Deep Convolutional Neural Networks for Thermal Infrared Object Tracking

Qiao Liu\textsuperscript{a,\ast}, Xiaohuan Lu\textsuperscript{a,\ast}, Zhenyu He\textsuperscript{a,\ast\ast}, Chunkai Zhang\textsuperscript{a,\ast\ast}, Wen-Sheng Chen\textsuperscript{b}

\textsuperscript{a}School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China
\textsuperscript{b}College of Mathematics and Statistics, Shenzhen Key Laboratory of Media Security, Shenzhen University, China

Abstract

Unlike the visual object tracking, thermal infrared object tracking can track a target object in total darkness. Therefore, it has broad applications, such as in rescue and video surveillance at night. However, there are few studies in this field mainly because thermal infrared images have several unwanted attributes, which make it difficult to obtain the discriminative features of the target. Considering the powerful representational ability of convolutional neural networks and their successful application in visual tracking, we transfer the pre-trained convolutional neural networks based on visible images to thermal infrared tracking. We observe that the features from the fully-connected layer are not suitable for thermal infrared tracking due to the lack of spatial information of the target, while the features from the convolution layers are. Besides, the features from a single convolution layer are not robust to various challenges. Based on this observation, we propose a correlation filter based ensemble tracker with multi-layer convolutional features for thermal infrared tracking (MCFTS). Firstly, we use pre-trained convolutional neural networks to extract the features of the multiple convolution layers of the thermal infrared target. Then, a correlation filter is used to construct multiple weak trackers with the corresponding convolution layer features. These weak trackers give the response maps of the target’s location. Finally, we propose an ensemble method that coalesces these response maps to get a stronger one. Furthermore, a simple but effective scale estimation strategy is exploited to boost the tracking accuracy. To evaluate the performance of the proposed tracker, we carry out experiments on two thermal infrared tracking benchmarks: VOT-TIR 2015 and VOT-TIR 2016. The experimental results demonstrate that our tracker is effective and achieves promising performance.

Keywords: Thermal infrared tracking, Convolutional features, Correlation filter, Ensemble method

\textsuperscript{\ast}These authors contributed equally to this work and should be considered co-first authors
\textsuperscript{\ast\ast}I am corresponding author

Email addresses: zhenyuhe@hit.edu.cn (Zhenyu He), ckzhang@hitsz.edu.cn (Chunkai Zhang), chenws@szu.edu.cn (Wen-Sheng Chen)
1. Introduction

Visual object tracking is an important computer vision problem with various applications. In the past decades, a large number of discriminative tracking methods [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] based on visible image sequences were proposed to solve various challenges, while the discriminative method has been used in classification [11, 12], face recognition [13, 14, 15, 16, 17, 18], action recognition [19, 20], handwriting identification [21, 22, 23], image segmentation [24, 25, 26, 27], image denoising [28, 29] and so on. Especially, in the recent two years, visual tracking has achieved a significant breakthrough by using deep learning. Despite much progress, visual tracking remains a largely unsolved problem due to negative factors such as changing appearance, occlusions, illumination variation, and background clutter. Compared with visual object tracking, thermal infrared tracking has several advantages. For instance, it is not influenced by illumination variation because the thermal infrared image does not rely on reflected light but mainly depends on the radiant temperature of the object. Thermal infrared tracking also can track the target in total darkness while visual tracking always fails in a bad visibility condition. Furthermore, in some real scenarios that are sensitive to personal information, thermal infrared tracking can protect privacy. Therefore, thermal infrared tracking is suitable for a variety of applications, such as rescue, video surveillance, and patrols at night.

Despite many superiorities, thermal infrared tracking is faced with many challenges in the meantime. First, thermal infrared images have some unwanted attributes, such as low resolution, a larger percentage of dead pixels, and no visual color patterns [30]. These unwanted attributes make it difficult to obtain the discriminative features of the target object and hence degrade the tracking performance. Additionally, without color patterns, there are many similar targets in the background to disturb the tracker. For example, two people, one dressed in green clothes, and the other in red clothes are almost the same in a thermal infrared image. There are also some other challenges faced by thermal infrared tracking, such as occlusion, changing appearance, and motion blur. Among these challenges, obtaining discriminative features is the basic and the most important one.

In the past several years, deep learning [31] has drawn a lot of attention from the computer vision community. It has been successfully applied in object recognition [32, 33, 34] and object detection [35, 36, 37], even electrocardiograph segmentation [38]. Most recently, deep learning has also been introduced into visual tracking and achieved promising results. Several convolutional neural networks (CNN) based trackers [39, 40, 41, 42, 43, 44] have been developed. Compared with the hand-crafted features based trackers, these trackers can easily obtain more superior tracking performance.

Inspired by these CNN-based trackers, we find that the powerful representational ability of a pre-trained CNN is the key to these methods and this CNN is also appropriate to thermal infrared tracking. Because CNN learns general features of the target object, it also can well represent the thermal infrared target. In this paper, we transfer VGG-Net [45] trained on ImageNet [46] to thermal infrared tracking. VGG-Net is used to extract the features of the thermal infrared target. We observe that the features from the fully-connected layer are not suitable for tracking on the thermal infrared video because they lack the spatial information of the target. Consequently, it is hard to determine the target’s locations in the tracking process. On the contrary, the convolutional features
from the convolution layers perform well in thermal infrared tracking mainly because they have rich spatial information of the target. Nevertheless, the features from a single convolution layer are not robust to the challenges, e.g., appearance variation and occlusion, in the tracking, because they have limited semantic information. Therefore, we propose a correlation filter based ensemble tracker (MCFTS) that leverages multi-layer convolutional features for the thermal infrared tracking. Specifically, we construct several weak trackers on multiple convolution layers by using the kernelized correlation filter (KCF) [1]. Each weak tracker gives a response map of the target’s location. Then, we coalesce these response maps using the Kullback-Leibler divergence to obtain a stronger response map of the target’s location. To improve the accuracy of the tracking, we also exploit a simple scale estimation strategy for our tracker. The experimental results demonstrate that the proposed method achieves promising performance.

The contributions of this paper are threefold:

• We transfer the pre-trained VGG-Net to thermal infrared tracking to extract the features of the thermal infrared target. Moreover, we find that the features from the convolution layer are more effective than a fully-connected layer for thermal infrared tracking.

• A correlation filter based ensemble tracker with multi-layer convolutional features is proposed. Meanwhile, a fusion method based Kullback-Leibler divergence and a simple scale estimation strategy are provided to improve the tracking performance.

• Extensive experiments are carried out on two thermal infrared tracking benchmarks: VOT-TIR 2015 [47] and VOT-TIR 2016 [48]. The experimental results demonstrate that the proposed method is effective and achieves promising results.

The rest of this paper is organized as follows. We briefly introduce some related works in Section 2. Then, we give a short introduction to the framework of the proposed algorithm in Section 3. Subsequently, we describe each component of the proposed method in detail in Section 4. Section 5 presents the experimental details and results. Finally, we draw a short conclusion in Section 6.

2. Related Work

Correlation filter based trackers. Recently, correlation filter tracking methods have become more popular mainly because of their desirable property of high computational efficiency. Since the first correlation filter, referred to as Minimum Output Sum of Squared Error (MOSSE) [7], was proposed for visual tracking, there has been a booming development of the correlation filter based trackers. Henriques et al. [49] improve the MOSSE filter by introducing kernel methods due to the limited ability of the linear classification in MOSSE. Subsequently, several multi-channel extension kernel methods [1, 50] have been proposed to adapt multi-dimension features. Additionally, a series of correlation filter based trackers have been proposed to handle a variety of challenges. For example, to adjust the scale variation of the target, Danelljan et al. [8] present a scale pyramid representation method for robust tracking. Several part-based tracking methods based on a correlation filter [51, 52] have been proposed to tackle occlusion. To handle long-term
tracking and re-detection, Ma et al. [53] present a long-term correlation tracking framework which decomposes the task of tracking into the translation and scale estimation of the objects. However, all of the above-mentioned correlation filter based trackers suffer an inherent unwanted boundary effect [54]. To overcome this problem, Danelljan et al. [55] exploit a spatial regularization to penalize the correlation filter coefficients (SRDCF). That tracker achieves promising results in all tracking benchmarks.

CNN based correlator trackers. Along with introducing the deep learning into the field of visual tracking, some researchers have combined CNN with a correlation filter for the tracking. Several CNN based correlator trackers have been presented. For instance, in [41], the authors combine the hierarchical convolutional features that are extracted from the multiple convolution layers of the CNN with kernelized correlation filters to construct a coarse-to-fine tracking model (HCF). This model can locate the target more precisely than the other models, which only use the features from the fully connected layer. Qi et al. [42] propose a Hedge tracker (HDT) that also uses multi-layer convolutional features with KCF and an adaptive Hedge method. Due to its exploitation of the multi-layer convolutional features that are robust to various challenges, this tracker achieves promising tracking performance. Danelljan et al. [56] investigate the impact of the convolutional features in a discriminative correlation filter (DCF) and the SRDCF framework. Subsequently, they extend the SRDCF using the convolutional features to obtain a superior deep tracker (Deep-SRDCF). All of these CNN based correlator trackers achieve state-of-the-art results.

Correlation filters based thermal infrared trackers. Influenced by the correlation filter successfully applied to visual tracking, there have been developed several thermal infrared trackers based on correlation filters. For example, Gundogdu et al. [57] employ multiple correlation filters as base trackers with different learning rates and an ensemble method to choose the optimal base tracker as the final decision result (TBOOST). This approach adapts to the changing appearance of the target with the aid of a switching mechanism. In [58], the authors leverage weighted multi-features (Gray, Spatial, Motion) to replace the single feature in the correlation filter framework. This algorithm has a performance superior to that of a correlator tracker that only uses a single feature. In [59], the authors investigate the impact of the convolutional features in the SRDCF [55] and DSST [8] frameworks. The convolutional features are extracted from the first layer of a pre-trained CNN based on a 16K infrared image dataset. The conclusion of this work finds that the convolutional features have richer information than hand-crafted features. Hence, it is more suitable for thermal infrared tracking within the correlation filter framework.

The most closely related studies to this paper are HDT [42] and TBOOST [57]. However, unlike HDT, our method mainly aims at thermal infrared tracking while HDT focuses on visual tracking. Additionally, in this paper, we propose a parameter-free ensemble method to replace the Hedge method in HDT. Unlike the TBOOST method, our tracker uses convolutional features while TBOOST uses hand-crafted features, e.g., HOG [60]. Furthermore, there is difference in the ensemble methods of the two papers. Our ensemble method exploits the responses of all weak trackers while the ensemble method of TBOOST only cares about the decision of the strongest base tracker.
Figure 1: The framework of the proposed method. The overall algorithm can be divided into two parts: training (top-left) and searching (bottom). **Training:** we use the pre-trained VGG-Net to extract the convolutional features of the thermal infrared target on the predicted frame (the $t$-th frame). Then, these convolutional features and a Gaussian shape label are used to get the filters on each convolution layer. **Searching:** The pre-trained VGG-Net is also used to extract the convolutional features on the search frame (the $(t+1)$-th frame). Then, the filters from the training and the feature maps from the search frame are gathered to calculate their cross-correlation. Each filter and feature map generate a response map. This process can be seen as a weak tracker. Finally, we fuse these response maps to get a stronger one by an ensemble method. And this process can be seen as an ensemble tracker.

3. Overview of Algorithm

We give a brief introduction to the overall framework of the algorithm, as shown in Fig. 1. We divide the algorithm into two stages: training and searching. In the training stage, our purpose is to get the filter templates on the predicted frame (the $t$-th frame). First, we utilize the pre-trained VGG-Net to extract the convolutional features of the thermal infrared target on multiple convolution layers. Then, we get the corresponding filter using these convolutional features and a Gaussian shape label. In the searching stage, our goal is to obtain the optimal tracking result on the search frame (the $(t+1)$-th frame). As in the training stage, we extract the feature map of the target on each convolution layer. Subsequently, these feature maps and the filter templates are gathered to calculate their cross-correlation. This operator can be seen as a weak correlation filter based tracker (see Section 4.1 for more details), and each weak tracker gives a response map. Finally, we propose an ensemble method (see Section 4.2) based on Kullback-Leibler divergence to fuse these response maps and return a stronger one.

4. The Proposed Approach

In this section, we give the main components of the proposed method. We first present the weak tracker using the correlation filter and convolutional features in Section 4.1, then give an
ensemble method to fuse the several response maps, and a scale estimation strategy in Section 4.2, and finally show the online model update in Section 4.3.

4.1. The Weak Trackers

Correlation filter based trackers [1, 8, 55, 61] have received considerable attention in recent years due to their lower computational expense. In this work, we also use a KCF [1] with convolutional features to construct a weak tracker. The pre-trained VGG-Net that is trained on the ImageNet dataset [46] is used to extract the convolutional features of the thermal infrared target. Given a feature map of a thermal infrared target region $X^k \in \mathbb{R}^{M \times N \times D}$ ($M$, $N$, $D$ denote the width, height, and the number of the channels respectively), which is extracted from the $k$-th convolution layer, and a corresponding Gaussian shape label matrix $Y \in \mathbb{R}^{M \times N}$. We set $X^k = F(X^k)$ and $Y = F(Y)$, where $F(\cdot)$ denotes the discrete Fourier transformation (DFT). Then, the corresponding filter of the $k$-th convolution layer can be formulated in the Fourier domain:

$$W^k = \arg\min_W \| Y - X^k \cdot W \|^2_F + \lambda \| W \|^2_F,$$  \hspace{1cm} (1)

where $\lambda$ is a regularization parameter, and

$$X^k \cdot W = \sum_{d=1}^{D} X^k \ast_{d} \odot W_{*,*,d},$$  \hspace{1cm} (2)

where $d = \{1, 2, \cdots, D\}$, and $\odot$ is the element-wise product.

Solving Eq. 1 corresponds to the training step in the framework of our method, as shown in Fig. 1. Fortunately, we can fast optimize Problem 1 in the Fourier domain since it has a simple closed form solution:

$$W^k_{*,*,d} = \frac{Y}{X^k \cdot X^k + \lambda} \odot X^k_{*,*,d}.$$  \hspace{1cm} (3)

In the searching stage, our purpose is to get the response map of the target’s location at search frame. Given a thermal infrared search region at search frame, we also use the pre-trained VGG-Net to extract the features of this search region. Let $V^k \in \mathbb{R}^{M \times N \times D}$ denote the feature map of the $k$-th convolution layer of this search region. We first transform it to the Fourier domain: $V = F(V)$. Then, the response map of the target’s location at search frame can be obtained by the following cross-correlation formulation:

$$P^k = F^{-1}(V \cdot W^k),$$  \hspace{1cm} (4)

where $P^k \in \mathbb{R}^{M \times N}$ is the response map of the target’s location at search frame, as shown in Fig. 1, and $F^{-1}$ denotes the inverse of the discrete Fourier transformation.

4.2. The Ensemble Tracker

Given $n$ response maps $P = \{P^1, P^2, \cdots, P^n\}$, where each response map $P^k \in \mathbb{R}^{M \times N}$ ($k = 1, 2, \cdots, n$) is generated from a weak tracker, our goal is to fuse these response maps to get a stronger response map $Q \in \mathbb{R}^{M \times N}$. In fact, each response map $P^k$ can be seen as a probability map, which consists of a probability distribution $p^k_{ij}$, $(i, j) \in \{1, 2, \cdots, M\} \times \{1, 2, \cdots, N\}$. This probability distribution
denotes the probability of position \((i, j)\) to be the center of the target and is subjected to \(\sum p_{ij}^k = 1\). Therefore, we can use the Kullback-Leibler divergence to measure the distance between the probability map \(P^k\) and the fused probability map \(Q\). Then, we minimize this distance to optimize the fused probability map \(Q\):

\[
\arg\min_Q \sum_{k=1}^{n} KL(P^k || Q)
\]

(5)

\[s.t. \sum q_{ij} = 1,
\]

where

\[
KL(P^k || Q) = \sum_{ij} p_{ij}^k \log \frac{p_{ij}^k}{q_{ij}}
\]

(6)

and \(p_{ij}, q_{ij}\) denote the \((i, j)\)-th elements of the probability maps \(P\) and \(Q\), respectively.

Usually, the feature map from the VGG-Net contains some noise. The response map obtained from Eq. 4 also has much noise. Therefore, we first filter the noise of the response map to get one with more confidence before we solve Problem 5. To achieve this, we just use a simple strategy that exploits another probability map to filter the current probability map. We formulate this by the following equation:

\[
P^{k,z} = p^k \odot p^z,
\]

(7)

where \(k = \{1, 2, \cdots, n - 1\}\), \(z = \{k + 1, k + 2, \cdots, n\}\). Eq. 7 means that if two probability maps have similar probability distribution at the same region, the filtered probability map has a higher output in this region and the other regions return lower values. After calculating Eq. 7 for \(n\) response maps, we get a set of filtered probability maps \(A = \{P^{1.2}, P^{1.3}, \cdots, P^{2.3}, \cdots, P^{n-1,n}\}\), which has \(\frac{n(n-1)}{2}\) response maps with less noise. Now, our purpose is to fuse these filtered response maps. Therefore, the objective function Eq. 5 can be rewritten as the following formulation:

\[
\arg\min_Q \sum_{p \in A} \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}}
\]

(8)

\[s.t. \sum q_{ij} = 1.
\]

The solution of Eq. 8 can be obtained by the Lagrange multiplier method, and the final result is as follows:

\[
Q = \frac{2}{n(n-1)} \sum_{p \in A} p.
\]

(9)

We can see that this solution 9 has a simple form: it is an average value of all the filtered response maps. This has a more intuitive explanation: the final result is strengthened by all the filtered response maps using a weighted sum.

Finally, we obtain the target’s location \((x, y)\) at the search frame by finding the maximum response of the fused response map \(Q\):

\[
(x, y) = \arg\max_{i, j} Q(i, j).
\]

(10)
However, the ensemble tracker can not adapt to the changing appearance of the target and hence it has limited tracking performance. In order to improve the accuracy of the ensemble tracker, an effective scale estimation method is necessary. In this paper, we adopt a simple but effective scale estimation strategy, which is presented in [61] by our group. For the given three different scale’s targets $I_{d_{t+1}} \in \{I_{t-1}, I_0, I_1\}$, the main purpose of this scale strategy is to find a corresponding scale changed direction of the maximal response map.

$$
\hat{d}_{t+1} = \arg\max_{d_{t+1} \in D} \beta_d \alpha(I_{d_{t+1}}),
$$

(11)

where $d_{t+1} \in D = \{-1, 0, 1\}$ denotes three scale change directions of the target at the frame $t + 1$, namely zooms out, no change, and zoom in, respectively. $\alpha(\cdot)$ denotes the maximal value of a response map which can be obtained by our tracker mentioned above, and $\beta_d \in \{\beta_{-1}, \beta_0, \beta_1\}$ represents the fixed weights of these scale change directions.

For each scale change direction, we give a fixed change step size $R_{d_{t+1}} \in \{R_{-1}, R_0, R_1\}$ at each frame to update the scale factor:

$$
f_{t+1} = f_t \times R_{d_{t+1}},
$$

(12)

where $f_t$ is the scale factor of the predicted frame $t$. Then we also can obtain the target size of the search frame $t + 1$:

$$
s_{t+1} = s_1 \times f_{t+1},
$$

(13)

where $s_1$ denotes the target size of the first frame. This scale estimation strategy is very simple, but it is effective at improving the accuracy of the tracking. The experimental results on the thermal infrared tracking benchmark VOT-TIR 2015 [47] prove conclusively that it works.

4.3. Model Update

The model update is a significant step in object tracking due to the fact that target object’s appearance changes dynamically. Therefore, the model needs to update to adapt the variation in the appearance. In this work, we use a simple linear update method, such as in [7], to update the filter. The method just exploits the current sample $\hat{X}^k$ to update the filter as follows:

$$
\mathcal{H}_{s+s, s}^k = \frac{\gamma}{\hat{X}^k \cdot \hat{X}^k + \lambda} \odot \hat{X}_{s+s, s}^k,
$$

$$
\mathcal{W}_t^k = (1 - \gamma)\mathcal{W}_{t-1}^k + \gamma \mathcal{H}_t^k,
$$

(14)

where $\gamma$ is the learning rate, which balances the proportion between the old filter and the new filter.

5. Experiments

To demonstrate the effectiveness and favorable results of the proposed method, we carry out two group experiments on thermal infrared benchmarks. We first introduce the implementation details of the experiment in Section 5.1, then present the evaluation criteria for the experiment in Section 5.2, and finally conduct two group experiments on the thermal infrared benchmarks VOT-TIR 2015 [47] in Section 5.3 and VOT-TIR 2016 [48] in Section 5.4. $^1$

$^1$If the paper is accepted, we will release the source codes.
5.1. Implementation Details

Algorithm 1 shows the main steps of the proposed tracker (MCFTS). We give the corresponding parameters as follows. The VGG-Net-19 [45], which is trained on ImageNet [46], is used to extract the features of the thermal infrared target object. Six convolution layers are exploited in our method, namely Con4_2, Con4_3, Con4_4, Con5_2, Con5_3, and Con5_4. We crop a search region with 1.5 times the size of the target bounding box. Then, we resize it to 224 × 224 pixels and sent it into the VGG-Net-19 to extract the feature maps of this search region. For a weak tracker based on a correlation filter in our method, we use the same parameters as KCF [1]: regularization parameter $\lambda = 10^{-4}$ in Eq. 3, the spatial bandwidth of the kernel is set to 0.1, the learning rate $\gamma = 0.01$ in Eq. 14, and cell size is set to 4. There are no parameters in the ensemble method. For the scale estimation strategy, we set the weights of the directions $\beta_d = \{0.94, 1, 0.94\}$ in Eq. 11 and the scale change step sizes $R_d = \{0.95, 1, 1.02\}$ in Eq. 12. We conduct the experiment in MATLAB R2015b and use the MatConvNet toolbox [62] in this work. The proposed method runs at 4.6 frames per second averagely on a PC with an Intel i7-6700K 4.0GHz CPU, 32G RAM, and a GeForce GTX 1080 GPU card.

**Algorithm 1** The proposed tracker (MCFTS)

1: **Inputs**: initial target position $(x_1, y_1)$, target size $s_1$, VGG-Net and convolution layers $nlayers = [layer1, layer2, \ldots, layern]$; scale factor $\gamma^1 = 1$.  
2: **Outputs**: The optimal target state $[position(x_t^1, y_t^1), s_t^1]$ at the search frame $t + 1$.  
3: Initialize $numel(nlayers)$ filters using Eq. 3.  
4: **for** $t = 2$ to length(sequence) **do**  
5: Get three different scale target’s search regions.  
6: Extract $numel(nlayers)$ convolutional feature maps of each scale’s target using the pre-trained VGG-Net.  
7: Calculate $numel(nlayers)$ response maps of each scale’s target using the weak tracker 4.  
8: Filter these response maps to get a more reliable response by Eq. 7.  
9: Exploit Eq. 9 to fuse the filtered response maps to obtain a stronger one.  
10: Estimate the scale change direction by Eq. 11 and get the current scale factor, target size by Eq. 12 and Eq. 13.  
11: Find the target’s optimal position by Eq. 10.  
12: Use Eq. 14 to update $numel(nlayers)$ filters.  
13: **end for**

5.2. Evaluation Criteria

Two weakly correlated evaluation criterion: accuracy and robustness, are exploited due to their high interpretability [63, 64]. Accuracy measurement is the average overlap, which is calculated from the overlap between the predicted bounding box and the ground-truth. Given an image sequence with $m$ frames, the predicted bounding box $B_t$ and the ground-truth $G_t$, where
\( t = \{1, 2, \cdots, m\} \) denotes the \( t \)-th frame of this sequence, the accuracy of this sequence is computed by:

\[
\text{accuracy} = \frac{1}{m} \sum_{t=1}^{m} \frac{\| B_t \cap G_t \|}{\| B_t \cup G_t \|}.
\]  

(15)

The final overall accuracy on the benchmark datasets is the average value of the accuracy of all the sequences. Robustness measurement is the average number of failures during the tracking. A failure occurs when the overlap between the predicted bounding box and the ground-truth turns out to be zero. To obtain a stochastic tracker, which returns different tracking results each time, this should be run 15 times and calculate the average value to reduce the bias. The accuracy/robustness (AR) ranking plot and the AR raw plot [65] for baseline experiments are used to visualize the tracking results. If a tracker performs better than the others, it should be positioned in the top-right part of the plot.

5.3. Experiments on VOT-TIR 2015

In this section, we mainly demonstrate the effectiveness of the proposed method, which uses the multi-layer convolutional features. We carry out the experiment on the thermal infrared object tracking benchmark VOT-TIR 2015 [47]. Three groups of comparison experiments are designed for this demonstration.

Datasets: VOT-TIR 2015 has 20 thermal infrared video sequences and is the first standard thermal infrared tracking benchmark. It was published by the visual object tracking (VOT) committee. The annotations of the sequences are in accordance with the VOT-standard. The benchmark has six local attributes, including occlusion, dynamics change, object motion, object size change, camera motion, and neutral. The local attribute is annotated per-frame in all sequences. It can also be used to evaluate the performance of the tracker on frames with specific attributes.

Compared trackers: To evaluate the proposed tracker, we compare it with three groups of trackers that use different features. The first group of trackers use one layer convolutional features, including KCF\_Deep\_Con3\_3, KCF\_Deep\_Con4\_2, and KCF\_Deep\_Con5\_4. The second group of trackers use multi-layer convolutional features, including MCFT (Ours) that does not use the scale estimation strategy, MCFTS-3s (Ours) that uses three-layer convolutional features (Con5\_2, Con5\_3, and Con5\_4) and contains the scale estimation strategy. The last group of trackers exploit hand-crafted features, including KCF\_HOG and KCF\_Gray. All the KCF trackers using different features employ the fixed parameters in the experiment.

Evaluation and results: Three group experiments are designed to demonstrate the effectiveness of the proposed tracker. First, we compare the KCF trackers that use single layer convolutional features, i.e., KCF\_Deep\_Con3\_3, KCF\_Deep\_Con4\_2, and KCF\_Deep\_Con5\_4 with MCFT that exploits the multiple layers convolutional features. The experimental results are shown in Fig. 2, which indicates that MCFT (Ours) achieves higher accuracy and robustness than KCF\_Deep\_Con3\_3, KCF\_Deep\_Con4\_2, and KCF\_Deep\_Con5\_4. Especially, its accuracy is higher than KCF\_Con3\_3 by about twenty percent. That illustrates that a tracker using multi-layer convolutional features has more stable tracking performance than a tracker that only uses single
layer convolutional features. We suggest that the main reason for this is that the features from a single convolutional layer have limited semantic information, which is crucial for various challenges. However, the proposed method adopts multiple convolutional features, which can strengthen the semantic information to cope with the various challenges. Then, we compare the KCF trackers that use hand-crafted features, \textit{i.e.}, KCF\_HOG and KCF\_Gray with MCFT and MCFTS-3s, which use three-layer convolutional features. The results are shown in Fig. 2. It is clear that MCFTS-3s achieves the second best results while the KCF\_HOG and KCF\_Gray get the fourth and fifth ranking. This means that the proposed method is effective and the convolutional features can represent the thermal infrared target object to some extent, despite the fact that the convolutional features are extracted from the VGG-Net trained on the visible image datasets. Finally, MCFTS and MCFT are compared to prove that the adopted scale estimation strategy is working. These results are also shown in Fig. 2. Obviously, the MCFTS obtains the optimal tracking performance and improves accuracy by five percent to the MCFT. The three group experiments results, yielding results relating to different aspects, demonstrate that the proposed method is reliable and effective.

5.4. Experiments on VOT-TIR 2016

In this section, we present the results of experiments on the thermal infrared tracking benchmark VOT-TIR 2016 [48] to demonstrate that the proposed tracker achieves promising results. We also use accuracy and robustness as evaluation criteria and use the AR raw plot and AR ranking plot to visualize the experiment results.

\textbf{Datasets}: The thermal infrared tracking benchmark VOT-TIR 2016 has 25 video sequences, and is an enhanced version of VOT-TIR 2015. Some easily tracked video sequences have been deleted from VOT-TIR 2015, and several more difficult video sequences added. The annotations of the sequences are also in accordance with the VOT-standard, and are the same as those of VOT-TIR 2015. It has six local attributes which can be used to evaluate the performance of the tracker on frames with specific attributes usually.

\textbf{Compared trackers}: Fourteen popular trackers are chosen to be compared with our tracker on VOT-TIR 2016. We divide these trackers into four categories: deep correlator trackers using deep features, correlation filter based trackers using hand-crafted features, part-based trackers, and fusion trackers. For the deep correlator trackers, we select three state-of-the-art trackers: deepMKCF [66], HCF [41], and HDT [42]. All of these trackers use multi-layer convolutional features and achieve state-of-the-art performance on the object tracking benchmark (OTB) 2013 [67]. Of the correlation filter based trackers, we choose three extensional KCF trackers: SKCF [68], DSST [8], NSAMF [69]. These trackers obtain promising results on the visual object tracking benchmark (VOT) 2014 [70]. Especially, DSST was the champion on VOT 2014. Part-based trackers always gain the better performance due to their favorable characteristic of having multiple cues. Six part-based trackers are selected to compare with our tracker. They are DPT [71], FCT [48], GGT2 [72], DPCF [73], BST [74], and BDF [75]. The last two trackers are MAD [76] and LOFT-Lite [77], which are based on the fusion method. More details of these trackers can be found in the VOT-TIR 2016 report [48].

\textbf{Evaluation and results}: We compare our tracker with trackers mentioned above on VOT-TIR 2016. Firstly, the accuracy and the robustness of all these trackers are compared on the bench-
Figure 2: The ranking plots and the accuracy/robustness (AR) raw plots of the compared trackers on the thermal infrared tracking benchmark VOT-TIR 2015 [47]. The better performance a tracker achieves, the closer it is to the top-right of the plot.
mark, of which the results are shown in Fig. 3. It is evident that MCFTS (Ours) obtains the best robustness and the third best accuracy. It should be noted that the deep correlator trackers get the optimal results out of the compared trackers, which indicates that the deep features learned from the visible image can also represent a thermal infrared target object to some extent. Furthermore, the correlation filter based trackers using hand-crafted features, i.e., DSST and NSAMF, achieve satisfactory results while our tracker gets preferable performance. This also illustrates how deep features have representation ability superior to that of the hand-crafted features in thermal infrared tracking. Despite fact that the part-based trackers always have better tracking performance, i.e., DPT and GGT2 et al., our MCFTS also performs favorably against these trackers.

Secondly, we further compare these trackers with the proposed tracker at each local attribute on the corresponding subset of VOT-TIR 2016, as shown in Fig. 4. With regard to the size change and occlusion challenges, we can see that our method obtains the best robustness, as illustrated in Fig. 4(a) and 4(b). This more intuitively shows that the proposed method is effective at handling these challenges. We suggest that the favorable performance mainly benefits from the multi-layer convolutional features and the ensemble strategy. With regard to motion change and changing dynamics, Fig. 4(c) and 4(d) show that our MCFTS achieves the best robustness and the third best accuracy. This also demonstrates that our tracker achieves satisfactory results. With regard to the other challenges, our tracker also has a decent performance, as shown in Fig. 4(e) and 4(f). These group experiments demonstrate more intuitively our MCFTS is effective and robust to various challenges.

6. Conclusion

In this paper, we transfer a pre-trained CNN, trained on visible image datasets, to the thermal infrared tracking task. We exploit the pre-trained CNN to extract the features of the thermal infrared target object. Then, we analyze the fully-connected layer features and convolution layer features, which are more suitable for thermal infrared tracking. As a consequence, we find that the convolutional features have a powerful ability to represent a thermal infrared object because they have richer spatial information than the features from the fully-connected layer. However, the features from a single convolution layer are not robust to various challenges due to their lacking the semantic information. Based on this observation, we propose a correlation filter based ensemble tracker using multi-layer convolutional features for thermal infrared tracking. In this tracking model, we also employ an ensemble method based on the Kullback-Leibler divergence to fuse each part. Furthermore, we also adopt a simple but effective scale estimation strategy to improve the accuracy of the tracking. We carry out experiments on the thermal infrared tracking benchmarks: VOT-TIR 2015 [47] and VOT-TIR 2016 [48]. The experimental results demonstrate that the proposed method is effective and achieves promising performance.

Acknowledgments

This work is supported by the Shenzhen Research Council(Grant No. JCYJ20170413104556946, JCYJ20160406161948211, JCYJ20160226201453085, JSGG20150331152017052), the National
Figure 3: The ranking plots and the accuracy/robustness (AR) raw plots of the compared trackers on the thermal infrared tracking benchmark VOT-TIR 2016 [48]. The better performance a tracker achieves, the closer it is to the top-right of the plot.
Figure 4: The ranking plots and the accuracy/robustness (AR) raw plots for the baseline experiments on each attribute subset of the thermal infrared tracking benchmark VOT-TIR 2016 [48]. The better performance a tracker achieves, the closer it is to the top-right of the plot.
Natural Science Foundation of China (Grant No. 61672183, 61272252), the Science and Technology Planning Project of Guangdong Province (Grant No. 2016B090918047), and the Natural Science Foundation of Guangdong Province (Grant No. 2015A030313544).

References


