Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 2437521, 11 pages
https://doi.org/10.1155/2019/2437521

Research Article

Discriminative Fusion Correlation Learning for Visible and Infrared Tracking

Xiao Yun, Yanjing Sun, Xuanxuan Yang, and Nannan Lu

China University of Mining and Technology, School of Information and Control Engineering, 1 Daxue Road, Xuzhou, Jiangsu 221116, China

Correspondence should be addressed to Yanjing Sun; yjsun@cumt.edu.cn

Received 22 January 2019; Revised 2 April 2019; Accepted 18 April 2019; Published 15 May 2019

Academic Editor: Volodymyr Ponomaryov

Discriminative correlation filter- (DCF-) based trackers are computationally efficient and achieve excellent tracking in challenging applications. However, most of them suffer low accuracy and robustness due to the lack of diversity information extracted from a single type of spectral image (visible spectrum). Fusion of visible and infrared imaging sensors, one of the typical multisensor cooperation, provides complementarily useful features and consistently helps recognize the target from the background efficiently in visual tracking. Therefore, this paper proposes a discriminative fusion correlation learning model to improve DCF-based tracking performance by efficiently combining multiple features from visible and infrared images. Fusion learning filters are extracted via late fusion with early estimation, in which the performances of the filters are weighted to improve the flexibility of fusion. Moreover, the proposed discriminative filter selection model considers the surrounding background information in order to increase the discriminability of the template filters so as to improve model learning. Extensive experiments showed that the proposed method achieves superior performances in challenging visible and infrared tracking tasks.

1. Introduction

Visual tracking has received widespread attention for its extensive applications in video surveillance, autonomous driving and human-machine interaction, military attack, robot vision, etc. [1, 2]. Depending on the appearance model, existing tracking algorithms can be categorized into two categories: generative and discriminative tracking. Generative tracking algorithms build a target model and search for the candidate image patch with maximal similarity. For example, Wang et al. [3] proposed a novel regression-based object tracking framework which successfully incorporates Lucas and Kanade algorithm into an end-to-end deep learning paradigm. Chi et al. [4] trained a dual network with random patches measuring the similarities between the network activation and target appearance to leverage the robustness of visual tracking. On the contrary, the goal of discriminative algorithms is to learn a classifier to discriminate between its appearance and that of the environment given an initial image patch containing the target. Yang et al. [5] proposed a temporal restricted reverse-low-rank learning algorithm for visual tracking to jointly represent target and background templates via candidates, which exploits the low-rank structure among consecutive target observations and enforces the temporal consistency of target in a global level. A new peak strength metric [6] is proposed to measure the discriminative capability of the learned correlation filter that can effectively strengthen the peak of the correlation response, leading to more discriminative performance than previous methods.

Besides these efforts, other researchers have worked on tracking methods that are both generative and discriminative. For instance, Zhang et al. [7] obtained an object likelihood map to adaptively regularize the correlation filter learning by suppressing the clutter background noises while making full use of the long-term stable target appearance information. Qi et al. [8] proposed a structure-aware local sparse coding algorithm, which encodes a target candidate using templates with both global and local sparsity constraints, and also obtains a more precise and discriminative sparse representation to account for appearance changes. In [9], an adaptive set of filtering templates is learned to alleviate drifting problem of tracking by carefully selecting object candidates in different situations to jointly capture the target appearance variations. Moreover, a variety of simple yet effective features are effectively integrated into the learning process of filters to further improve the discriminative power of the filters. In the salient-sparse-collaborative tracker [10], an object salient feature map is built to create a salient-sparse discriminative model and a salient-sparse generative model to both handle the appearance variation and reduce tracking drifts effectively. A multilayer convolutional network-based visual tracking algorithm based on important region selection [11] is proposed to build high entropy selection and background discrimination models and to obtain the feature maps by weighting the template filters with cluster weights, which enables the training samples to be informative in order to provide enough stable information and also be discriminative so as to resist distractors. Generally speaking, discriminative and generative methods have complementary advantages in appearance modeling, and the success of a visual tracking method depends not only on its representation ability against appearance variations but also on the discriminability between target and background, thus leading to the requirement of a more robust training model [12].

Recently, discriminative correlation filter- (DCF-) based visual tracking methods [13–18] have shown excellent performances on real-time visual tracking for its advantage of robustness and computational efficiency. The DCF-based methods work by learning an optimal correlation filter used to locate the target in the next frame. The significant gain in speed is obtained by exploiting the fast Fourier transform (FFT) at both learning and detection stages [14]. Bolme et al. [13] presented an adaptive correlation filter, named Minimum Output Sum of Squared Error (MOSSE) filter, which produces stable correlation filters by optimizing the output sum of squared error. Based on MOSSE, Danelljan et al. [14, 15] proposed a novel scale adaptive tracking approach by learning separate discriminative correlation filters for translation and scale estimation, which achieves accurate and robust scale estimation in a tracking-by-detection framework. Galoogahi et al. [16] proposed a computationally efficient Background-aware correlation filter-based on handcrafted features that can efficiently model how both the foreground and background of the object varies over time. The work in [17] reformulates DCFs as a one-layer convolutional neural network composed of integrates feature extraction, response map generation, and model update with residual learning. Johnander et al. [18] proposed a unified formulation for learning a deformable convolution filter in which the deformable filter is represented as a linear combination of subfilters, and both the subfilter coefficients and their relative locations are inferred jointly in our formulation. However, the above trackers fail when the target undergoes severe appearance changes due to limited data supplied by single features.

Multiple feature fusion contains more useful information than single feature, thus providing higher precision, certainty, and reliability for visual tracking. Wu et al. [19] proposed a data fusion approach via sparse representation with applications to robust visual tracking. Uzkent et al. [20] proposed an adaptive fusion tracking method that combines likelihood maps from multiple bands of hyperspectral imagery into one single more distinctive representation, which increases the margin between mean value of foreground and background pixels in the fused map. Chan et al. [21] proposed a robust adaptive fusion tracking method, which incorporates a novel complex cell into the group of object representation to enhance the global distinctiveness. Feature fusion also achieves superior performances on correlation filter-based tracking. For example, Rapuru et al. [22] proposed a robust tracking algorithm by efficiently fusing tracking, learning, and detection with the systematic model update strategy of kernelized correlation filter tracker.

Although much efforts have been made, single-sensor feature fusion-based tracking suffer low accuracy and robustness due to the lack of diversity information. Fusion of visible and infrared sensors, one of the typical multisensor cooperations, provides complementarily useful features, which is able to achieve a more robust and accurate tracking result [23]. Li et al. [24] designed a fusion scheme containing joint sparse representation and colearning update model to fuse color visual spectrum and thermal spectrum images for object tracking. Li et al. [25] proposed an adaptive fusion scheme based on collaborative sparse representation in Bayesian filtering framework for online tracking. Mangale and Khambete [26] developed reliable camouflaged target detection and tracking system using fusion of visible and infrared imaging. Yun et al. [23] proposed a compressive time-space Kalman fusion tracking with time-space adaptability for visible and infrared images and introduced extended Kalman filter to update fusion coefficients optimally. A visible and infrared fusion tracking algorithm based on multiview multikernel fusion model is presented in [27]. Zhang et al. [28] transferred visible tracking data to infrared data to obtain better tracking performances. Lan et al. [29] proposed joint feature learning and discriminative classifier framework for multimodality tracking, which jointly eliminate outlier samples caused by large variations and learn discriminability-consistent features from heterogeneous modalities. Li et al. [30] proposed a convolutional neural network architecture including a two-stream ConvNet and a FusionNet, which proves that tracking with visible and infrared fusion outperforms that with single sensor in terms of accuracy and robustness.

DCF-based trackers have significant low computational load and are especially suitable for a variety of real-time challenging applications. However, most of the DCF-based trackers suffer low accuracy and robustness due to the lack of diversity information extracted from a single type of spectral image (visible spectrum). Therefore, this paper proposes a discriminative fusion correlation learning model to improve DCF-based tracking performance by combining multiple features from visible and infrared imaging sensors. The main contributions of our work are summarized as follows:

(i) A discriminative fusion correlation learning model is presented to fuse visible and infrared features such that valuable information from all sensors is preserved.

(ii) The proposed fusion learning filters are obtained via late fusion with early estimation, in which the

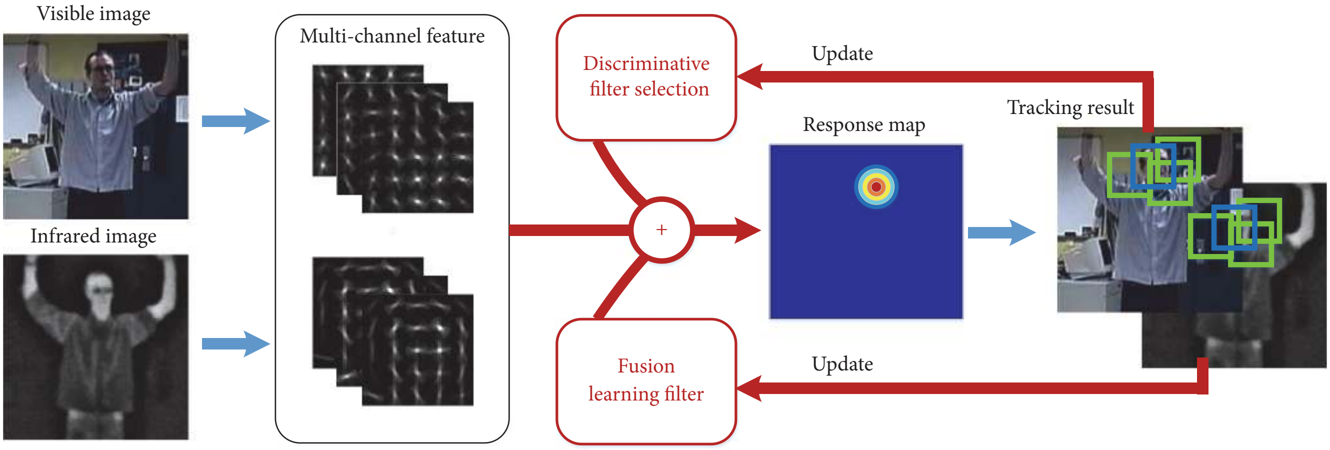


FIGURE 1: Discriminative fusion correlation learning for visible and infrared tracking. Blue and green boxes denote the target and background samples extracted from the tracking result, respectively.

performances of the filters are weighted to improve the flexibility of fusion.

(iii) The proposed discriminative filter selection model considers the surrounding background information in order to increase the discriminability of the template filters so as to improve model learning.

The remainder of this paper is organized as follows. In Section 2, the multichannel discriminative correlation filter is introduced. In Section 3, we describe our work in detail. The experimental results are presented in Section 4. Section 5 concludes with a general discussion.

2. Multichannel Discriminative Correlation Filter

Multichannel DCF provides superior robustness and efficiency in dealing with challenging tracking tasks [14]. In the multichannel DCF-based tracking algorithm, d channel Histogram of Oriented Gradient (HOG) features [14] from the target sample s are extracted to maintain diverse information. During training process, the goal is to learn correlation filter h, which is achieved by minimizing the error of the correlation response compared to the desired correlation output g as

$$\varepsilon = \left\| g - \sum_{l=1}^{d} h^l * s^l \right\|^2 + \lambda \sum_{l=1}^{d} \left\| h^l \right\|^2, \quad (1)$$

where $*$ denotes circular correlation and λ is the weight parameter [14]. s^l and h^l ($l = 1, \cdots, d$) are the l-th channel feature and the corresponding correlation filter, respectively. The correlation output g is supposed to be a Gaussian function with a parametrized standard deviation [14].

The minimization of (1) can be solved by minimizing (2) in the Fourier domain as

$$H^l = \frac{\overline{G} S^l}{\sum_{k=1}^{d} \overline{S}^k S^k + \lambda}, \quad l = 1, \cdots, d, \quad (2)$$

where H, G, and S are the discrete Fourier transform (DFT) of h, g, and s, respectively. The symbol bar $\bar{}$ denotes complex conjugation. The multiplications and divisions in (2) are performed pointwise. The numerator A_t^l and denominator B_t of the filter H_t^l in (2) are updated as

$$\begin{aligned} A_t^l &= (1 - \eta) A_{t-1}^l + \eta \overline{G} S_t^l, \quad l = 1, \cdots, d, \\ B_t &= (1 - \eta) A_{t-1} + \eta \sum_{k=1}^{d} \overline{S}^k S^k, \end{aligned} \quad (3)$$

where η is a learning rate parameter.

During tracking process, the DFT of the correlation score y_t of the test sample z_t is computed in the Fourier domain as

$$Y_t = \frac{\sum_{l=1}^{d} \overline{A}_{t-1}^l Z_t^l}{B_{t-1} + \lambda}, \quad (4)$$

where Y_t and Z_t^l are the DFTs of y_t and z_t^l, respectively. A_t^l and B_t are the numerator and denominator of the filter updated in the previous frame, respectively. Then the correlation scores y_t is obtained by taking the inverse DFT $y_t = \mathcal{F}^{-1}\{Y_t\}$. The estimate of the current target state is obtained by finding the maximum correlation score among the test samples.

3. Proposed Discriminative Fusion Correlation Learning

In this section, we introduce the tracking framework of the proposed algorithm, the general scheme of which is described in Figure 1. Firstly, the multichannel features are extracted, respectively, from the visible and infrared images according to [14]. Secondly, the proposed discriminative filter selection and the fusion filter learning are applied to get the fusion response map. Finally, the discriminative filters and fusion filters are updated via the tracking result obtained by the response map. We will discuss them specifically below.

3.1. Discriminative Filter Selection. According to DCF-based trackers, we obtain the correlation output g by

$$g = \sum_{l=1}^{d} h_T^l * s_T^l, \quad l = 1, \cdots, d, \tag{5}$$

where s_T^l and h_T^l are the target sample and target correlation filter corresponding to the l-th channel feature among d channels, respectively. In this paper, g is selected as a 2D Gaussian function where $\sigma = 2.0$ [13].

Before tracking, we need to choose the optimal target correlation filters in the training step via minimizing (6) as

$$H_T^l = \frac{\overline{G} S_T^l}{\sum_{k=1}^{d} \overline{S}_T^k S_T^k + \lambda}, \quad l = 1, \cdots, d, \tag{6}$$

where S_T^l, H_T^l, and G are the DFTs of s_T^l, h_T^l, and g, respectively.

Different from a single training sample of the target appearance, multiple background samples at different locations around target need to be considered to maintain a stable model. However, extracting multichannel features from each background sample increase computational complex significantly. Moreover, in practice, single channel features from multiple background samples are enough to present satisfied performances. Therefore, in this paper, we extract M background samples randomly in the range of an annulus around the target location [11] and obtain the correlation output g as

$$g = \sum_{m=1}^{M} h_B * s_B^m, \quad m = 1, \cdots, M, \tag{7}$$

where s_B^m, $m = 1, \cdots, M$ denotes the m-th background sample.

Similarly, the optimal background correlation filters in the training step are selected via minimizing (8) as

$$H_B^m = \frac{\overline{G} S_B^m}{\sum_{j=1}^{M} \overline{S}_B^j S_B^j + \lambda}, \quad m = 1, \cdots, M, \tag{8}$$

where S_B^m and H_B^m are the DFTs of s_B^m and h_B^m, respectively.

While tracking, DFT $Y_{t,es}$ of the estimated discriminative correlation score $y_{t,es}$ of the test sample z_t is defined as

$$\begin{aligned} Y_{t,es} = \frac{Y_{t,es,T}}{Y_{t,es,B}} &= \frac{\left(\sum_{l=1}^{d} \overline{A}_{t-1,T}^{l} Z_t^l\right) / \left(B_{t-1,T} + \lambda\right)}{\left(\sum_{m=1}^{M} \overline{A}_{t-1,B}^{m} Z_t\right) / \left(B_{t-1,B} + \lambda\right)} \\ &= \frac{\left(\sum_{l=1}^{d} \overline{A}_{t-1,T}^{l} Z_t^l\right) B_{t-1,B}}{\left(Z_t \sum_{m=1}^{M} \overline{A}_{t-1,B}^{m}\right) \left(B_{t-1,T} + \lambda\right)}, \end{aligned} \tag{9}$$

where $Y_{t,es,T}$ and $Y_{t,es,B}$ are the DFTs of the target and background correlation scores $y_{t,es,T}$ and $y_{t,es,B}$, respectively, and Z_t^l is the DFTs of z_t^l. $A_{t-1,T}^l$ and $B_{t-1,T}$ denote the numerator and denominator of the filter H_T^l in (6). $A_{t-1,B}^m$ and $B_{t-1,B}$ denote the numerator and denominator of the filter H_B^m in (8). Then the discriminative correlation scores $y_{t,es}$ are obtained by taking the inverse DFT $y_{t,es} = \mathscr{F}^{-1}\{Y_{t,es}\}$. The estimate of the current target state $z_{t_0,es}$ is obtained by finding the maximum correlation score among the test samples as $z_{t_0,es} \longrightarrow t_0 : t_0 = \arg\max_t y_{t,es}$.

3.2. Fusion Learning Filter. As proved by Wagner et al. [31], late fusion with early estimation provides better performance than early fusion with late estimation. Based on this conclusion, we use the discriminative correlation filters to obtain the estimate of target location in visible and infrared images, respectively, and then do fusion with the fusion correlation filters. Let $z_{t,es,i}, i \in \{vi, ir\}$ denote the estimate of target location of visible vi or infrared ir images. Then we define the region $R_\circ(\cdot, \bullet)$ denoting the minimum bounding rectangle that contains the regions of samples $\cdot$ and $\bullet$ in image $\circ$. Thus, we extract the fusion test samples through $z_{t,i} \in R_i(z_{t,es,vi}, z_{t,es,ir})$, $i \in \{vi, ir\}$ and define the DFT $Y_{t,fu}$ of the fusion correlation score $y_{t,fu}$ of fusion sample $z_{t,i}$ that is defined as

$$\begin{aligned} Y_{t,fu} = \frac{Y_{t,T,fu}}{Y_{t,B,fu}} &= \frac{\sum_{i\in\{vi,ir\}} \alpha_i Y_{t,es,T,i}}{\sum_{i\in\{vi,ir\}} \alpha_i Y_{t,es,B,i}} \\ &= \frac{\sum_{i\in\{vi,ir\}} \alpha_i \left(\left(\sum_{l=1}^{d} \overline{A}_{t-1,T,i}^{l} Z_{t,i}^l\right) / \left(B_{t-1,T,i} + \lambda\right)\right)}{\sum_{i\in\{vi,ir\}} \alpha_i \left(\left(\sum_{m=1}^{M} \overline{A}_{t-1,B,i}^{m} Z_{t,i}\right) / \left(B_{t-1,B,i} + \lambda\right)\right)} \\ &= \sum_{i\in\{vi,ir\}} \frac{\alpha_i \left(\sum_{l=1}^{d} \overline{A}_{t-1,T,i}^{l} Z_{t,i}^l\right) B_{t-1,B,i}}{\left(Z_{t,i} \sum_{m=1}^{M} \overline{A}_{t-1,B,i}^{m}\right) \left(B_{t-1,T,i} + \lambda\right)}, \end{aligned} \tag{10}$$

where $Y_{t,fu,T}$ and $Y_{t,fu,B}$ are the DFTs of the target and background fusion correlation scores $y_{t,fu,T}$ and $y_{t,fu,B}$, respectively. $z_{t,i}^l \in R_{i^l}(z_{t,es,vi}, z_{t,es,ir})$, $i \in \{vi, ir\}$ in which i^l is the l-th channel feature of image i. $Z_{t,i}^l$ and $Z_{t,i}$ are the DFTs of samples $z_{t,i}^l$ and $z_{t,i}$, respectively. α_i denotes the image weights that are denoted as

$$\alpha_{\mathrm{i}} = \frac{\max_i y_{t,es,i}}{\sum_{\bar{i}\in\{vi,ir\}} \max_{\bar{i}} y_{t,es,\bar{i}}}, \tag{11}$$

where $y_{t,es,i}$ is the correlation score of the i-th image computed by (9).

After obtaining $Y_{t,fu}$, the fusion correlation score $y_{t,fu}$ is obtained by taking the inverse DFT $y_{t,fu} = \mathscr{F}^{-1}\{Y_{t,fu}\}$. The fusion location of the current target state is obtained by finding the maximum correlation score among the test samples.

The whole tracking process of DFCL is summarized in Algorithm 1.

4. Experiments

The proposed DFCL algorithm was tested on several challenging real-world sequences, and some qualitative and quantitative analyses were performed on the tracking results in this section.

```
Input: The $t$-th visible and infrared images
For $t = 1$ to number of frames do
    1. Crop the samples and extract the $l$-th ($l = 1, \cdots, d$) channel features $z_t^l$ for visible and
    infrared images, respectively.
    2. Compute the discriminative correlation scores $y_{t,es}$ using Eq. (9).
    3. Compute the fusion correlation scores $y_{t,fu}$ using Eq. (10).
    4. Obtain the tracking result by maximizing $y_{t,fu}$.
    5. Extract the $l$-th ($l = 1, \cdots, d$) channel feature $s_T^l$ of the target samples and the $m$-th
    ($m = 1, \cdots, M$) sample $s_B^l$.
    6. Update the discriminative correlation filters $H_T^l$ and $H_B^m$ using Eq. (6) and Eq. (8),
    respectively.
end for
Output: Target result and the discriminative correlation filters $H_T^l$ and $H_B^m$
```

Algorithm 1: Discriminative fusion correlation learning for visible and infrared tracking.

4.1. Experimental Environment and Evaluation Criteria. DFCL was implemented with C++ programming language and.Net Framework 4.0 in Visual Studio 2010 on an Intel Dual-Core 1.70GHz CPU with 4 GB RAM. Two metrics, i.e., location error (pixel) and overlapping rate, are used to evaluate the tracking results quantitatively. The location error is computed as $error = \sqrt{(x_G - x_T)^2 + (y_G - y_T)^2}$, where (x_G, y_G) and (x_T, y_T) are the ground truth (either downloaded from a standard database or located manually) and tracking bounding box centers, respectively. The tracking overlapping rate is defined as $overlapping = area(ROI_G \cap ROI_T)/area(ROI_G \cup ROI_T)$, where ROI_G and ROI_T denote the ground truth and tracking bounding box, respectively, and $area(\cdot)$ is the rectangular area function. A smaller location error and a larger overlapping rate indicate higher accuracy and robustness.

4.2. Experimental Results. The performance of DFCL was compared with state-of-the-art trackers Struck [32], ODFS [33], STC [34], KCF [35], ROT [36], DCF-based trackers MOSSE [13], DSST [14], fDSST [15], and visible-infrared fusion trackers TSKF [23], MVMKF [27], L1-PF [19], JSR [24], and CSR [25]. Figures 2–6 present the experimental results of the test trackers in challenging visible sequences named *Biker* [37], *Campus* [37], *Car* [38], *Crossroad* [39], *Hotkettle* [39], *Inglassandmobile* [39], *Labman* [40], *Pedestrian* [41], and *Runner* [38], as well as their corresponding infrared sequences *Biker-ir*, *Campus-ir*, *Car-ir*, *Crossroad-ir*, *Labman-ir*, *Hotkettle-ir*, *Inglassandmobile-ir*, *Pedestrian-ir*, and *Runner-ir*. Single-sensor trackers were separately tested on visible and the corresponding infrared sequences, while visible-infrared fusion trackers obtain the results with information from both visible and infrared sequences. For the convenience of presentation, some tracking curves are not shown entirely in the Figures. Next, the performance of the trackers in each sequence is described in detail.

(a) Sequences *Biker* and *Biker-ir*: *Biker* presents the example of complex background clutters. The target human in the visible sequence encounters similar background disturbance (i.e., bikes), which causes the ODFS, MOSSE, fDSST, TSKF, and MVMKF trackers to drift away from the target. The corresponding infrared sequence *Biker-ir* provides temperature information that eliminates the background clutter in *Biker*. But when the target is approaching another person at around Frame #20, Struck, ODFS, STC, MOSSE, TSKF, and MVMKF do not perform well because they are not able to distinguish target from persons with similar temperature in infrared sequences. Only KCF, ROT, DSST, and our DFCL have achieved precise and robust performances in these sequences.

(b) Sequences *Campus* and *Campus-ir*: the target in *Campus* and *Campus-ir* undergoes background clutters, occlusion, and scale variation. At the beginning of *Campus*, ODFS, STC, KCF, and ROT lose the target due to background disturbance. Only TSKF and DFCL perform well, while Struck, fDSST, and MVMKF do not achieve accurate results. Because of the infrared information provided by *Campus-ir*, fewer test trackers lose tracking when background clutters happen as shown in Figure 2. But Struck, KCF, and ROT mistake another person for the target. As shown in Figure 2, most of the trackers result in tracking failures, whereas DFCL outperforms the others in most metrics (location accuracy and success rate).

(c) Sequences *Car* and *Car-ir*: *Car* and *Car-ir* demonstrate the efficiency of DFCL on coping with heavy occlusions. The target driving car is occluded by lampposts and trees many times, which cause tracking failures of most trackers. Only TSKF, MVMKF, and DFCL are able to handle the occlusion throughout the tracking process in this sequence. As shown in Figure 2, most trackers perform better in *Car-ir* than in *Car* because the infrared features can overcome the difficulties of target detection among surrounding similar background. STC, TSKF, MVMKF, and DFCL are able to handle this problem, whereas the result of DFCL is the most accurate, as shown in Figure 2.

(d) Sequences *Crossroad* and *Crossroad-ir*: the target in *Crossroad* and *Crossroad-ir* undergoes heavy background clutters when she crosses the road. While the target is passing by the road lamp, both ODFS and JSR lose the target. Then, when a car passes by the target, Struck, TSKF, and MVMKF drift away from the target. When the target goes toward

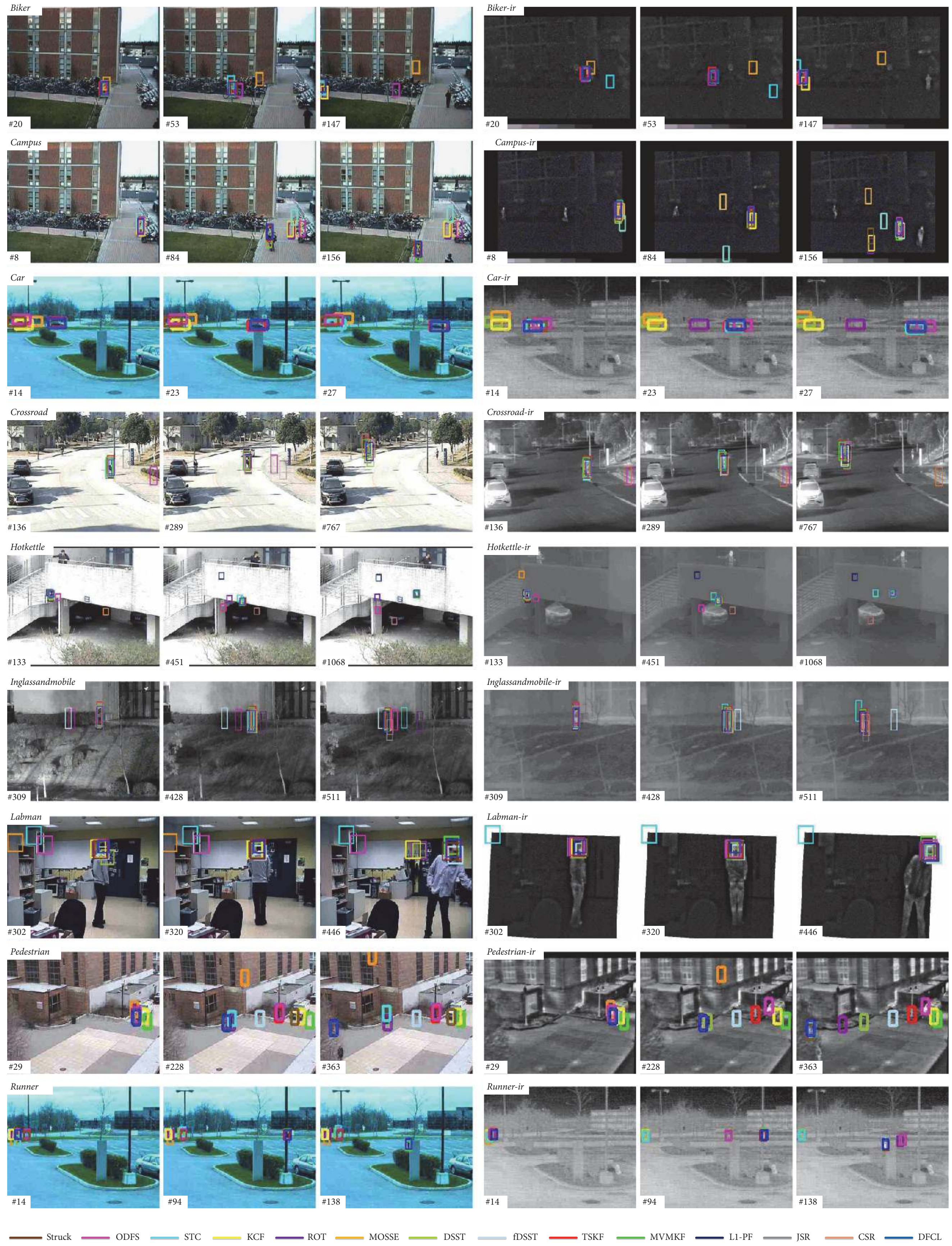


Figure 2: Tracking performances of the test sequences.

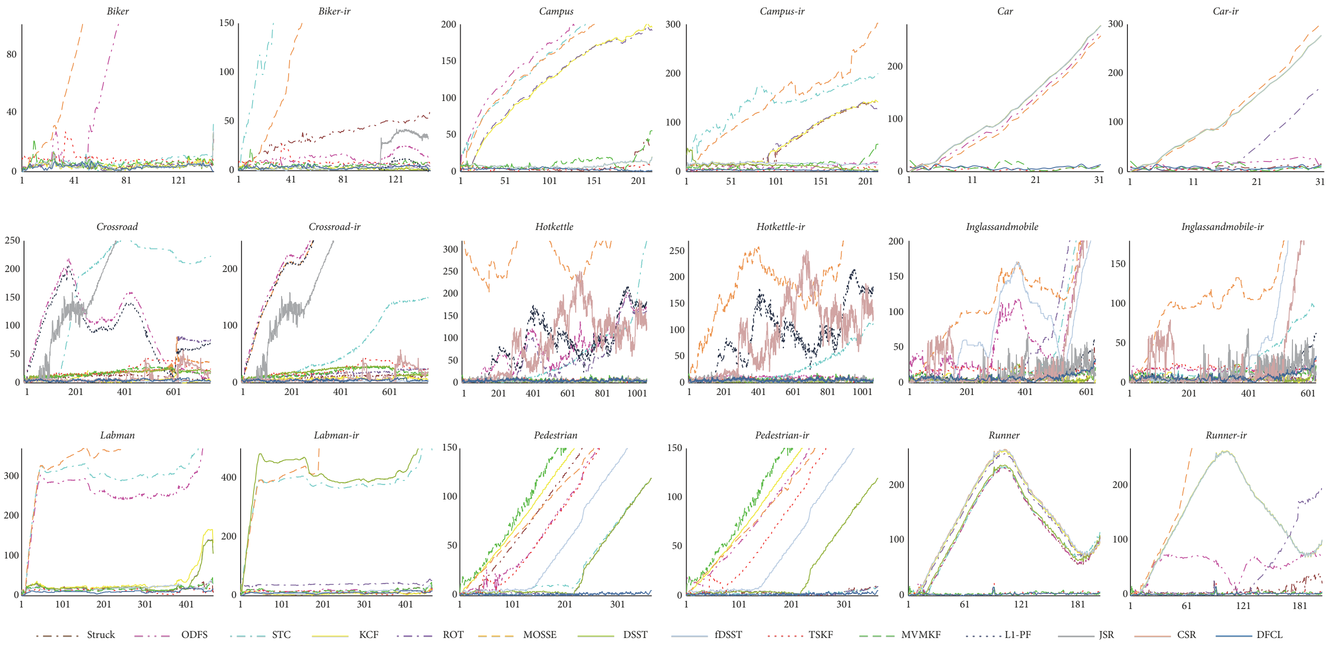


Figure 3: Location error (pixel) of the test sequences.

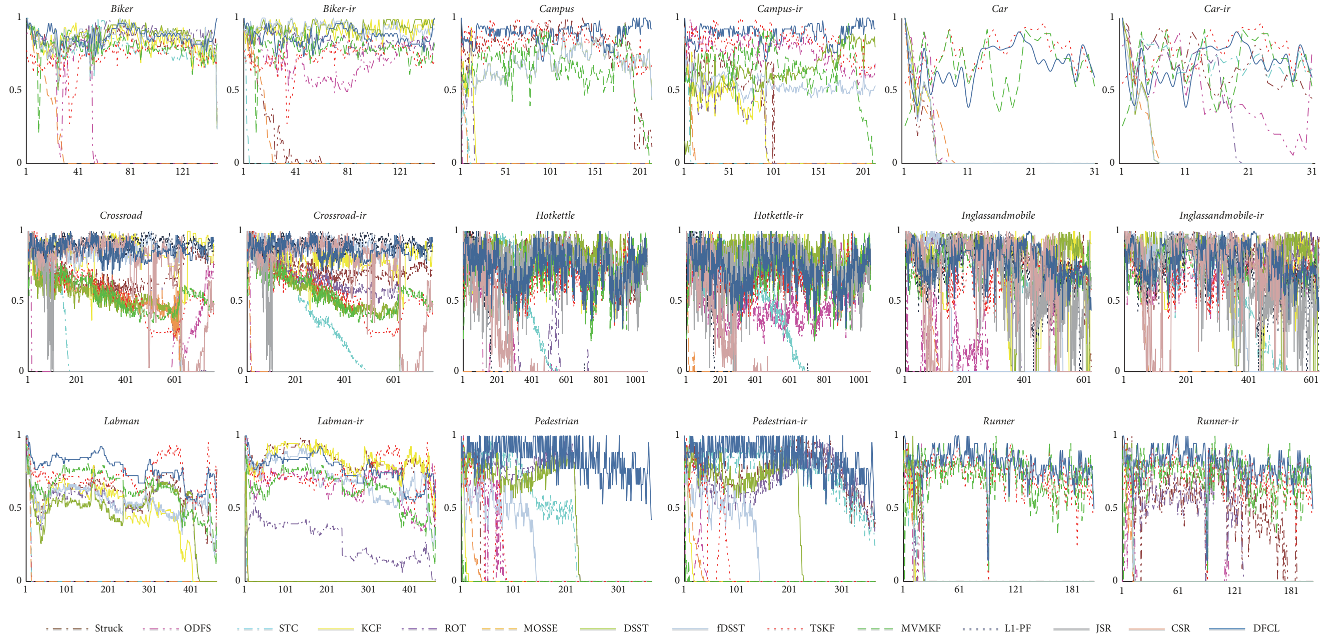


Figure 4: Overlapping rate of the test sequences.

the sidewalk, most of the trackers are not able to handle the problem of heavy background clutters, but our tracker performs satisfying tracking results as shown in Figures 2–4.

(e) Sequences *Hotkettle* and *Hotkettle-ir*: in these sequences, tracking is hard because of the changes of the complex background clutters. Most trackers perform better in *Hotkettle-ir* than in *Hotkettle* for the reason that the temperature diverge makes the hot target more distinct in the cold background. Struck, KCF, DSST, fDSST, and DFCL can achieve robust and accurate tracking performances as shown in Figures 2–4.

(f) Sequences *Inglassandmobile* and *Inglassandmobile-ir*: Sequences *Inglassandmobile* and *Inglassandmobile-ir* demonstrate the performances of the 14 trackers under the circumstances of background clutters, illumination changes, and occlusion. As shown in Figure 2, when the illumination changes at around Frames #300, ODFS and fDSST lose the target, and KCF, TSKF, and L1-PF drift a little away from the

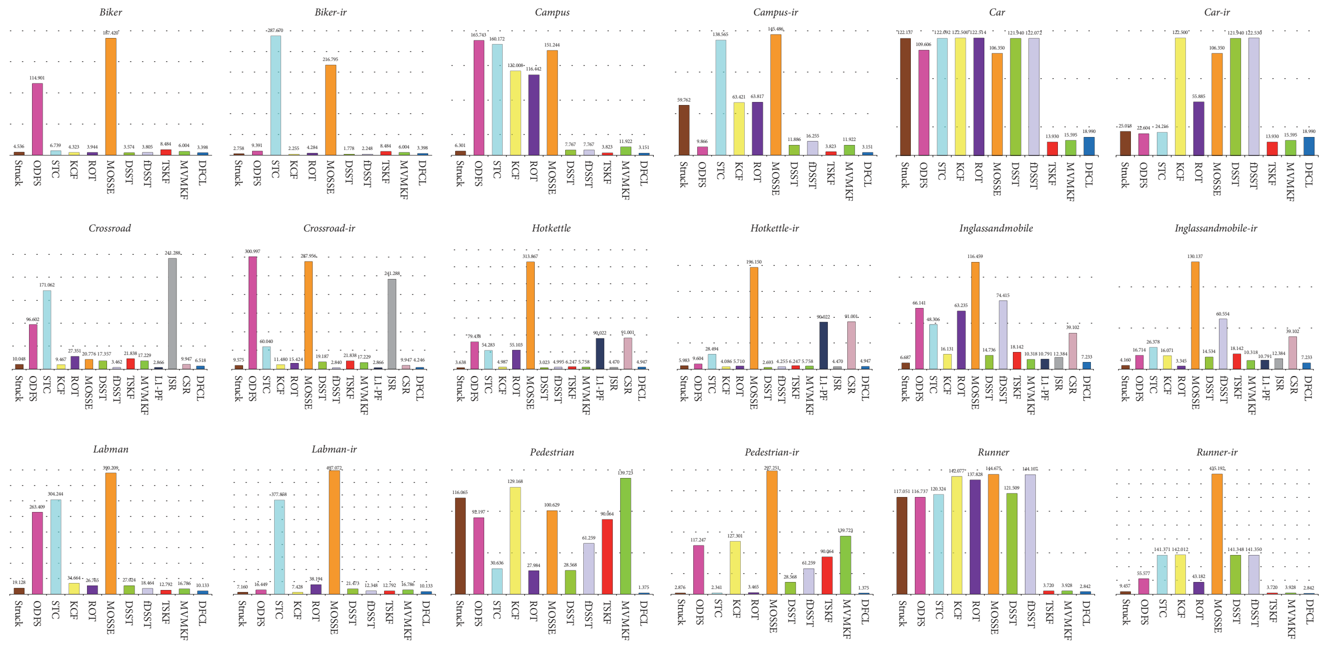


FIGURE 5: Average location error (pixel) of the test sequences.

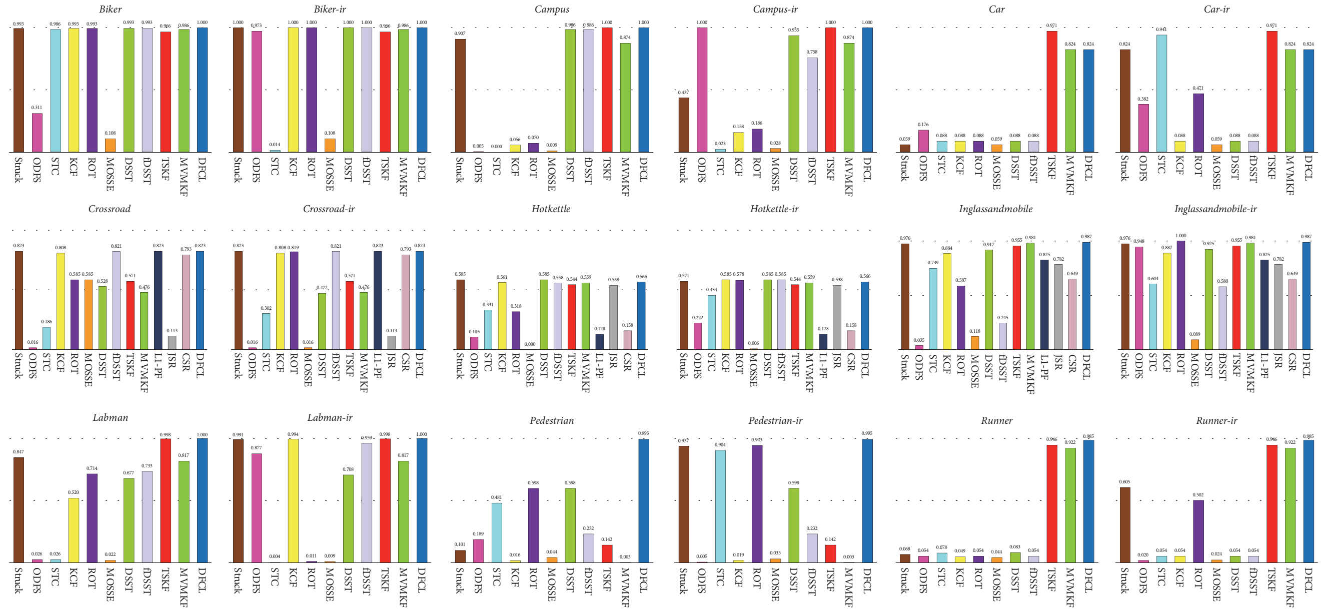


FIGURE 6: Success rate of the test sequences.

target. When the target is approaching a tree, the background clutters makes most of trackers cause tracking failures that can be seen from Figure 2. Our DFCL can overcome these challenges and perform well in these sequences.

(g) Sequences *Labman* and *Labman-ir*: the experiments in Sequences *Labman* and *Labman-ir* aim to evaluate the performances on tracking under appearance variation, rotation, scale variation, and background clutter. In *Labman*, when the target man is walking into the laboratory, ODFS, STC, and MOSSE lose the target. When the man keeps shaking and turning around his head at around Frame #400, KCF, ROT, and DSST cause tracking failures. Also, most trackers achieve better tracking performances in *Labman-ir* as shown in Figure 2.

(h) Sequences *Pedestrian* and *Pedestrian-ir*: the target in *Pedestrian* and *Pedestrian-ir* undergoes heavy background clutters and occlusion. As shown in Figure 2, other trackers result in tracking failures in *Pedestrian*, whereas our tracker shows satisfying performances in terms of both accuracy and robustness. The efficient infrared features extracted from

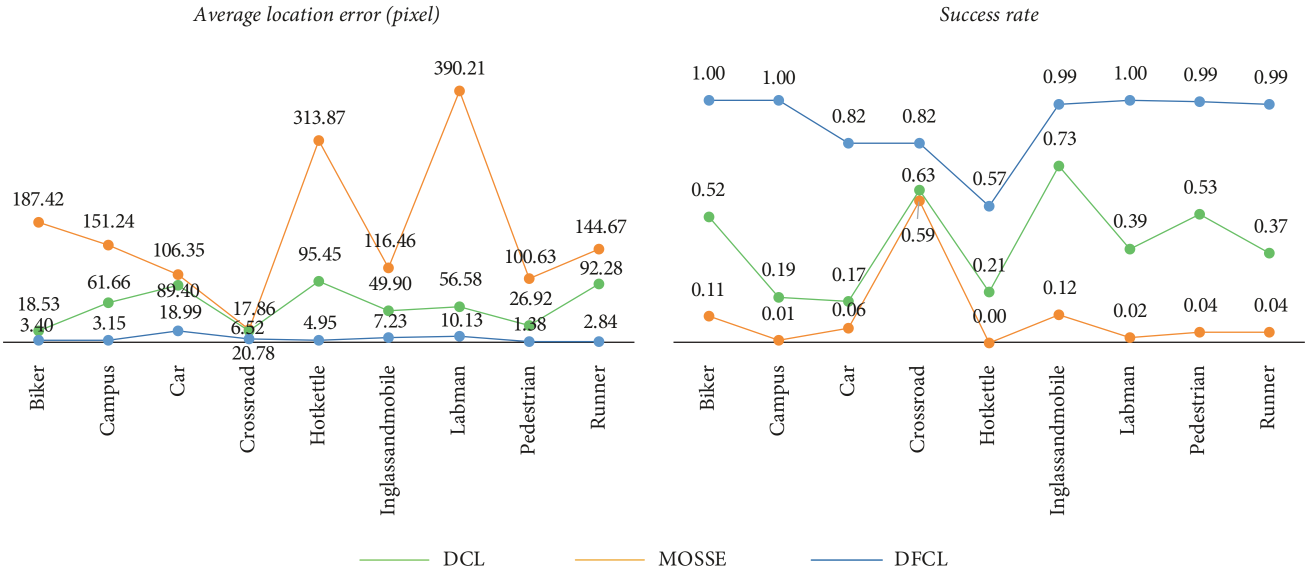


Figure 7: Average location error (pixel) and success rate of DCL, DFCL, and MOSSE on the test sequences.

Pedestrian-ir ensure the tracking successes of Struck, STC, and DFCL, as can be seen from Figures 2–4.

(i) Sequences *Runner* and *Runner-ir*: *Runner* and *Runner-ir* contain examples of heavy occlusion, abrupt movement, and scale variation. The target running man is occluded by lampposts, trees, stone tablet, and bushes many times, resulting in tracking failures of most trackers. Also, the abrupt movement and scale variation cause many trackers to drift away the target in both *Runner* and *Runner-ir* as shown in Figure 2. Once again, our DFCL is able to overcome the above problems and achieve good performances.

Figures 5 and 6 are included here to demonstrate quantitatively the performances on average location error (pixel) and success rate. The success rate is defined as the number of times success is achieved in the whole tracking process by considering one frame as a success if the overlapping rate exceeds 0.5 [33]. A smaller average location error and a larger success rate indicate increased accuracy and robustness. Figures 5 and 6 show that DFCL performs satisfying most of the tracking sequences.

To validate the effectiveness of the discriminative filter selection model of DFCL, we compare the tracker DCL (the proposed DFCL without the fusion learning model) with DFCL and the original DCF tracker MOSSE on visible sequences. The performances shown in Figure 7 demonstrate the efficiency of the discriminative filter selection model especially in the sequences with background clutters, i.e., Sequences *Biker*, *Hotkettle*, *Inglassandmobile*, and *Pedestrian*.

5. Conclusion

Discriminative correlation filter- (DCF-) based trackers have the advantage of being computationally efficient and more robust than most of the other state-of-the-art trackers in challenging tracking tasks, thereby making them especially suitable for a variety of real-time challenging applications. However, most of the DCF-based trackers suffer low accuracy due to the lack of diversity information extracted from a single type of spectral image (visible spectrum). Fusion of visible and infrared sensors, one of the typical multisensor cooperation, provides complementarily useful features and consistently helps recognize the target from the background efficiently in visual tracking. For the above reasons, this paper proposes a discriminative fusion correlation learning model to improve DCF-based tracking performance by combining multiple features from visible and infrared imaging sensors. The proposed fusion learning filters are obtained via late fusion with early estimation, in which the performances of the filters are weighted to improve the flexibility of fusion. Moreover, the proposed discriminative filter selection model considers the surrounding background information in order to increase the discriminability of the template filters so as to improve model learning. Numerous real-world video sequences were used to test DFCL and other state-of-the-art algorithms, and here we only selected representative videos for presentation. Experimental results demonstrated that DFCL is highly accurate and robust.

Data Availability

The data used to support the findings of this study were supplied by China University of Mining and Technology under license and so cannot be made freely available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the Natural Science Foundation of Jiangsu Province (BK20180640, BK20150204), Research Development Programme of Jiangsu Province (BE2015040), the State Key Research Development Program (2016YFC0801403), and the National Natural Science Foundation of China (51504214, 51504255, 51734009, 61771417, and 61873246).

References

[1] S. A. Wibowo, H. Lee, E. K. Kim, and S. Kim, "Visual tracking based on complementary learners with distractor handling," *Mathematical Problems in Engineering*, vol. 2017, Article ID 5295601, 13 pages, 2017.

[2] T. Zhou, M. Zhu, D. Zeng, and H. Yang, "Scale adaptive kernelized correlation filter tracker with feature fusion," *Mathematical Problems in Engineering*, vol. 2017, Article ID 1605959, 8 pages, 2017.

[3] C. Wang, H. K. Galoogahi, C. Lin, and S. Lucey, "Deep-LK for efficient adaptive object tracking," *Computer Vision and Pattern Recognition*, 2017.

[4] Z. Chi, H. Li, H. Lu, and M.-H. Yang, "Dual deep network for visual tracking," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 2005–2015, 2017.

[5] Y. Yang, W. Hu, Y. Xie, W. Zhang, and T. Zhang, "Temporal restricted visual tracking via reverse-low-rank sparse learning," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 485–498, 2017.

[6] Y. Sui, G. Wang, and L. Zhang, "Correlation filter learning toward peak strength for visual tracking," *IEEE Transactions on Cybernetics*, vol. 48, no. 4, pp. 1290–1303, 2018.

[7] K. Zhang, X. Li, H. Song, Q. Liu, and W. Lian, "Visual tracking using spatio-temporally nonlocally regularized correlation filter," *Pattern Recognition*, vol. 83, pp. 185–195, 2018.

[8] Y. Qi, L. Qin, J. Zhang, S. Zhang, Q. Huang, and M.-H. Yang, "Structure-aware local sparse coding for visual tracking," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3857–3869, 2018.

[9] B. Bai, B. Zhong, G. Ouyang et al., "Kernel correlation filters for visual tracking with adaptive fusion of heterogeneous cues," *Neurocomputing*, vol. 286, pp. 109–120, 2018.

[10] Y. Liu, F. Yang, C. Zhong, Y. Tao, B. Dai, and M. Yin, "Visual tracking via salient feature extraction and sparse collaborative model," *AEÜ - International Journal of Electronics and Communications*, vol. 87, pp. 134–143, 2018.

[11] X. Yun, Y. Sun, S. Wang, Y. Shi, and N. Lu, "Multi-layer convolutional network-based visual tracking via important region selection," *Neurocomputing*, vol. 315, pp. 145–156, 2018.

[12] X. Lan, S. Zhang, and P. C. Yuen, "Robust joint discriminative feature learning for visual tracking," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016*, pp. 3403–3410, AAAI Press, New York, NY, USA, July 2016.

[13] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2544–2550, IEEE, San Francisco, Calif, USA, June 2010.

[14] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proceedings of the British Machine Vision Conference*, pp. 65.1-65.11, BMVA Press, Nottingham, UK, 2014.

[15] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561–1575, 2017.

[16] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proceedings of the 16th IEEE International Conference on Computer Vision, ICCV 2017*, pp. 1144–1152, IEEE, Istanbul, Turkey, 2018.

[17] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M.-H. Yang, "Crest: convolutional residual learning for visual tracking," in *Proceedings of the 16th IEEE International Conference on Computer Vision, ICCV 2017*, pp. 2574–2583, IEEE, Venice, Italy, October 2017.

[18] J. Johnander, M. Danelljan, F. . Khan, and M. Felsberg, "DCCO: towards deformable continuous convolution operators for visual tracking," in *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, pp. 55–67, Springer, Ystad, Sweden, 2017.

[19] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple source data fusion via sparse representation for robust visual tracking," in *Proceedings of the 14th International Conference on Information Fusion, Fusion 2011*, IEEE, Chicago, IL, USA, July 2011.

[20] B. Uzkent, A. Rangnekar, and M. J. Hoffman, "Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps," in *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2017*, pp. 233–242, IEEE, Honolulu, HI, USA, July 2017.

[21] S. Chan, X. Zhou, and S. Chen, "Robust adaptive fusion tracking based on complex cells and keypoints," *IEEE Access*, vol. 5, pp. 20985–21001, 2017.

[22] M. K. Rapuru, S. Kakanuru, P. M. Venugopal, D. Mishra, and G. R. Subrahmanyam, "Correlation-based tracker-level fusion for robust visual tracking," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4832–4842, 2017.

[23] X. Yun, Z. Jing, G. Xiao, B. Jin, and C. Zhang, "A compressive tracking based on time-space Kalman fusion model," *Science China Information Sciences*, vol. 59, no. 1, pp. 1–15, 2016.

[24] H. P. Liu and F. C. Sun, "Fusion tracking in color and infrared images using joint sparse representation," *Science China Information Sciences*, vol. 55, no. 3, pp. 590–599, 2012.

[25] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5743–5756, 2016.

[26] S. Mangale and M. Khambete, "Camouflaged target detection and tracking using thermal infrared and visible spectrum imaging," in *Automatic Diagnosis of Breast Cancer using Thermographic Color Analysis and SVM Classifier*, vol. 530 of *Advances in Intelligent Systems and Computing*, pp. 193–207, Springer International Publishing, 2016.

[27] X. Yun, Z. Jing, and B. Jin, "Visible and infrared tracking based on multi-view multi-kernel fusion model," *Optical Review*, vol. 23, no. 2, pp. 244–253, 2016.

[28] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, and F. S. Khan, "Synthetic data generation for end-to-end thermal infrared tracking," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1837–1850, 2019.

[29] X. Lan, M. Ye, S. Zhang, and P. C. Yuen, "Robust collaborative discriminative learning for rgb-infrared tracking," in *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, pp. 7008–7015, AAAI, New Orleans, LA, USA, 2018.

[30] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang, “Fusing two-stream convolutional neural networks for RGB-T object tracking,” *Neurocomputing*, vol. 281, pp. 78–85, 2018.

[31] J. Wagner, V. Fischer, M. Herman, and S. Behnke, “Multispectral pedestrian detection using deep fusion convolutional neural networks,” in *Proceedings of the 24th European Symposium on Artificial Neural Networks*, pp. 1–6, Bruges, Belgium, 2016.

[32] S. Hare, A. Saffari, and P. H. S. Torr, “Struck: structured output tracking with kernels,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 263–270, IEEE, Providence, RI, USA, 2012.

[33] K. Zhang, L. Zhang, and M.-H. Yang, “Real-time object tracking via online discriminative feature selection,” *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4664–4677, 2013.

[34] K. Zhang, L. Zhang, M. Yang, and D. Zhang, “Fast tracking via spatio- temporal context learning,” *Computer Vision and Pattern Recognition*, 2013.

[35] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[36] X. Dong, J. Shen, D. Yu, W. Wang, J. Liu, and H. Huang, “Occlusion-aware real-time object tracking,” *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 763–771, 2017.

[37] C. Ó. Conaire, N. E. O’Connor, and A. Smeaton, “Thermo-visual feature fusion for object tracking using multiple spatiogram trackers,” *Machine Vision & Applications*, vol. 19, no. 5-6, pp. 483–494, 2008.

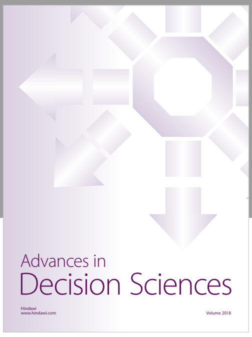
[38] INO, “Ino’s video analytics dataset,” https://www.ino.ca/en/video-analytics-dataset/.

[39] C. Li, X. Liang, Y. Lu, Z. Nan, and T. Jin, “RGB-T object tracking: benchmark and baseline,” *IEEE Transactions on Image Processing*, 2018.

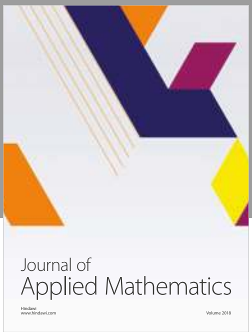
[40] C. Ó. Conaire, N. E. O’Connor, and A. F. Smeaton, “Detector adaptation by maximising agreement between independent data sources,” in *Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR’07*, pp. 1–6, IEEE, Minneapolis, MN, USA, June 2007.

[41] J. W. Davis and V. Sharma, “Background-subtraction using contour-based fusion of thermal and visible imagery,” *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 162–182, 2007.

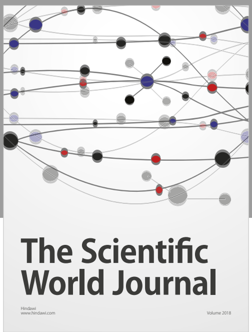
Advances in Operations Research



Advances in Decision Sciences

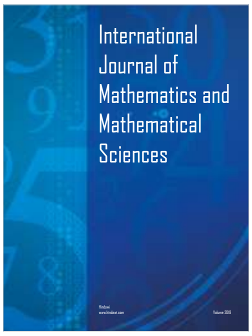


Journal of Applied Mathematics



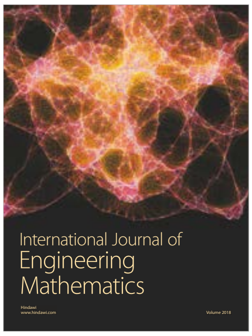
The Scientific World Journal

Journal of Probability and Statistics



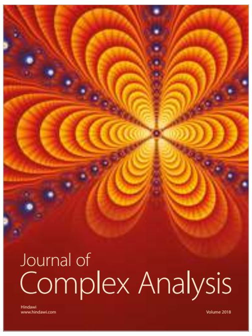
International Journal of Mathematics and Mathematical Sciences

Journal of Optimization

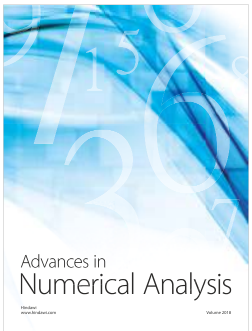


International Journal of Engineering Mathematics

International Journal of Analysis

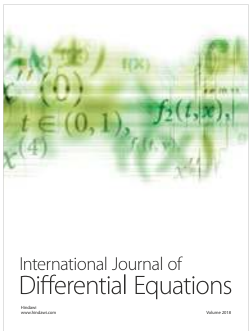


Journal of Complex Analysis

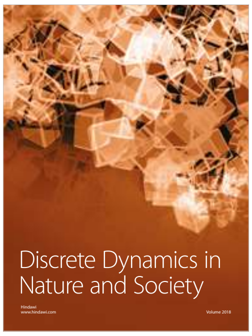


Advances in Numerical Analysis

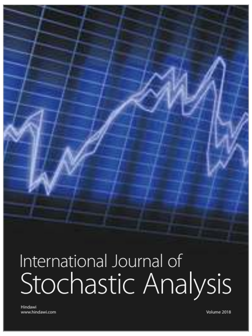
Mathematical Problems in Engineering



International Journal of Differential Equations



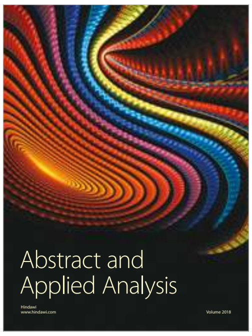
Discrete Dynamics in Nature and Society



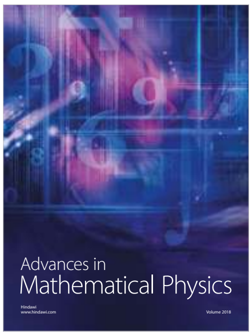
International Journal of Stochastic Analysis

Journal of Mathematics

Journal of Function Spaces



Abstract and Applied Analysis



Advances in Mathematical Physics