Real-Time Visual Tracking: Promoting the Robustness of Correlation Filter Learning

Yao Sui[1](✉), Ziming Zhang[2], Guanghui Wang[1], Yafei Tang[3], and Li Zhang[4]

[1] Department of EECS, University of Kansas, Lawrence, KS 66045, USA
suiyao@gmail.com, ghwang@ku.edu

[2] Department of ECE, Boston University, Boston, MA 02215, USA
zzhang14@bu.edu

[3] China Unicom Research Institute, Beijing 100032, China
tangyf24@chinaunicom.cn

[4] Department of EE, Tsinghua University, Beijing 100084, China
chinazhangli@tsinghua.edu.cn

Abstract. Correlation filtering based tracking model has received lots of attention and achieved great success in real-time tracking, however, the lost function in current correlation filtering paradigm could not reliably response to the appearance changes caused by occlusion and illumination variations. This study intends to promote the robustness of the correlation filter learning. By exploiting the anisotropy of the filter response, three sparsity related loss functions are proposed to alleviate the overfitting issue of previous methods and improve the overall tracking performance. As a result, three real-time trackers are implemented. Extensive experiments in various challenging situations demonstrate that the robustness of the learned correlation filter has been greatly improved via the designed loss functions. In addition, the study reveals, from an experimental perspective, how different loss functions essentially influence the tracking performance. An important conclusion is that the sensitivity of the peak values of the filter in successive frames is consistent with the tracking performance. This is a useful reference criterion in designing a robust correlation filter for visual tracking.

Keywords: Visual tracking · Correlation filtering · Sparsity regularization · Loss function · Robustness

1 Introduction

In recent years, there is a significant interest in correlation filtering based tracking. Under this paradigm, a correlation filter is efficiently learned online from previously obtained target regions, and the target is located according to the magnitude of the filter response over a large number of target candidates. The main strength of this paradigm is its high computational efficiency, because the target and the candidate regions can be represented in frequency domain and manipulated by fast Fourier transform (FFT), which yields $\mathcal{O}(n \log n)$

B. Leibe et al. (Eds.): ECCV 2016, Part VIII, LNCS 9912, pp. 662–678, 2016.
DOI: 10.1007/978-3-319-46484-8_40

computational complexity for a region of $\sqrt{n} \times \sqrt{n}$ pixels. For this reason, extensive real-time trackers [1–9] have been proposed within the correlation filtering paradigm.

Specifically, a correlation filter is learned from previously obtained target regions to approximate an expected filter response, such that the peak of the response is located at the center of the target region. The response used in previous methods is often assigned to be of Gaussian shaped, which is treated as a continuous version of an impulse signal. For this reason, the learned filter is encouraged to produce Gaussian shaped response. The candidate region with the strongest filter response is determined as the target.

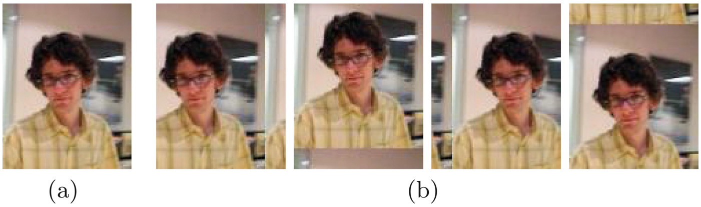


Fig. 1. Illustration of the cyclic shift method. (a) The base image. (b) The cyclic shifts of the base image with ±15 pixels in horizontal and vertical directions, respectively.

Note that the Gaussian shaped response, from a signal processing perspective, is *isotropic*, *i.e.*, all the regions that deviate the same distance away from the center of the target are assigned to the same response values. However, it has been demonstrated that the anisotropic response values can significantly improve the tracking performance from a regression point of view [10,11][1], *e.g.*, using the overlap rates between the training image samples and the target as the response values. Figure 1 illustrates a popular approach for samples generation adopted by previous correlation filtering based trackers [2,7]. It is evident from Fig. 1(b) that the regions of interest are not continuous. This will bring challenges to the correlation filter learning if the response values of the four significantly different regions are enforced to be the same, easily leading to an overfitting.

In addition, from a loss function point of view, the correlation filter is always learned under the squared loss (*i.e.*, ℓ_2-loss) in the previous methods. The choice for the squared loss is limited by the Parseval's identity, by which the learning problem can be exactly transferred into frequency domain. Moreover, the squared loss can lead to a closed-form solution, which guarantees the high computational efficiency. Nevertheless, the target appearance may change significantly during tracking in various challenging situations, such as occlusion and illumination variation. A robust loss function is required to reliably respond to these appearance changes, and avoid the overfitting. The squared loss allows the filter response to fit the expected response with small errors, *i.e.*, stochastically yields Gaussian errors with a small variance. In the presence of significant appearance

[1] The exact equivalence between regression and correlation filtering is proved in [7].

changes, the errors may be extremely large in some feature dimensions, leading to an instability of the squared loss.

Inspired by the previous success, an *anisotropy* of the filter response is exploited in this work by means of an adaptive learning approach via robust loss functions, including ℓ_1-, $\ell_1\ell_2$-, and $\ell_{2,1}$-loss functions. These loss functions will increase the robustness, since they allow large errors in the filter learning in the presence of significant appearance changes. As a result, three real-time trackers are proposed in this study, and it is demonstrated how the loss functions essentially influence the tracking performance. An interesting observation is obtained from the experimental results, which can be taken as a reference criterion in designing a robust correlation filter for visual tracking: the sensitivity of the peak values of the filter in successive frames is consistent with the tracking performance. The proposed algorithms are evaluated by extensive experiments on a popular benchmark [12], and they outperform the competing counterparts.

2 Related Work

Recently, correlation filtering [1] has received much attention in visual tracking. It achieves state-of-the-art tracking performance in terms of both accuracy and running speed. By exploiting the circulant structure [2], visual tracking can be described as a correlation filtering problem, which is also demonstrated to be equivalent to a ridge regression problem [2]. In this paradigm, the cyclic shifts of the latest target region (a base image) is utilized to generate a large number of training samples, essentially as the dense sampling method does, as illustrated in Fig. 1. The cyclic shifts lead to the fact that the sample matrix has a circulant structure. To efficient solve the correlation filtering, the tracking is conducted in frequency domain by fast Fourier transform (FFT) under the Parseval's identity. Because the filter response is considered to be of Gaussian shaped, there is a closed-form solution to the problem of the correlation filter learning. This is why the correlation filtering based tracking methods significantly improve the tracking speed.

There is extensive literature on correlation filtering based tracking methods in recent years. Henriques *et al.* [7] proposed to incorporating the kernel trick with the correlation filter learning, leading to kernelized version of [2]. Since the scale variations of the target appearance between successive frames are not considered in [7], Danelljan *et al.* [4] and Li and Zhu [5] integrated adaptive scale estimations with the correlation filter learning, respectively. An approach leveraging adaptive color attributes [3] was proposed for real-time visual tracking within the correlation filtering framework. Ma *et al.* [13] developed a long-term correlation tracking method by decomposing visual tracking into translation and scale estimations. Liu *et al.* [8] designed an adaptive correlation filter to exploit the part-based information of the target. Tang and Feng [14] proposed a multi-kernel correlation filter for visual tracking, which fully takes advantage of the invariance-discriminative power spectrums of various features. Danelljan *et al.* [9] leveraged a spatial regularization for correlation filter learning, leading to an impressive tracking performance.

Beyond the correlation filter based method, extensive tracking approaches were proposed and achieved state-of-the-art performance, such as structural learning [11,15,16], sparse and low-rank learning [17–20], subspace learning [21–23], and deep learning [24,25]. Readers are recommended to refer to [26,27] for a thorough review of visual tracking.

3 Proposed Approach

3.1 Formulation

The typical correlation filtering based model focuses on solving the following ridge regression problem

$$\min_{\mathbf{w}} \sum_{i} \left(f\left(\mathbf{x}_i\right) - y_i\right)^2 + \lambda \left\|\mathbf{w}\right\|_2^2, \tag{1}$$

where a regression function $f\left(\mathbf{x}_i\right) = \mathbf{w}^T \varphi\left(\mathbf{x}_i\right)$ is trained with a feature-space projector $\varphi\left(\cdot\right)$; the objective values y_i are specified to be of Gaussian shaped; and $\lambda > 0$ is a weight parameter. The training samples $\{\mathbf{x}_i\}$ consists of the cyclically shifted image patches of the base image (*i.e.*, the latest target). With the learned regression model, the target is localized by selecting the candidate with the largest regression value (filter response in the frequency domain) from a set of target candidates that are generated by the cyclically shifted patches of the latest target region in the current frame.

The goal of the proposed approach is to promote the robustness of the correlation filter learning. An anisotropy of the filter response, from a signal processing perspective, is exploited for visual tracking, and the robust loss functions, from a overfitting point of view, are utilized to deal with the significant appearance changes. To this end, an adaptive approach is leveraged in this work, which employs different sparsity related loss functions to adaptively fit the Gaussian shaped objective values. Similar to the previous work [2,7], the proposed approach is modeled from the regression perspective and solved via the correlation filtering method. Generally, the regression in this work is defined as

$$\min_{\mathbf{w}} \sum_{i} \ell\left(f\left(\mathbf{x}_i\right) - y_i\right) + \lambda \left\|\mathbf{w}\right\|_2^2, \tag{2}$$

where $\ell\left(\cdot\right)$ is a loss function, and the regularization $\left\|\mathbf{w}\right\|_2^2$ is reserved to make the regression stable. In order to promote the robustness of the above model against the significant target appearance changes, the sparsity related loss function [28] is encouraged. Three loss functions, ℓ_1-, $\ell_1\ell_2$- and $\ell_{2,1}$-loss, are leveraged in this work, which exploit the sparsity, elastic net and group sparsity structures of the loss values. Note that the problem in Eq. (1) is also described via the above model when the loss function is set as ℓ_2-loss.

3.2 Evaluation Algorithm

The problem in Eq. (2) is NP-hard [28] because the sparsity related constraints on the data fitting term are involved. For this reason, it is equivalently reformulated as

$$\min_{\mathbf{w},\mathbf{e}} \sum_{i} \ell\left(e_i\right) + \lambda \left\|\mathbf{w}\right\|_2^2, \quad s.t.\ e_i = y_i - f\left(\mathbf{x}_i\right), \tag{3}$$

where e_i denotes the difference between the regression values $f\left(\mathbf{x}_i\right)$ and the objective values y_i, and y_i is of Gaussian shaped. Notice that the reformulated problem is convex with respect to either $\mathbf{w}$ or $\mathbf{e}$. However, it is still NP-hard with respect to both $\mathbf{w}$ and $\mathbf{e}$. As a result, an iterative algorithm is required to approximate the solution. First, an equivalent form is employed to represent the above problem as

$$\min_{\mathbf{w},\mathbf{e}} \sum_{i} \left(f\left(\mathbf{x}_i\right) + e_i - y_i\right)^2 + \lambda \left\|\mathbf{w}\right\|_2^2 + \tau \sum_{i} \ell\left(e_i\right), \tag{4}$$

where τ is a weight parameter. Note that Eq. (4) can be split into two subproblems:

$$\min_{\mathbf{w}} \left\|\mathbf{f}\left(\mathbf{X}\right) + \mathbf{e} - \mathbf{y}\right\|_2^2 + \lambda \left\|\mathbf{w}\right\|_2^2 \tag{5}$$

$$\min_{\mathbf{e}} \left\|\mathbf{f}\left(\mathbf{X}\right) + \mathbf{e} - \mathbf{y}\right\|_2^2 + \tau\ell\left(\mathbf{e}\right), \tag{6}$$

where $\mathbf{X}$ denotes the sample matrix, of which each row denotes a sample. Both the above two subproblems have globally optimal solutions. The problem in Eq. (4) can be solved by alternately optimizing the two subproblems until the objective function values converged.

The dual space is leveraged to solve Eq. (5). The dual conjugate of $\mathbf{w}$, denoted by $\boldsymbol{\alpha}$, is introduced, such that $\mathbf{w} = \sum_i \alpha_i \varphi\left(\mathbf{x}_i\right)$. The problem with respect to $\boldsymbol{\alpha}$ is squared. It indicates that there is a closed-form solution

$$\hat{\boldsymbol{\alpha}} = \frac{\hat{\mathbf{y}} - \hat{\mathbf{e}}}{\hat{\mathbf{k}}_1 + \lambda}, \tag{7}$$

where $\mathbf{k}_1$ denotes the first row of the kernel matrix $\mathbf{K}$ whose element $k_{ij} = \varphi^T\left(\mathbf{x}_i\right)\varphi\left(\mathbf{x}_j\right)$, the fraction means element-wise division, and the hat ˆ stands for the discrete Fourier transform (DFT) and hereafter. Note that because the sample matrix $\mathbf{X}$ denotes all the training samples that are generated by cyclically shifting the latest target, some kernels, such as Gaussian, and polynomial, can lead to a circulant kernel matrix, as demonstrated in [7]. Based on such a circulant structure, the kernel matrix $\mathbf{K}$ can be diagonalized as

$$\mathbf{K} = \mathbf{D} diag\left(\hat{\mathbf{k}}_1\right)\mathbf{D}^H, \tag{8}$$

where $\mathbf{D}$ denotes the DFT matrix, $\mathbf{k}_1$ denotes the first row[2] of the kernel matrix $\mathbf{K}$, and $\mathbf{D}^H$ denotes the Hermitian transpose of $\mathbf{D}$. Note that the above diagonalization significantly improves the computational efficiency.

[2] The rows of the kernel matrix $\mathbf{K}$ are actually obtained from the cyclic shifts of the vector $\mathbf{k}_1$.

Three algorithms are employed to solve Eq. (6), corresponding to the three loss functions used in Eq. (4).

(1) ℓ_1-*loss.* In this case, the sparsity constraint is imposed on $\mathbf{e}$. By using the shrinkage thresholding algorithm [29], the globally optimal solution of $\mathbf{e}$ can be obtained from

$$\mathbf{e} = \sigma\left(\frac{1}{2}\tau, \mathcal{F}^{-1}\left(\hat{\mathbf{y}} - \hat{\boldsymbol{\alpha}} \odot \hat{\mathbf{k}}_1\right)\right), \tag{9}$$

where $\mathcal{F}^{-1}(\cdot)$ denotes the inverse Fourier transform, and $\odot$ denotes the element-wise multiplication, and the function σ is a shrinkage operator, defined as

$$\sigma(\varepsilon, x) = sign(x) \max(0, |x| - \varepsilon). \tag{10}$$

(2) $\ell_1\ell_2$-*loss.* In this case, the elastic net constraint is enforced on $\mathbf{e}$. By completing the square, Eq. (6) can be solved in a similar way as using ℓ_1-loss. The globally optimal solution of $\mathbf{e}$ is obtained from

$$\mathbf{e} = \sigma\left(\frac{\tau}{4+2\tau}, \frac{2}{2+\tau}\mathcal{F}^{-1}\left(\hat{\mathbf{y}} - \hat{\boldsymbol{\alpha}} \odot \hat{\mathbf{k}}_1\right)\right). \tag{11}$$

The coefficients of the ℓ_1- and ℓ_2-regularization terms in the elastic net constraint are set to be equal in the experiments.

(3) $\ell_{2,1}$-*loss.* In this case, the variables are considered to be two-dimensional (*i.e.*, matrix variables). Under the $\ell_{2,1}$-loss, the group sparsity of $\mathbf{e}$ is exploited. By using the accelerated proximal gradient method [30], the globally optimal solution of $\mathbf{e}$ is obtained from

$$\mathbf{e}_j = \begin{cases} \left(1 - \frac{1}{\tau\|\mathbf{q}_j\|_2}\right)\mathbf{q}_j, & \frac{1}{\tau} < \|\mathbf{q}_j\|_2 \\ \mathbf{0}, & otherwise, \end{cases} \tag{12}$$

where $\mathbf{e}_j$ denotes the j-th column of the matrix $\mathbf{e}$, and $\mathbf{q} = \mathcal{F}^{-1}\left(\hat{\mathbf{y}} - \hat{\boldsymbol{\alpha}} \odot \hat{\mathbf{k}}_1\right)$. In addition, considering the symmetry of the matrix $\mathbf{e}$, the j-th row of $\mathbf{e}$ is also zeroed for all $j \in \{k|\mathbf{e}_k = \mathbf{0}\}$.

The computational cost in each iteration comes from the Fourier and the inverse Fourier transforms of $\mathbf{e}$, which yield $\mathcal{O}(n \log n)$ complexity. The empirical results in this work show that the algorithm converges within tens of iterations. Thus, the efficiency of the proposed approach can be satisfied for a real-time tracker.

3.3 Target Localization

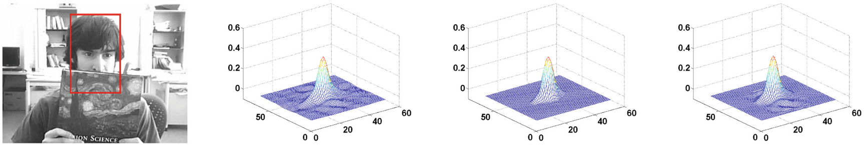
In each frame, a large number of training samples are generated by cyclically shifting the latest target region (a base image), essentially as the dense sampling method does. Given a target candidate $\mathbf{x}'$, the regression value of this candidate is computed in the frequency domain from

$$\hat{\mathbf{f}}(\mathbf{x}') = \hat{\mathbf{k}}' \odot \hat{\boldsymbol{\alpha}}, \tag{13}$$

where $\hat{\mathbf{k}}' = \varphi^T(\mathbf{x})\,\varphi(\mathbf{x}')$ denotes the kernel correlation of the latest target region $\mathbf{x}$ and the candidate region $\mathbf{x}'$. The candidate with the largest regression value (filter response) f is determined as the current target. Note that the above operation in Eq. (13) is actually a spatial correlation filtering over $\mathbf{k}'$ using the filter $\boldsymbol{\alpha}$ in frequency domain, because the frequency representation can lead to significant improvement in the running speed.

3.4 Explanation on the Loss Functions

The different sparsity related loss functions are leveraged in this work, in order to promote the robustness of the filter learning. The ℓ_1-loss allows the errors $\mathbf{e}$ to be extremely large but sparse, such that the learned filter $\boldsymbol{\alpha}$ may ignore the significant appearance changes of the target. The $\ell_1\ell_2$-loss appends an additional ℓ_2-loss to the ℓ_1-loss. Note that because the ℓ_2-loss always leads to small and dense errors, the globally uniform appearance changes, *e.g.*, in the case of illumination variation, can be dealt with effectively. For this reason, the $\ell_1\ell_2$-loss allows for both the abrupt and the slow appearance changes. The $\ell_{2,1}$-loss exploits the relationship between the errors, such that the appearance changes in local patches can be well handled.



(a) (b) (c) (d)

Fig. 2. The anisotropy of the expected filter response exploited in the frame shown in (a) with respect to the ℓ_1-loss (b), the $\ell_1\ell_2$-loss (c), and the $\ell_{2,1}$-loss (d).

As discussed above, the three loss functions can tolerate the large errors during the filter learning, leading to promoted robustness. Referring to Eq. (3), it indicates that the difference e_i between the filter response $f(\mathbf{x}_i)$ and the Gaussian shaped response y_i can be large, leading to an anisotropic expected response $y_i - e_i$. In fact, such an anisotropy essentially facilitates tracking. Figure 2 illustrates the anisotropic expected filter response adaptively learned via the three loss functions in a representative frame. It is evident that the three loss functions result in relatively larger filter responses in the horizontal direction. It suggests that because the distractive object (the book) moves vertically, the loss functions punish the regions vertically deviating away from the target region more severely.

3.5 Implementation Details

The training samples $\mathbf{X}$ in each frame are the fully cyclic shifts of the image region centered at the current target region with the size of 1.5 times of the target. A cosine window is leveraged in the based image to avoid the discontinuity caused by the cyclic shifts. Histogram of orientation gradient (HOG) feature is employed to describe the samples. Gaussian kernel is adopted to map the samples into a non-linear high-dimensional feature space. The above operations are also imposed on the target candidates in each frame, which cyclically shifted from the image centered at the latest target region. As recommended in [7], the parameter λ in Eq. (4) is set to 10^{-4}. Another parameter τ in Eq. (4) is set to be equal to λ in the experiments.

4 Experiments

Three trackers are implemented, corresponding to the ℓ_1-, $\ell_1\ell_2$- and $\ell_{2,1}$-loss functions, denoted by Ours_{S} (sparsity), Ours_{EN} (elastic net), and Ours_{GS} (group sparsity), respectively. The proposed trackers were evaluated on a popular benchmark [12], which contains 51 video sequences with various challenging situations, such as illumination change, non-rigid deformation, and occlusion. The target region in each frame of the 51 video sequences is labeled manually and used as the ground truth. Although many real-time trackers [31–39] have been proposed recently, the 12 most related state-of-the-art trackers, which are publicly provided by the authors, were compared in the experiments. Two criteria of performance evaluation were used in the comparisons, which are defined as follows.

- *Precision.* The percentage of frames where the center location errors (CLE) are less than a predefined threshold. The CLE in each frame is measured by the Euclidean distance between the centers of the tracking and the ground truth regions.
- *Success Rate.* The percentage of frames where the overlap rates (OR) are greater than predefined threshold. The OR in each frame is computed from $\frac{A_t \bigcap A_g}{A_t \bigcup A_g}$ for A_t and A_g are the areas of the tracking and the ground truth regions, respectively.

4.1 Comparison with the State-of-the-Art Trackers

We compared the proposed trackers to the top five trackers in [12], including Struck [11], SCM [40], TLD [41], ASLA [42], and CXT [43]. Figure 3 shows the precision plots and success rate plots of the proposed and the top five trackers in [12] on the 51 video sequences. It is evident that the proposed trackers significantly outperform the top five trackers, yielding the improvements of 14 % in precision ($\rho = 20$) and 5 % in success rate (average). This is attributed to the advantage of the correlation filtering paradigm.

4.2 Comparison with Trackers Within Correlation Filter Learning

It is also desired to investigate the performance of the proposed trackers within the correlation filtering based methods. 7 popular correlation filtering based

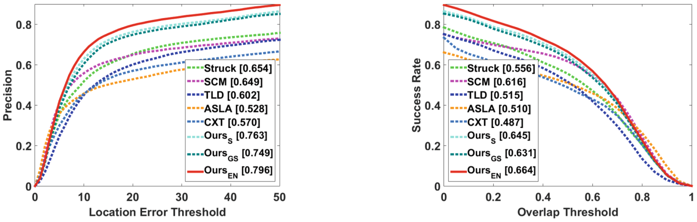


Fig. 3. Tracking performance of the proposed and the top ten trackers in [12] on the 51 video sequences.

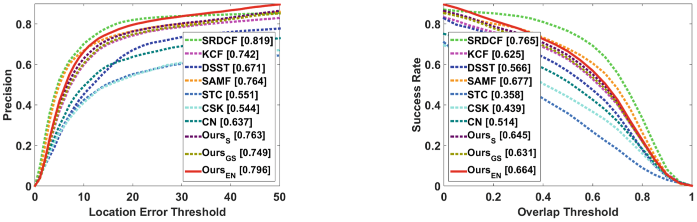


Fig. 4. Tracking performance of the proposed and the popular correlation filtering based trackers on the 51 video sequences.

trackers are referred to for the comparisons. Figure 4 shows the precision plots and success rate plots of the proposed and the 7 correlation filtering based trackers on the 51 video sequences.

It is evident that the proposed tracker, Ours$_{\mathrm{EN}}$, obtains the second best results in terms of precisions. This is attributed to the robustness of the $\ell_1\ell_2$-loss used in the correlation filter learning. Because the scale variation is not considered, the performance of the proposed trackers is inferior to the SMAF tracker and the SRDCF tracker in terms of success rate, which adopt ad hoc strategies to deal with the scale change. It is also necessary to note that the success rate of the proposed trackers are still superior to other five competing counterparts (except for SMAF and SRDCF), because of the improved accuracy on the target localization.

The computational efficiency should be compared within the correlation filtering based methods for a fair evaluation. Table 1 shows the running speeds (in frames per second) of the proposed and the popular correlation filtering based trackers. Because the iterative algorithm can converge within tens of iterations, the proposed trackers run faster than their counterparts like SRDCF, SMAF, and DSST.

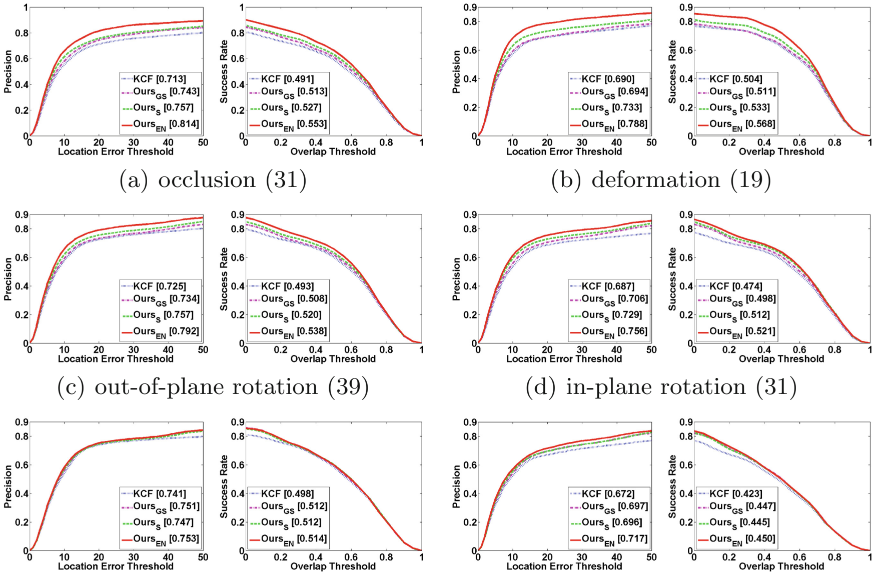
4.3 Evaluations in Various Situations

The tracking performance in various challenging situations is analyzed to thoroughly evaluate the proposed trackers. In order to investigate the effectiveness of

Table 1. Running speeds (in frames per second) of the proposed and the popular correlation filtering based trackers.

Tracker	Ours	SRDCF [9]	SAMF [5]	DSST [4]	KCF [7]	CN [44]	CSK [2]	STC [6]
FPS	37	5	15	25	172	135	154	181

the three (ℓ_1-, $\ell_1\ell_2$- and $\ell_{2,1}$-) loss functions, the KCF tracker (ℓ_2-loss) is referred to as the base line method. Figure 5 shows the results in the six challenging situations, respectively.



(a) occlusion (31)

(b) deformation (19)

(c) out-of-plane rotation (39)

(d) in-plane rotation (31)

(e) illumination change (25)

(f) scale variation (28)

Fig. 5. Tracking performance of the three proposed trackers and the KCF tracker on various challenging situations. In the caption of each sub-figure, the number in parentheses denotes the number of the video sequences in the corresponding situation.

Occlusion. In the case of occlusion, the target is occluded by the other objects, leading to abrupt appearance changes. It is evident that the three proposed trackers significant outperform the KCF tracker in this case. This is attributed to that the sparsity related loss functions of the proposed trackers are more robust to the abrupt appearance changes than the squared loss, resulting in more reliable filter response.

Deformation. The target suffers from the non-rigid deformation in some complicated factors, like motion, pose change, and viewpoint variation. In this case, the target appearance often partially changes significantly. Note that the sparsity related loss functions work more robustly in the presence of significant appearance change, while the squared loss is more effective to deal with globally uniform appearance change. It is evident from the results that the proposed tracker, Ours_{EN}, achieves the best results in this case, because the significant changes are well handled by its sparsity constraint and the small changes are dealt with by its squared regularization. Ours_{S} also obtains better results than KCF. In contrast, the $\ell_{2,1}$-loss (Ours_{GS}) does not improve the results obviously because of its sensitivity to this complicated situation.

In-Plane/Out-of-Plane Rotation. This challenge is often caused by the target motion and/or viewpoint change. It is evident that the three proposed trackers improve the KCF tracker to different extent. This benefits from the robustness of the sparsity based loss functions.

Illumination Change. In this case, the target appearance changes as the lighting condition of the scene varies. This challenge often causes uniform changes in target appearance, *i.e.*, the illumination change influences in the entire target appearance. For this reason, the squared loss is very efficient to deal with this case. As a result, the proposed approach does not improve the KCF tracker significantly.

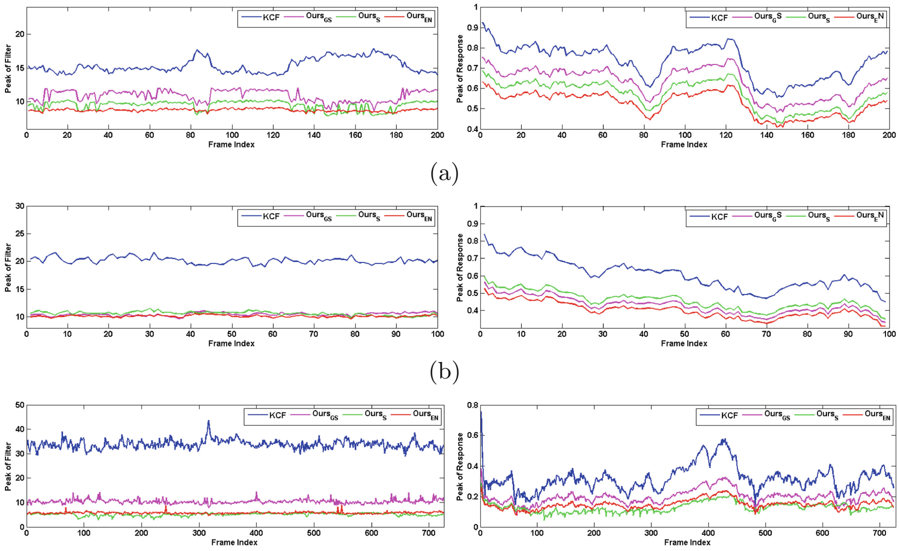
Scale Variation. During tracking, the scale of the target appearance is inevitably changed. If the tracker does not adjust the size of the target window appropriately, tracking failure will be possibly caused because more background information is unexpectedly acquired by the tracker. Unfortunately, considering the efficiency, the KCF and the proposed trackers does not deal with the scales. It can be seen from the results that the proposed trackers improve the precisions but obtain similar success rates as the KCF tracker in this situation.

4.4 Analysis of the Proposed Approach

The goal of the proposed approach is to improve the robustness of the correlation filter learning by means of different loss functions. The different loss functions lead to different anisotropic filter responses. In this section, we interpret how the different loss functions essentially influence the tracking performance via the anisotropic filter responses.

Intuitively, the peak value of an online learned correlation filter, which is responsible for the accuracy of the target localization in each frame, should be stable enough between successive frames in the presence of various challenges. To this end, we analyze the peak values of the filter on three representative video sequences, which include the challenges of occlusion, illumination change, and deformation, respectively. Because there are also other challenges on the video sequences of *faceocc2* and *david*, only the first 200 and 100 frames are selected, respectively. For the convenience of discussion, we use the KCF tracker [7] as a baseline method in the analysis.

Qualitatively, Fig. 6 plots the peak value of the correlation filter and the filter responses, respectively, in each frame of the three video sequences with respect to the KCF and the proposed trackers. Note that the abrupt changes of the peak values correspond to the significant appearance changes in the corresponding



(a)

(b)

(c)

Fig. 6. Peak values of the online learned filters (left column) and the responses (right column) with respect to the four trackers in different challenging cases. The curves are expected to be as smooth as possible. (a) occlusion (first 200 frames of *faceocc2*); (b) illumination change (first 100 frames of *david*); and (c) deformation (all the 725 frames of *basketball*).

frames. If the peak values are sensitive in successive frames, the corresponding filter responses will be unstable, leading to lower accuracy of target location.

In the case of occlusion, as shown in Fig. 6(a), the peak values of the filter with respect to the $\ell_1\ell_2$-loss (Ours$_{\mathrm{EN}}$) varies the most slowly between the frames. The ℓ_1-loss (Ours$_{\mathrm{S}}$) also achieves smoother plot of the peak values than the $\ell_{2,1}$- (Ours$_{\mathrm{GS}}$) and the ℓ_2-loss (the KCF tracker). The analysis results on the sensitivity of the peak values in the successive frames are consistent with the tracking performance evaluations shown in Fig. 5(a).

In the case of illumination change, it is evident in Fig. 6(b) that the peak values of the filter with respect to the four trackers have the similar sensitivities in the successive frames. It is also verified in Fig. 5(e) that the four trackers achieve similar tracking performance.

When non-rigid deformation is involved, as shown in Fig. 6(c), the proposed tracker, Ours$_{\mathrm{EN}}$, produces the most stable filter peak values, achieving the best tracking performance, as shown in Fig. 5(b). In contrast, the peak values of the filter with respect to the KCF tracker are very sensitive between the frames, resulting in the interior tracking performance, as evaluated in Fig. 5(b).

Quantitatively, a metric is required to measure the sensitivity of the filter. It is discussed in [1] that, from a signal processing perspective, a good correlation filter often has a large peak-to-sidelobe ratio (PSR) value. The PSR only

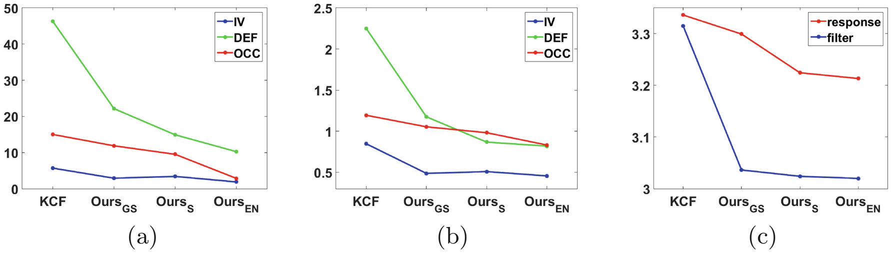


Fig. 7. Sensitivity of the peak values of the filters (a) and the responses (b) with respect to the four trackers in different challenging cases. (c) Average sensitivity of the peak values of the filters and the responses with respect to the four trackers on all the 51 video sequences.

considers the performance of the filter in one frame, while a measurement focusing on the performance in successive frames is more desired for tracking analysis. To this end, the following metric is defined, from a visual tracking point of view, to measure the sensitivity of a correlation filter:

$$s = \sum_{i=1}^{n} (p_i - p_m)^2, \tag{14}$$

where p_i denotes the peak value of the correlation filter in the i-th frame, p_m denotes the mean of the peak values in the n frames, and the n peak values are normalized by their squared norm. As discussed above, the value of s is expected to be small for a good correlation filter.

Figures 7(a) and (b) plot the sensitivity s of the learned correlation filters and the filter responses with respect to the four trackers in the above three challenging situations, respectively. To thoroughly verify the sensitivity, in Fig. 7(c), we show the average sensitivity of the filter and the response on all the 51 video sequences. It is evident that the sensitivity analysis of the correlation filter in successive frames is consistent with the tracking performance evaluations (refer to the results shown in Fig. 5).

From both the qualitative and the quantitative analysis, a conclusion can be drawn to explain how the loss functions essentially influence the tracking performance: the lower the sensitivity s of the learned correlation filter in successive frames is, the higher the tracking performance is achieved. This also can be used as a criterion to design a robust correlation filter for visual tracking.

Revisiting the proposed approach, because the sparsity related loss functions allow large errors in the correlation filter learning, the appearance changes will not cause the significant changes in filter peak values, leading to low sensitivity values. In contrast, the squared loss used by the KCF tracker enforces small errors in the correlation filter learning, such that the filter is always adjusted to fit all the small appearance changes, leading to high sensitivity values. This explains why the proposed trackers perform better than the KCF tracker from the sensitivity perspective.

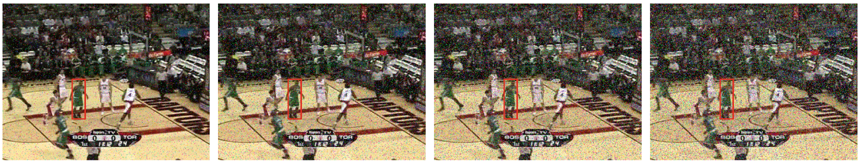
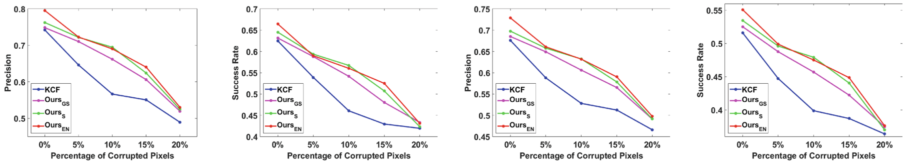


Fig. 8. Representative frames with 5 %, 10 % , 15 % and 20 % corrupted pixels (from left to right).



(a) (b) (c) (d)

Fig. 9. Tracking performance of the three proposed trackers and the KCF tracker in the presence of noise with different amounts on the 51 video sequences. (a) precision plots with $\theta = 20$; (b) success plot with $\rho = 0.5$; (c) precision plots in average; and (d) success plot in average.

4.5 Tracking in Noise Contaminated Frames

In the practical applications, the quality of the video sequences cannot be guaranteed, *i.e.*, the frames are often corrupted by noise. Thus, a visual tracker is expected to be robust to the noise contaminated frames. For this reason, to thoroughly evaluate the robustness of the proposed approach, the tracking is investigated in the noise contaminated frames. The representative noise contaminated frames are shown in Fig. 8. Figure 9 shows the tracking performance of the three proposed trackers and the KCF tracker in the presence of noise with different amounts. It is evident that, in the case that even small number of pixels are corrupted, the performance of the KCF tracker decreases significantly. In contrast, the proposed trackers are not influenced by the noise so drastically as the KCF tracker. The performance of the proposed trackers decreases sharply until a relative large number of pixels (20 %) are corrupted. As a result, it can be observed that the proposed trackers perform more robustly than the KCF tracker in the noise contaminated frames. This also suggests that the proposed approach is closer to the practical applications.

5 Conclusion

Three real-time trackers have been proposed in this work within the correlation filtering paradigm. The robustness of the filter learning has been successfully promoted by employing three sparsity related loss functions. It has been shown that the tracking performance in various challenging situations has been improved via the proposed approach. Through analyzing how the different loss functions essentially influenced the tracking performance, we have found that the analysis result on the sensitivity of the peak values of the filter is consistent with the tracking performance evaluations. This is a very useful reference criterion to design a robust correlation filter for tracking.

Acknowledgment. This work is partly supported by the National Natural Science Foundation of China (NSFC) under grants 61573351 and 61132007, the National Aeronautics and Space Administration (NASA) LEARN II program under grant number NNX15AN94N, and the joint fund of Civil Aviation Research by the National Natural Science Foundation of China (NSFC) and Civil Aviation Administration under grant U1533132.

References

1. Bolme, D., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
2. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 702–715. Springer, Heidelberg (2012)
3. Danelljan, M., Khan, F.S., Felsberg, M., Weijer, J.V.D.: Adaptive color attributes for real-time visual tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1090–1097 (2014)
4. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: British Machine Vision Conference (BMVC) (2014)
5. Li, Y., Zhu, J.: A Scale adaptive kernel correlation filter tracker with feature integration. In: European Conference on Computer Vision Workshop (2014)
6. Zhang, K., Zhang, L., Liu, Q., Zhang, D., Yang, M.-H.: Fast visual tracking via dense spatio-temporal context learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 127–141. Springer, Heidelberg (2014)
7. Henriques, J., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **37**(3), 583–596 (2015)
8. Liu, T., Wnag, G., Yang, Q.: Real-time part-based visual tracking via adaptive correlation filters. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4902–4912 (2015)
9. Danelljan, M., Gustav, H., Khan, F.S., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: IEEE International Conference on Computer Vision (ICCV), pp. 4310–4318 (2015)
10. Zhang, S., Zhao, S., Sui, Y., Zhang, L.: Single object tracking with fuzzy least squares support vector machine. IEEE Trans. Image Process. (TIP) **24**(12), 5723–5738 (2015)
11. Hare, S., Saffari, A., Torr, P.: Struck: structured output tracking with kernels. In: IEEE International Conference on Computer Vision (ICCV), pp. 263–270 (2011)
12. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2411–2418 (2013)

13. Ma, C., Yang, X., Zhang, C., Yang, M.h.: Long-term correlation tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5388–5396 (2015)
14. Tang, M., Feng, J.: Multi-kernel correlation filter for visual tracking. In: IEEE International Conference on Computer Vision (ICCV), pp. 3038–3046 (2015)
15. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **34**(7), 1409–1422 (2012)
16. Zhang, T., Liu, S., Xu, C., Yan, S., Ghanem, B., Ahuja, N., Yang, M.H.: Structural sparse tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 150–158 (2015)
17. Mei, X., Ling, H.: Robust visual tracking and vehicle classification via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **33**(11), 2259–2272 (2011)
18. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Low-rank sparse learning for robust visual tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 470–484. Springer, Heidelberg (2012)
19. Sui, Y., Tang, Y., Zhang, L.: Discriminative low-rank tracking. In: IEEE International Conference on Computer Vision (ICCV), pp. 3002–3010 (2015)
20. Sui, Y., Zhang, L.: Robust tracking via locally structured representation. Int. J. Comput. Vis. (IJCV) **119**(2), 110–144 (2016)
21. Kwon, J., Lee, K.: Visual tracking decomposition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1269–1276 (2010)
22. Wang, D., Lu, H., Yang, M.H.: Least soft-thresold squares tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2371–2378 (2013)
23. Sui, Y., Zhang, S., Zhang, L.: Robust visual tracking via sparsity-induced subspace learning. IEEE Trans. Image Process. (TIP) **24**(12), 4686–4700 (2015)
24. Ma, C., Huang, J.b., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: IEEE International Conference on Computer Vision (ICCV), pp. 3074–3082 (2015)
25. Wang, L., Ouyang, W., Wang, X., Lu, H.: Visual tracking with fully convolutional networks. In: IEEE International Conference on Computer Vision (ICCV), pp. 3119–3127 (2015)
26. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. ACM Comput. Surv. **38**(4), 13–57 (2006)
27. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: an experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **36**(7), 1442–1468 (2014)
28. Wright, J., Ma, Y., Mairal, J., Sapiro, G.: Sparse representation for computer vision and pattern recognition. Proc. IEEE **98**(6), 1031–1044 (2010)
29. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
30. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Convex optimization with sparsity-inducing norms. In: Optimization for Machine Learning, pp. 1–35 (2011)
31. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. Br. Mach. Vis. Conf. (BMVC) **6**(1–6), 10 (2006)
32. Zhang, K., Zhang, L., Yang, M.H.: Real-time object tracking via online discriminative feature selection. IEEE Trans. Image Process. (TIP) **22**(12), 4664–4677 (2013)

33. Zhang, K., Zhang, L., Yang, M.-H.: Real-time compressive tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 864–877. Springer, Heidelberg (2012)
34. Li, H., Shen, C., Shi, Q.: Real-time visual tracking using compressive sensing. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
35. Hall, D., Perona, P.: Online, real-time tracking using a category-to-individual detector. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part I. LNCS, vol. 8689, pp. 361–376. Springer, Heidelberg (2014)
36. Wu, Y., Cheng, J., Wang, J., Lu, H.: Real-time visual tracking via incremental covariance tensor learning. In: IEEE International Conference on Computer Vision (ICCV) (2009)
37. Bao, C., Wu, Y., Ling, H., Ji, H.: Real time robust L1 tracker using accelerated proximal gradient approach. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1830–1837, June 2012
38. Holzer, S., Pollefeys, M., Ilic, S., Tan, D.J., Navab, N.: Online learning of linear predictors for real-time tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 470–483. Springer, Heidelberg (2012)
39. Hager, G.D., Belhumeur, P.N.: Real-time tracking of image regions with changes in geometry and illumination. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 403–410 (1996)
40. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparsity-based collaborative model. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1838–1845 (2012)
41. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N learning: Bootstrapping binary classifiers by structural constraints. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 49–56., June 2010
42. Jia, X., Lu, H., Yang, M.H.: Visual tracking via adaptive structural local sparse appearance model. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1822–1829 (2012)
43. Dinh, T.B., Vo, N., Medioni, G.: Context tracker: exploring supporters and distracters in unconstrained environments. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1177–1184, June 2011
44. Danelljan, M., Khan, F.S., Felsberg, M., Weijer, J.V.D.: Adaptive color attributes for real-time visual tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2014)