21.51	18.85	21.19	19.99	20.59
DynMS	22.82	19.47	21.51	19.51	22.16	19.49	20.83

and the ± 0.1 radian range.

In these two experiments, our tracker achieves the best performance in both accuracy and robustness. The overlap ratio and the number of tracking failure for each sequence and each tracker in the baseline experiment is shown in Table 1. Compared to the best performance from the other four trackers which provides average failure times of 1.16 and 1.283 in two experiments, our tracker fails respectively 0.84 and 1.197 times. Although the scale and rotation adaptivity is considered in our tracker, it remains stable and reliable. Since dense samples are taken from the joint scale-spatial space and the orientation space, the difference between target patches and noisy background patches is carefully learned. Thus, the tracker is less confused when the external disturbance occurs and is less likely to drift. The average tracking accuracy of our tracker in these two experiments is presented with overlap ratios of 64.5% and 59.0% respectively, while the best results given by the other four trackers are 62.9% and 57.8%. Although they are re-initialized more frequently with the ground truth bounding boxes, their tracking precisions are still poorer than ours.

The scale variation is estimated in these trackers except the general KCF. SAMF models each scale level individually and then makes a comparison among the maximal responses of these scale levels. DSST trains a scale filter after the target position is obtained. The scale adaptation scheme of the enhanced KCF is not given. We model the relationship of features from different scale levels and positions with the block-circulant kernel matrix and get the template matching scores from the joint scale-spatial space. So we infer the target's displacements and scale variations simultaneously and thus obtain obviously improved accuracy. For the scale change attribute, we compare the four trackers in Table 2. Our tracker can well handle the rotation variation of the rigid targets which can be seen from the video *motocross* with higher accuracy and robustness. There is less improvement for roughly deformed targets like the cat in video *fernando*, because the transformations for a non-rigid target are so complex and massive that cyclic shifts are not enough to extract all the examples.

4.3. Compare with other state-of-the-art algorithms

To ensure a fair and unbiased comparison, we use the original results provided by the VOT committee. We compare our approach to recent state-of-the-art algorithms including the winner of the VOT2014 challenge, DSST and other competitive trackers such as PLT_13 [15], Struck [10]. The AR-rank plots and raw plots for the baseline and bounding box perturbation experiments with per-attribute normalization are shown in Figure 1. Table 3 shows the exact per-visual attribute normalized accuracy and robustness ranks of the top twenty trackers. The top performing tracker in robustness is PLT_13, the winner of the VOT2013 challenge. It is an extension of the Struck tracker which uses a structured SVM on gray-scale patches to learn a regression from intensity to center of object displacement. However, PLT_13 applies histogram back-projection as feature selection strategy in the SVM training. The DGT tracker uses superpixels to decompose the target into parts and constructs a graph to represent the structural relationship of target parts. It casts tracking as graph matching across consecutive frames. The complex model ensures its robustness. Since correlation filter based trackers

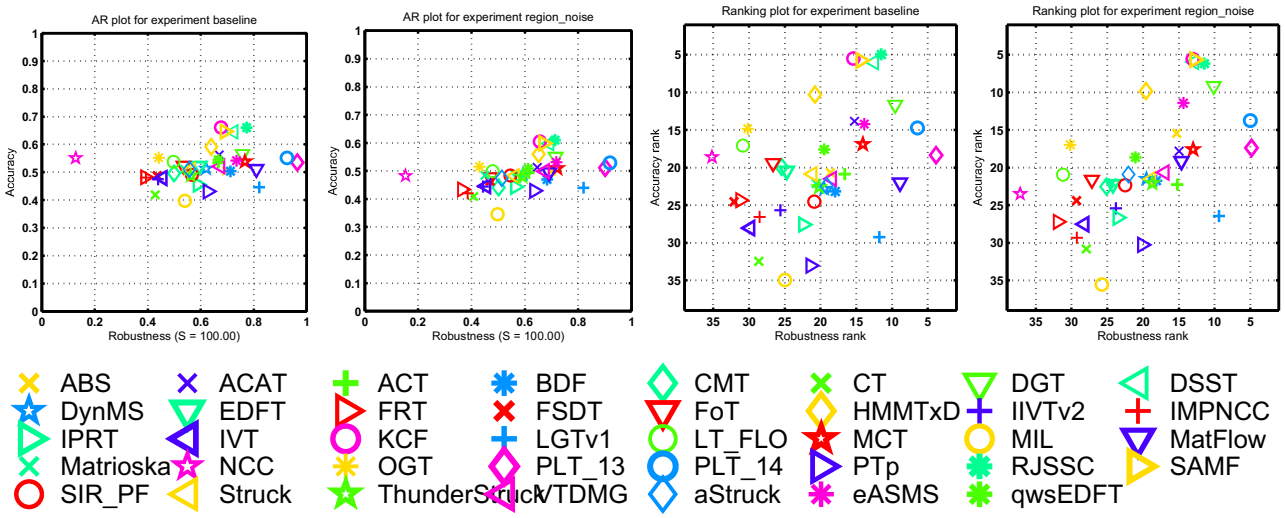

Figure 1. The raw plots and the per-visual attribute AR-rank plots for the baseline and bounding box perturbation experiments. A tracker is competitive if it resides close to the top-right corner of the plot

model the visual tracking as simple template matching problems, their templates are easier contaminated by the background clutters and occlusions. Nonetheless, our tracker still gains some robustness compared with other correlation filter based trackers. In terms of accuracy, the top-performing trackers are these correlation filter based trackers which apply holistic models. They form a cluster in the raw plots and AR-rank plots. It seems dense template matching is useful compared to the sparse sampling in the SVM based trackers. Our tracker ranks first in the baseline experiment while ranks forth in the bounding box perturbation experiment. One reason is that we sacrifice the tracking accuracy for higher robustness. Another reason is that the orientation estimation is taken after the displacements and scale estimation. Once the tracking drift occurs, the base sample in the Log-Polar space maybe changes a lot which accumulate the tracking error.

The raw AR plots for each visual attribute are shown in Figure 2. At illumination changes, the correlation filter based tracker shows their superiority in both accuracy and robustness. The trackers which rely heavily on the color information such as DGT, eASMS show poorer performance. Our tracker gains high accuracy in terms of motion and scale change, while loses some robustness compared with the DGT tracker and the PLT_13 tracker. The adaptability and reliability of the correlation filter based trackers are further enhanced taking account of these three attributes. It seems that our scale and rotation estimation scheme is effective, while the regression framework itself lacks a bit robustness. Moreover, problem is that the occlusion-handling module is hard to be integrated into the correlation filtering framework. The neutral visual attribute does not present particular difficulties in robustness analysis for most trackers, but the tracking accuracy varies apparently.

4.4. VOT2015 challenge results

The VOT2015 challenge database comprises 60 short challenging sequences where the targets undergo severe deformation and external disturbance. The tracking results of our tracker are compared with the NCC tracker and summarized in Table 4. In terms of robustness, we have more failure in video *birds1* and *soccer2*. The reason is that the target is too small for our tracker to capture the structure information by cyclic shifts. For most videos, our tracker fails much less than the NCC tracker, especially when the target is undergoing large deformation like the *gymnastics* and *ice skaters*. For the accuracy estimation, we have an overlap ratio gain of 16% compared with the NCC tracker. The targets' sizes change a lot in video *bag*, *graduate*, *helicopter*, *pedestrian2*, and is well estimated by our trackers. Video *motocross1* and *bmx* prove our rotation adaptation scheme feasible.

Table 4. VOT2015 competition report.

	Acc. Rank	Overlap ratio	Rob. Rank	Failures
RAJSSC	1.15	0.52	1.11	1.63
NCC	1.88	0.36	1.89	10.74

5. Conclusions and future work

In this paper, we have presented a new correlation filter based tracking algorithm designed to perform effective and efficient tracking. The proposed algorithm simultaneously models the spatial displacements, scale variations, and rotation transformations of the object in a unified correlation

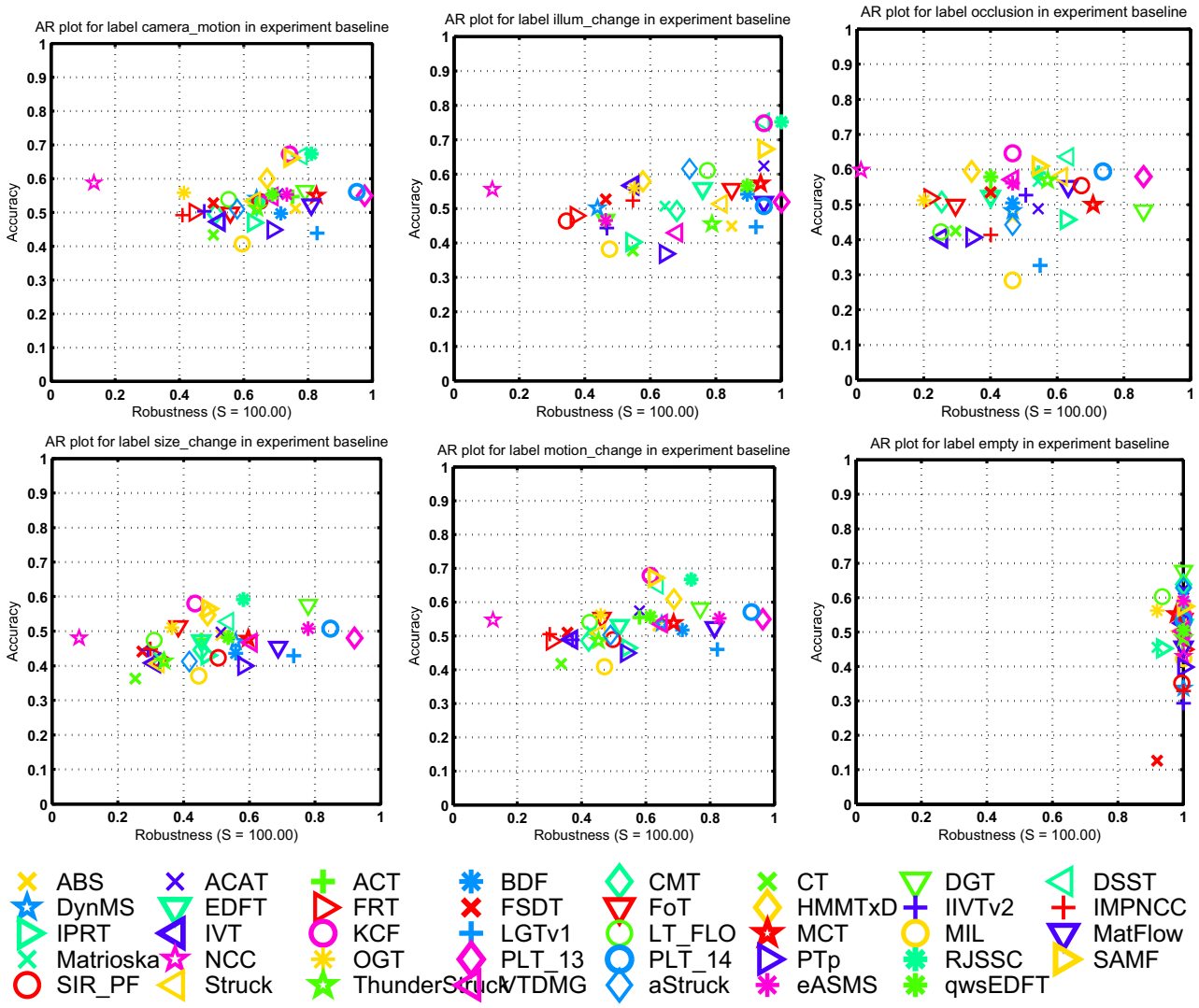

Figure 2. The normalized AR-rank plots for the baseline experiment with respect to the six sequence attributes.

tracking framework, thus can perform joint optimization of the object tracking state. In the VOT2014 Challenge, our tracker demonstrate very competing performance over many state-of-the-art tracking algorithms. In the future, we plan to model the temporal variations of the object appearance in a similar framework.

6. Acknowledgements

This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the Natural Science Foundation of China (Grant No. 61472421 and Grant No. 60935002), the Project Supported by CAS Center for Excellence in Brain Science and Intelligence Technology, and the Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogni.*, pages 798–805, 2006.

[2] B. Babenko, M. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009.

[3] D. Bolme, J. Beveridge, B. Draper, and Y. Lui. Visual object tracking using adaptive correlation filters. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2010.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. Pattern Anal. Mach. Intel.*, 25(5):564–577, 2003.

[5] M. Danelljan, G. Häger, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *Proc. British*

Mach. Vis. Conf., 2014.

[6] M. Danelljan, F. Khan, M. Felsberg, and J. Van de Weijer. Adaptive color attributes for real-time visual tracking. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014.

[7] P. Davis. *Circulant matrices*. American Mathematical Society, New York, 1979.

[8] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vision*, 88(2):303–338, 2010.

[9] J. Gao, J. Xing, W. Hu, and S. Maybank. Discriminant tracking using tensor representation with semi-supervised improvement. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1569–1576, 2013.

[10] S. Hare, A. Saffari, and P. H. S. Torr. Structured output tracking with kernels. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2011.

[11] J. Henriques, R. Caseiro, P. Martins, and J. Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *Proc. Eur. Conf. Comput. Vis.*, 2012.

[12] J. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intel.*, 37(3):583–596, 2015.

[13] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao. Multi-store tracker (MUSTer): A cognitive psychology inspired approach to object tracking. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015.

[14] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Cehovin, G. Nebehay, G. Fernandez, T. Vojíř, A. Gatt, A. Khajenezhad, A. Salahledin, A. Soltani-Farani, A. Zarezade, A. Petrosino, A. Milton, B. Bozorgtabar, B. Li, C. S. Chan, C. Heng, D. Ward, D. Kearney, D. Monekosso, H. C. Karaimer, H. R. Rabiee, J. Zhu, J. Gao, J. Xiao, J. Zhang, J. Xing, K. Huang, K. Lebeda, L. Cao, M. E. Maresca, M. K. Lim, M. ELHelw, M. Felsberg, P. Remagnino, R. Bowden, R. Goecke, R. Stolkin, S. Y. Lim, S. Maher, S. Poullot, S. Wong, S. Satoh, W. Chen, W. Hu, X. Zhang, Y. Li, and Z. Niu. The visual object tracking VOT2013 challenge results. In *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013.

[15] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, T. Vojíř, G. Fernández, A. Lukežič, A. Dimitriev, A. Petrosino, A. Saffari, B. Li, B. Han, C. Heng, C. Garcia, D. Pangeršič, G. Häger, F. S. Khan, F. Oven, H. Possegger, H. Bischof, H. Nam, J. Zhu, J. Li, J. Y. Choi, J.-W. Choi, J. ao F. Henriques, J. van de Weijer, J. Batista, K. Lebeda, K. Öfjäll, K. M. Yi, L. Qin, L. Wen, M. E. Maresca, M. Danelljan, M. Felsberg, M.-M. Cheng, P. Torr, Q. Huang, R. Bowden, S. Hare, S. Y. Lim, S. Hong, S. Liao, S. Hadfield, S. Z. Li, S. Duffner, S. Golodetz, T. Mauthner, V. Vineet, W. Lin, Y. Li, Y. Qi, Z. Lei, and Z. Niu. The visual object tracking VOT2014 challenge results. In *Proc. Eur. Conf. Comput. Vis. Workshops*, 2014.

[16] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015.

[17] Y. Li and J. Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *Proc. Eur. Conf. Comput. Vis. Workshop*, 2014.

[18] Y. Li, J. Zhu, and S. C. Hoi. Reliable patch trackers: Robust visual tracking by exploiting reliable patches. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015.

[19] T. Liu, G. Wang, and Q. Yang. Real-time part-based visual tracking via adaptive correlation filters. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015.

[20] C. Ma, X. Yang, C. Zhang, and M.-H. Yang. Learning a temporally invariant representation for visual tracking. In *Proc. IEEE Int. Conf. Image Process.*, 2015.

[21] K. Murphy. *Machine learning: a probabilistic perspective*. MIT press, London, 2012.

[22] H. Possegger, T. Mauthner, and H. Bischof. In defense of color-based model-free tracking. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015.

[23] J. Shi and C. Tomasi. Good features to track. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pages 593–600, 1994.

[24] S. Song and J. Xiao. Tracking revisited using RGBD camera: Unified benchmark and baselines. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2013.

[25] H. Tao, S. H.S, and R. Kumar. Object tracking with Bayesian estimation of dynamic layer representations. *IEEE Trans. Pattern Anal. Mach. Intel.*, 24(1):75–89, 2002.

[26] Y. Wu, J. Lim, and M. Yang. Online object tracking: A benchmark. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2013.

[27] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13, 2006.

[28] M. Zhang, J. Xing, J. Gao, and W. Hu. Robust visual tracking using joint scale-spatial correlation filters. In *Proc. IEEE Int. Conf. Image Process.*, 2015.