IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 30, 2019, accepted July 15, 2019, date of publication July 18, 2019, date of current version August 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929749

An Optimal Long-Term Aerial Infrared Object Tracking Algorithm With Re-Detection

XIAOTIAN WANG[1], KAI ZHANG[1], SHAOYI LI[1], YANGGUANG HU[2], AND JIE YAN[1]
[1]School of Astronautics, Northwestern Polytechnical University, Xi'an 710072, China
[2]School of Astronautics Engineering, Air Force Engineering University, Xi'an 710072, China

Corresponding author: Xiaotian Wang (18710993786@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61703337 and in part by the Shanghai Aerospace Science and Technology Innovation Fund under Grant SAST2017-082.

ABSTRACT In the field of automatic target recognition and tracking, long-term tracking for aerial infrared target has been recently seen with great interest. Although deep trackers and correlation filtering trackers offer competitive results on performance, the problems of deformation, abrupt motion, heavy occlusion, and out of view still remain unsolved. In addition, since this paper focus on infrared images, it is also important to consider that infrared images have a significant drawbacks, such as low resolution, low contrast, and lack of textures. In this paper, we adopt correlation filtering trackers and deep learning detection method to achieve accurate tracking results. Our tracking system composed of three parts: the DTB correlation filtering tracker (DTB-CF), a better regression model to discriminate the target from the background with adjustable Gaussian window functions; the UTA correlation filtering tracker (UTA-CF), an optimum regression model to update the target appearance with simultaneously optimal in position, scale, and integration of multi-feature fusion; and the YOLOv3 re-detector, which ensures re-location of the correct position of the target when the tracking fails. In addition, we introduce the ratio between average peak-to-correlation energy (APCE) of the current frame and average APCE of former frames as a criterion to update the UTA-CF tracker to maintain the target model stability. And we combine the nearest neighbor maximum value method with APCE as criterion together to initialize the YOLOv3 re-detector. We evaluate our algorithm on real aerial infrared target thermal image sequences in terms of precision plot, success plot, and speed. The experimental results show that our method has a significant improvement than the state-of-the-art methods for long-term tracking both in accuracy and robustness for aerial infrared object tracking.

INDEX TERMS Aerial infrared object tracking, correlation filtering, deep learning detection, multi-feature fusion, APCE criterion.

I. INTRODUCTION

Object tracking is one of the most fundamental concerns in computer vision, which has been widely used in the fields of surveillance, human computer interaction, behavior recognition and unmanned driving. Given the initialized state (e.g., position and size) of a target object in the first frame, the task of object tracking is to predict the states of the target in the subsequent frames. Infrared image has the advantages of unsusceptible to illumination, strong anti-interference ability and all-weather working. Infrared object tracking is becoming a popular research topic and is also employed in various military, scientific and medical areas. In this paper, we focus on the problem of long-term aerial infrared object tracking.

Compared with visual tracking, infrared object tracking is more challenging. It needs not only to solve the problem of universal tracking (e.g., deformation, abrupt motion, heavy occlusion and out of view), but also to consider its significant defects with low resolution, low contrast and lack of textures. An effective and real-time tracking algorithm should be able to consistently track the infrared object for a long time without failing under these situations.

At present, the mainstream methods of tracking algorithms are based on two types: the first is the traditional correlation filtering method, and the other is the convolutional neural network method. The convolutional neural network method has powerful capability of feature extraction.

The associate editor coordinating the review of this manuscript and approving it for publication was Shangce Gao.

Instead of extracting hand-crafted features, the tracking algorithms benefits from this method can directly use pre-trained convolution neural network (CNN) to extract convolutional features [1]–[3]. Moreover, in [4]–[10], online updating model for tracking is built without pre-trained CNN. Whereas, the lack of training data and real-time requirement limit the application of convolutional neural network method in engineering. Correlation filters are not sensitive to visual and infrared images of different spectra, which are more suitable for infrared object tracking [11]. In addition, the methods based on correlation filtering are very effective for their fast calculation speed, outstanding real-time performance, and high precision. The basic idea of correlation filtering method is to find a filtering template and make this filtering template convolve with next frame, and the region with the largest response is the predicted target. According to this idea, a large number of tracking algorithms based on correlation filtering have been proposed. Bolme *et al.* [12] propose to learn a minimum output sum of squared error (MOSSE), which produces stable correlation filters when initialized by using a single frame, where the learned filter encodes target appearance with updates on every frame. MOSSE algorithm forms multiple samples with random affine, which will cause redundancy; Henriques *et al.* [13] propose to adopt the theory of circulant matrices for fast detection and update with the Fast Fourier Transform. The CSK method builds on illumination intensity feature, as an enhancement, is further improved by using HOG feature in the KCF tracking algorithm [13]. In order to solve scale change problem, Danelljan *et al.* [14], [15] present a novel approach by learning separate filters for translation and scale estimation. Good feature is important for object tracking, so it is essential that the target is represented by multi-feature fusion. Li and Zhu [16] present a scale adaptive kernel correlation filter tracker with feature integration (SAMF) by integrating the powerful features including HOG and color-naming to further boost the overall tracking performance. The boundary effect seriously affects the performance of the correlation filtering algorithm. Danelljan *et al.* [17] adopt a spatial regularization technique to weigh the filter coefficients, and boundary effect is effectively alleviated.

Advancements in the correlation filtering method mostly focus on incorporating robustness to specific challenges such as scale change, illumination change, boundary effect, but fail to track under other conditions like deformation, occlusion, varied distractors etc. To successfully track the target all the time, three key issues should be addressed in long-term tracking. The first one is the well-known stability-plasticity dilemma. It is related to online update mechanism of correlation filtering algorithm. If the learning rate is too high and frequent, filtering template easily result in drifting due to noisy updates. On the contrary, if the learning rate is too low and conservative, the target has been distorted, but the filtering template is still the same template. It is a fact that the target will not be recognized. The second issue is to extract better features to make the target easier to identify and multi-feature fusion can improve tracking performance [18]. The third issue is to judge whether the tracker fails to locate the target and add a re-detector in tracking process. When the tracking fails, the re-detection is performed to obtain the correction position for successful tracking.

Our approach builds on these three observed issues. We adopt correlation filter and re-detection to track the target that may disappear, deform, or occlusion heavily in long-term tracking. Aiming at stability-plasticity dilemma, our algorithm effectively alleviates this dilemma by using two regression models based on correlation filters with different learning and updated rates to update the target appearance and discriminate the target from the background. The regression model to discriminate the target from the background is called DTB-CF based on KCF. It is designed to aggressively adapt to translation estimation against significant deformation and heavy occlusion, and we extend the DTB-CF trackers with the capability of handling scale change by introducing adjustable Gaussian window functions for better back-and-foreground separation around the target, leading to increased accuracy and robustness. The regression model to update the target appearance, UTA-CF is sampled with different scales, and is conservatively adapted and applied for optimal location and optimal scale. Aiming at improving tracking performance with better features, we can employ various powerful features to exploit the advantages of multi-feature fusion. We integrate HOG, histogram feature of intensity, and histogram feature of local intensity for DTB-CF tracker, and HOG, color-naming, and gray feature for UTA-CF tracker. In the literature of [19], it merely adopted the maximum response as the re-detection criterion, however, the information of the fluctuating response map will be lost. Therefore, we introduce the APCE criterion to evaluate the fluctuation degree in the tracking process to determine whether the tracker needs to be updated or initialized by re-detector. As we know, deep learning could act as state-of-the-art in the detecting process. It could learn very general representation of objects. YOLOv3 [20]–[22], one of the most balanced target detection networks for speed and accuracy, is used as the re-detector. This work aims to address the problem of deformation, occlusion, varied distractors etc., and achieve better performance in long-term aerial infrared object tracking.

The structure of this paper is as follows. Section II first present the closely related works to our long-term aerial infrared object tracking framework. Section III presents the general framework of the proposed tracking approach and its main modules in detail. The proposed approach is validated by experiments in Section IV. Section V draws the conclusions of this paper.

II. RELATED WORKS

This section introduces the state-of-the-art tracking and detection approaches, which are closely related to this work. Moreover, we explain the reason for using a structure of combing correlation filter and re-detection to achieve long-term track, and introduce the idea of verification in tracking, which inspired us to design the structure of our approach.

A. TRACKING ALGORITHMS

A lot of tracking algorithms can be found in current literature. Many of them have been evaluated with the Visual Tracker Benchmark [23], [24] and in the VOT challenge [25]. We restrict our review here to those tracking algorithms that are close to our work, including Struck [26], TLD [27], MOSSE [12], CSK [28], KCF [13], CN [29], DSST [15], SAMF [16], BACF [30], SKCF [31], OMFL [32], C-COT [33], ECO [34], LCT [19], LMCF [34], CA-CF-SVM [35], and CTAD [36].

Struck, one of the most representative trackers relies on a kernelized structured output Support Vector Machine (SVM) to distinguish between the tracked object and the background. It achieves appealing results. TLD combines the traditional tracking algorithm with detection algorithm to solve the problem in long-term tracking. At the same time, the parameters of the tracking module and the target template of the detection module are constantly updated by the online learning mechanism, making the tracking outcome more stable, robust and reliable.

However, the computational cost of Struck and TLD is high which makes them less appealing compared to the correlation filtering algorithm. In addition, the correlation filtering algorithms MOSSE, CSK, and KCF outperform both of the trackers mentioned above in Visual Tracker Benchmark and VOT challenges, while also being significantly faster. Last but not least, when the correlation filtering algorithm is applied to the infrared object tracking, the tracking performance will hardly decline [11]. Therefore, we focus on correlation filtering tracking algorithms.

Danelljan *et al.* [29] exploit the color attributes of a target object and learn an adaptive correlation filter by mapping a multi-channel features into a Gaussian kernel space. The DSST tracker learns adaptive multi-scale correlation filters using HOG features to handle the scale change of objects. The BACF tracker is capable of learning/updating filters from real negative examples densely extracted from the background instead of shifted foreground patches, achieving superior accuracy with real-time performance. The SKCF tracker uses the Gaussian window to create a better separation of the target and the background, improving accuracy and getting the same result with improved BACF algorithm [37]. The OMFL tracker uses novel features, i.e., intensity, color names, and saliency, to respectively represent both the tracking object and its background information in a background-aware correlation filter (BACF) framework instead of only using the histogram of oriented gradient (HOG) feature to get better tracking effect. The multi-feature learning approach is able to improve the object tracking performance. However, these methods do not resolve the critical issues regarding online model update. We also explored the implication of C-COT and ECO which are excellent algorithms while the high computational complexity cannot meet real-time demand. Therefore, these correlation filtering tracking methods are susceptible to drifting and less effective for handing long-term occlusion and out-of-view problems. Our DTB-CF tracker is designed to aggressively adapt to translation estimation against significant deformation and heavy occlusion with high and frequent learning rate. Based on the result of DTB-CF tracker, our UTA-CF tracker uses target regressor to get optimal target position and scale information with low and conservative learning rate, achieving target accurate evaluation. Meanwhile we integrate HOG, histogram feature of intensity, and gray feature for DTB-CF tracker, and HOG, color-naming, and gray feature for UTA-CF tracker. Therefore, our approach effectively adapts to appearance changes and alleviates the risk of drifting. Virtually, our DTB-CF is derived from the fusion of BACF and SKCF trackers, and it relies on the Gaussian window to retain foreground information, exploiting background patches together with the target patch to train the tracker. The UTA-CF tracker is derived from the SAMF tracker, which can get optimal object tracking result with simultaneously optimal in position and scale.

To judge whether the tracker needs to be updated or initialized by the re-detector, the LCT tracker adopts the maximum response as a criterion. However, it would lose the information of the fluctuating response map. The LMCF and CA-CF-SVM trackers use APCE as the criterion, which indicates the fluctuated degree of response maps and the confidence level of the detected targets. The CTAD tracker proposes an algorithm composed of a tracker based on the structure of the LCT tracker and a deep learning detecting method based on the YOLOv3, which improves the tracking performance. Therefore, we adopt correlation filter and re-detection to track the target, where we also use YOLOv3 as the re-detector, which could take advantage of both deep tracker and high-speed correlation filter. At the same time, we innovatively improve the APCE criterion to initial the re-detector (YOLOv3) and update the UTA-CF tracker, achieving better experimental results.

B. DEEP DETECTION ALGORITHMS

Due to the advantages of deep convolutional network in feature expression, researchers have gradually combined deep convolutional network with target detection. R-CNN [38], using convolutional neural network to express the nature attributes of the target, is the earliest method to achieve better detection effect. Based on the R-CNN detector, Fast R-CNN [39] and Faster R-CNN [40] appeared one after another, but they still cannot satisfy the real-time demand. YOLO [20] regards the target detection process as a regression problem, which has satisfying real-time performance. Aiming at the poor detection rate of the YOLO algorithm, the author of YOLO proposed various improvements to the YOLO detection method called YOLO9000 [21]. YOLO9000 is greatly advanced in terms of recognition rate. In [22], the author makes some changes to YOLO9000. Compared with the two previous methods, the YOLOv3 is faster and more accurate. The real-time performance and accuracy of YOLOv3 is the best.

In summary, we adopt two correlation filters (DTB-CF and UTA-CF) and a re-detection method based on the YOLOv3 to track the target. As a result, it could take advantage of both deep tracker and high-speed correlation filter. At the same time, we innovatively improve the APCE criterion to initialize the YOLOv3 re-detector and update the UTA-CF tracker.

III. METHODOLOGY

In the following section, we would like to introduce the details of our algorithm. The basic idea of our algorithm is the combination of tracker and re-detector. We evaluate the tracking effect and re-detection initialization by its confidence level. Our DTB-CF tracker is designed to aggressively adapt to translation estimation against significant deformation and heavy occlusion. The confidence level will not influence online update mechanism of DTB-CF correlation filtering algorithm whose filtering template will be updated all the time. Our UTA-CF tracker uses target regression to get optimal target position and scale information. When the confidence level is higher than the threshold, the UTA-CF will be updated, getting accurate target filtering template. When the confidence level is lower than the threshold, the target may be significantly deformed and heavily occluded, and the tracker cannot continuously track the target; we re-detect the target in the current frame and re-locate the correct position of the target. Therefore, judging the tracking confidence of the target is an important aspect. We use the APCE criterion to make judgments. Firstly, we would like to introduce the framework of our algorithm. Then we would like to present the specific implementation procedures. The DTB-CF correlation filtering algorithm is introduced in section B, and we show UTA-CF correlation filtering algorithm in section C. In section D, re-detection based on the YOLOv3 is presented. In section E we improve the APCE criterion and give the details of its usage.

A. FRAMEWORK

The algorithm mainly contains four parts: DTB-CF, UTA-CF, YOLOv3 and APCE criterion. These four parts are associated with each other to achieve real-time and efficient infrared target long-term tracking. Illustration of the algorithm is show in Figure 1.

- DTB-CF tracker: Regression model to discriminate the target from background. Our DTB-CF tracker is designed to aggressively adapt to translation estimation against significant deformation and heavy occlusion. The adjustable Gaussian window functions are introduced for better back-and foreground separation around the target, leading to increased accuracy and robustness.
- UTA-CF tracker: Regression model to update the target appearance. The features of HOG, color-naming and gray are integrated for UTA-CF tracker. An optimal object tracking result with simultaneously optimal in position and scale was obtained, and we can get accurate tracking result.
- YOLOv3 re-detector: Re-detector to re-locate the correct position of the target. YOLOv3 is based on deep learning, which can support an accurate detection result for its strong ability of characteristic expression. When the tracking fails, it is important for YOLOv3 to search for the target again.
- APCE: Criterion that the tracker needs to be updated or initialized by the re-detector. We use the ratio between APCE of the current frame and their historical average values as criterion to update the tracker to maintain the target model stability. And we combine the nearest neighbor maximum value method with APCE as criterion together to initialize the YOLOv3 re-detector.

The DTB-CF tracker is initialized in the first frame and we can get the translation position of the target. Based on the information of translation position of the target, UTA-CF tracker realizes an optimal object tracking result with simultaneously optimal in position and scale. Meanwhile we can achieve the confidence values, and we use the improved APCE criterion to make judgment about when the tracker needs to be updated or initialized by the re-detector. With the implication of the re-detector, the tracker could get verification. Then, we still use the tracker to track the target until the end of the process. The DTB-CF model is updated frame by frame, which is not affected by the APCE criterion. The tracking algorithm has been evaluated in our infrared image sequences, which could track the target properly, even in a complex background.

B. DTB-CF TRACKER

The DTB-CF tracker is proposed to aggressively adapt to translation estimation against significant deformation and heavy occlusion. It is derived from the fusion of BACF and SKCF trackers, and which has the advantages of the two algorithms, so it is not only robust to the problems such as rotation, fast motion, and background clutter, but also has strong capability for occlusion and deformation in long-term tracking. We set it as the baseline algorithm to aggressively adapt to translation estimation against significant deformation and heavy occlusion.

A typical tracker based on correlation filtering models the appearance of a target object by using a filter trained on an image patch x of M*N pixels, where all the circular shifts of $x_{m,n}$, $(m, n) \in$ {0, 1, . . . , M-1}*{0, 1, . . . , N-1}, are generated as training samples with Gaussian function label $y(m, n)$,i.e.,

$$w = \arg\min_{w} = \sum_{m,n} (f(x_{m,n}) - y(m, n))^2 + \lambda \|w\| \qquad (1)$$

where f denotes the result of correlation between the training samples and the filter ω, and λ is the regularization coefficient which prevents overfitting. According to the Representer theorem [41], the objective function can be expressed as

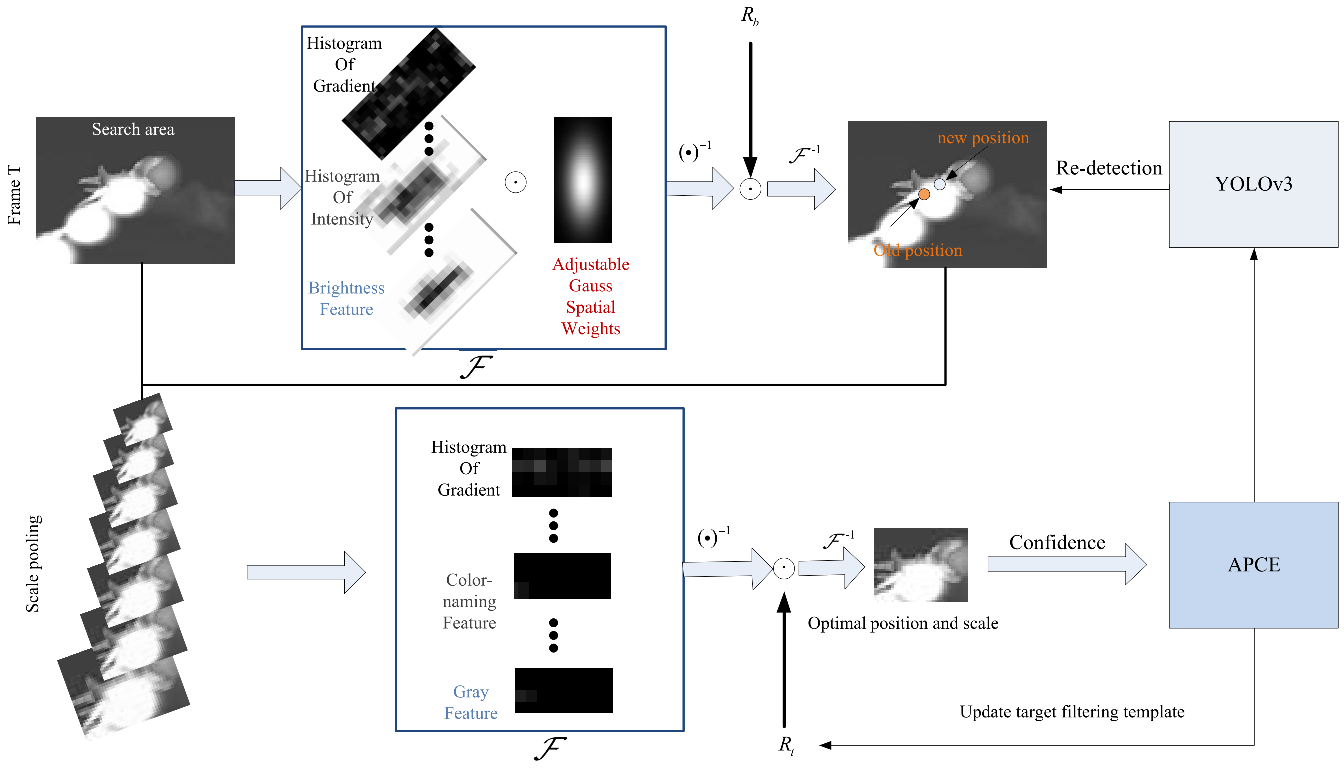


FIGURE 1. Algorithm framework.

follows:

$$w = \sum_{m,n} \alpha(m,n) f(x_{m,n}) \tag{2}$$

where α is the filter coefficient, which is defined by

$$\alpha = \mathcal{F}^{-1}[\frac{\mathcal{F}(y)}{\mathcal{F}(k_{xx}) + \lambda}] \tag{3}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denotes the discrete Fourier operator and inverse Fourier operator and $\{y(m,n)| (m,n) \in \{0, 1, \ldots, \text{M-1}\}*\{0, 1, \ldots, \text{N-1}\}\}$. k_{xx} is defined as kernel correlation in KCF tracker. In this paper, we adopt the Gaussian kernel which can apply the circulant matric trick as below:

$$k_{xx} = \exp(-\frac{1}{\sigma^2}(\|x\|^2 + \|x\|^2 - 2\mathcal{F}^{-1}(\mathcal{F}(x) \odot \mathcal{F}(x^*)))) \tag{4}$$

where $\odot$ is the Hadamard product. The tracking task is carried out on an image patch z in the new frame with the search window size M*N by computing the response map as

$$y = \mathcal{F}^{-1}[\mathcal{F}(k_{xz}) \odot \mathcal{F}(a)] \tag{5}$$

Therefore, the new position of target is detected by searching for the location of the maximal value of y.

Contrasting from prior work, as shown in Figure 2, we train the DTB-CF tracker by taking both the target and surrounding background into account to get the filtering model R_b, since this information remains temporally stable and useful to discriminate the target from the background in the event of heavy occlusion and significant deformation. Meanwhile with the capability of handling scale change by introducing adjustable Gaussian window functions for better back-and-foreground separation around the target leading to increased accuracy and robustness. The multi-feature fusion is the same with LCT tracker, which use feature vectors with 47 channels, including Hog, histogram feature of intensity and the histogram of local intensity. The R_b model is updated with a big and frequent learning rate η frame by frame as

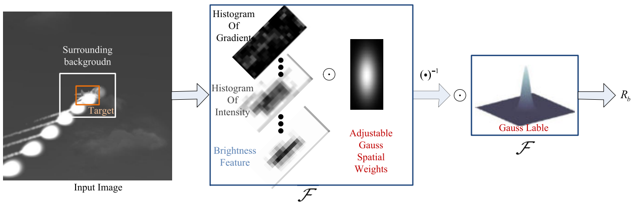


FIGURE 2. DTB-CF regression model learned from a single frame.

$$\begin{aligned} \mathcal{F}^t(x) &= (1-\eta)\mathcal{F}^{t-1}(x) + \eta\mathcal{F}^t(x) \\ \mathcal{F}^t(\alpha) &= (1-\eta)\mathcal{F}^{t-1}(\alpha) + \eta\mathcal{F}^t(\alpha) \end{aligned} \tag{6}$$

where t is the index of the current frame.

It is well known that the purpose of windowing is usual to isolate the signal of interest while reducing the frequency leakage in image processing. It is important for tracking to separate the target signal from the background. When the frequency spectrum of the original target signal mixes other frequencies, it easily results in frequency leakage. Adjustable windows are functions that are capable of reducing the frequency leakage while controlling the bandwidth, which can guarantee that the frequency spectrum of a measured input is what we want to analyze. The Fourier operator of Gaussian is also a Gaussian which ensures the separation between foreground and the background while reducing the frequency leakage. However, the cosine window has no such property. Therefore, Gaussian window has been widely used for window filter for its benefits.

$$G(m, n, \sigma_w, \sigma_h) = g(m, \sigma_w)^* g(n, \sigma_h) \tag{7}$$

The function $g(N, \sigma)$ returns a vector of size N computed as follow:

$$g(N, \sigma) = \exp(-\frac{1}{2}(\frac{i-(N+1)/2}{\sigma})^2), \quad 1 \leq i \leq N \tag{8}$$

The bandwidth σ of the Gaussian function $g(N, \sigma)$ is computed independently for the horizontal and vertical orientations. σ_w and σ_h are calculated by the side length ratio of target to tracking area in horizontal and vertical orientations. The replaced cosine window is computed as

$$\cos\_window = \frac{1}{2}(1 - \cos(2\pi \frac{i}{N})), \quad 1 \leq i \leq N \tag{9}$$

The gaussian and cosine window filtering raw pixel value example is shown in Figure 3.

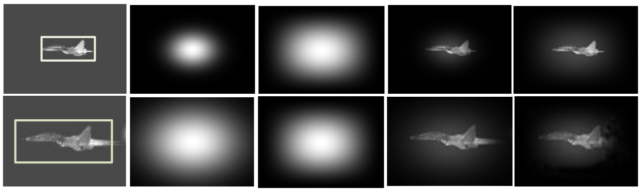


FIGURE 3. Gaussian and cosine window filtering raw pixel value example. The tracked region has a size of (130*186). First column: Same region with targets at two different scales (i.e., small (70*34), large (124*56)). Second column: Gaussian windows according the target size (i.e., small ($\sigma_w = 0.54$, $\sigma_h = 0.18$), large ($\sigma_w = 0.96$, $\sigma_h = 0.3$)). Third column: Cosine window (i.e., size (130*186)). Fourth and fifth columns: Images filtered with Gaussian and cosine windows respectively. Figure shows how the fixed cosine window fails to represent the target compared to the Gaussian windows. The cosine window includes background for small targets and discards information for big targets.

C. UTA-CF TRACKER

As shown in Figure 4, the UTA-CF tracker is derived from the SAMF tracker, which can get optimal object tracking result with simultaneously optimal in position and scale. We integrate HOG, color-naming and gray feature for the UTA-CF tracker, and train UTA-CF by taking the most reliable target appearance into account to get the filtering model R_t, which is an accurate target filtering template. Specifically, we use improved APCE criterion to update the tracker to maintain the stability of the model. Only the filtering model with high confidence could be updated. It is worth noting that the cosine window is not being used in the UTA-CF tracker. We use the scaling pool S = {1 0.975 0.98 0.985 0.99 0.995 1.005 1.01 1.015 1.02 1.025 1.03 1.035}. The σ used in Gaussian function is set to 0.8.

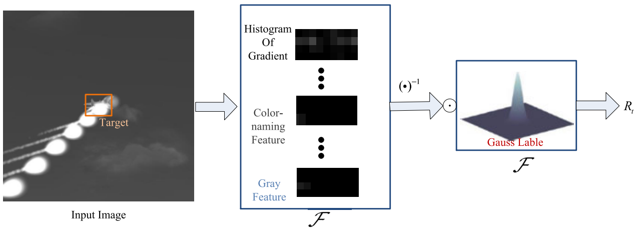


FIGURE 4. UTA-CF regression model learned from a single frame.

D. YOLOV3 RE-DETECTOR

As we all know, the real-time performance and accuracy of YOLOv3 is the best. It uses a single convolutional network which simultaneously predicts multiple bounding boxes and class probabilities for those boxes, then trains the network on full images, and directly optimizes detection. YOLOv3 re-detector has many advantages, such as high speed and robustness.

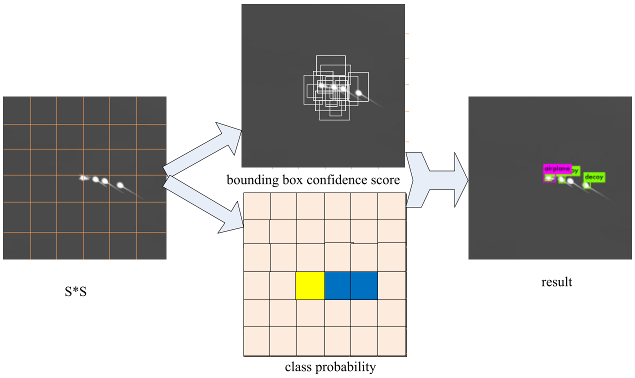


FIGURE 5. YOLOv3 block diagram.

As shown in Figure 5, firstly, the input image is divided into an S*S grid. If the center of an object falls into a grid cell, that grid cell is responsible for detecting that object. Each grid cell predicts B bounding boxes and confidence scores for those boxes. These confidence scores reflect how confident the model is that the box contains an object and also how accurate it thinks the box is that it predicts. At the same time, each grid cell also needs to predict C class probability, which predict the percentage of the grid cell containing different objects.

YOLOv3 predicts 4 coordinates for each bounding box, t_x, t_y, t_w, t_h. If the cell is offset from the top-left corner of the image by (c_x, c_y) and the bounding box prior has width and height (p_w, p_h), then the predictions correspond to:

$$\begin{cases} b_x = \sigma(t_x) + c_x \\ b_y = \sigma(t_y) + c_y \\ b_w = p_w e^{t_w} \\ b_h = p_h e^{t_h}. \end{cases} \tag{10}$$

YOLOv3 predicts an objectness score for each bounding box using logistic regression. If the bounding box prior overlaps a ground truth object by more than any other bounding box prior, it should be 1, and vice versa. In summary, YOLOv3 is good re-detector, which is fast and accurate. And we use improved APCE criterion to initialize the YOLOv3 re-detector.

E. UPDATE MECHANISM

In order to get accurate target filtering template, when the target disappear, deform, or occlusion in the view, we should not update object tracking model of the UTA-CF tracker. This may ensure the target filtering template is less prone to drifting caused by model update with noisy samples. At the same time, when the tracking fails, it is important for tracker to search the target again. Aiming at this problem, a good re-detector is essential. Therefore, the criterion for updating target filtering template and initializing the re-detector is very important. There are two schemes to deal with the tracking confidence of the target. The first is the maximum response value of the current frame, and the second is the discriminative algorithm defined by the response map named average peak-to-correlation energy (APCE) [34]. When the target is occluded severely, the response map fluctuates fiercely, while the maximum response value remains strong enough. If we choose to update the model in this frame, then the tracking model will be corrupted. At present, the APCE is regard as the best criteria, so we use APCE to deal with the tracking confidence of the target. We introduce the APCE criterion to evaluate the fluctuation degree in the tracking process to determine whether the tracker needs to be updated or initialized by re-detector. Specifically, we use the ratio between APCE of the current frame and their historical average values as criterion to update the UTA-CF tracker to maintain the target model stability. And we combine the nearest neighbor maximum value method with APCE as criterion together to initialize the YOLOv3 re-detector.

The APCE criterion is defined as follows:

$$APCE = \frac{|F_{\max} - F_{\min}|^2}{mean(\sum\limits_{wid,hei} (F_{wid,hei} - F_{\min})^2)} \tag{11}$$

where $F_{\min}$, $F_{\max}$, and $F_{wid,hei}$ denote the minimum, maximum and width row height column response value respectively. The maximum response value of the current frame cannot assess the confidence of the tracking for it ignores the fluctuations in the tracking process. The APCE value reflects the degree of fluctuation of the response map and the confidence of the target object. When there is less noise, the value of APCE becomes larger, and the response map gets smoother except for only one peak, and vice versa.

When the ratio between APCE of the current frame and their historical average value is larger than a certain threshold 0.8, the tracking result in the current frame is considered to be high-confidence, which shows that the target filtering template is accurate without any deform or occlusion. Then the UTA-CF tracker will be updated online with a learning rate parameter ξ. Despite UTA-CF tracker has high confidence of the target model, the confidence maximum shows a decreasing trend in the process of occlusion and re-capture. A certain threshold cannot be a criterion for tracking failure, but tracking failure is related to the nearest neighbor maximum value of APCE. When the ratio between their historical average value and the nearest neighbor maximum value of APCE in the current frame is less than a certain threshold 0.8, the re-detector is performed. This can increase the performance of the algorithm in long-term tracking. Figure 6 illustrates the update mechanism process. And an outline of our method is presented in Algorithm 1.

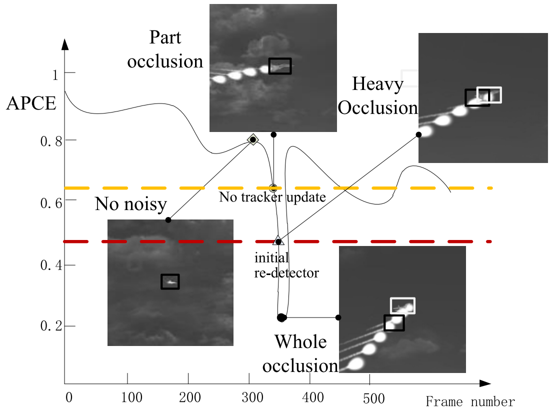


FIGURE 6. **Online update mechanism.**

Algorithm 1 Proposed Tracking Algorithm

```
Input: Initial target bounding box $X_0$.
Output: Estimated object state $X_t = (x_t, y_t, s_t)$,
        DTB-CF regression model $R_b$,
        UTA-CF regression model $R_t$, and YOLOv3 detector.
repeat
    1) Crop out the search window in frame t according to
       $(x_{t-1}, y_{t-1})$ and extract the features.
    2) Do translation estimation by DTB-CF regression
       model Rb to get the new position $(x_t^*, y_t^*)$.
    3) Building scaling pool around $(x_t^*, y_t^*)$. and do
       optimal estimation by UTA-CF regression model
       Rt to get optimal simultaneously position and scale
       $(x_t, y_t, s_t)$, and APCE also can de achieved.
    4) Improved APCE satisfies UTA-CF update mechanism,
       the target filtering template $R_t$ will be updated.
    5) Improved APCE satisfies YOLOv3 re-detector
       update mechanism, the candidate states $x'$ will be
       used.
Until End of video sequences.
```

IV. EXPERIMENTAL RESULTS

We conduct four experiments to evaluate the efficacy of our proposed algorithm. Firstly, we implemented four trackers with various settings, including Adaptive Gauss Weighted tracker (AGW), Optimal Position and Scale tracker (OPS), Re-detector (LCT-YOLOv3) and the proposed integrated algorithm. Finally, we compare them with a relevant tracker LCT. Secondly, we evaluate our proposed tracker against the state-of-the-art tracker to show the effectiveness of algorithm. Additionally, we compare our algorithm with five representative trackers on challenging sequences for qualitative evaluation and compare the center location error frame-by-frame to show the validity of the proposed method. Experimental data is derived from real aerial infrared thermal images made up of 8 sequences, public database named The Thermal Infrared Visual Object Tracking challenge 2016 (VOT-TIR2016) and Army Missile Command (AMCOM), and infrared simulation images. Figure 7 shows some samples of aerial infrared thermal image sequences.

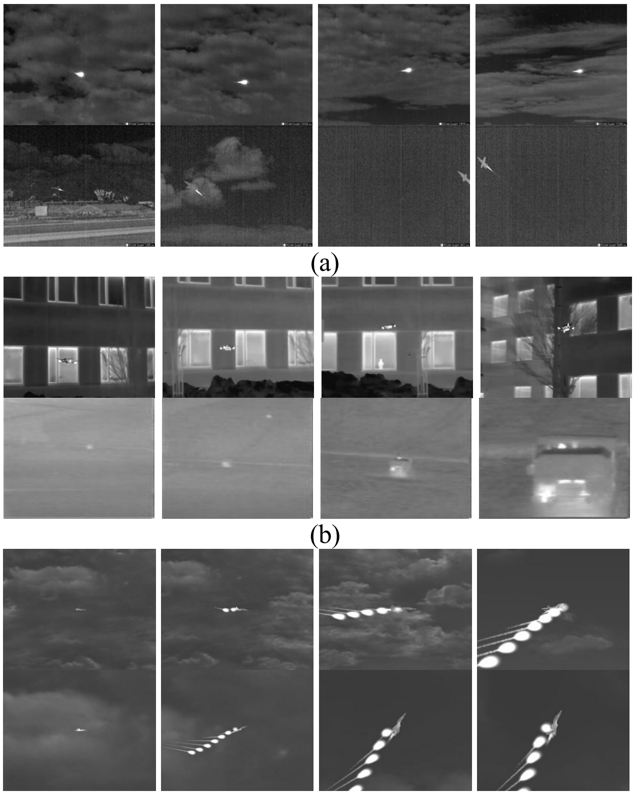


FIGURE 7. Some samples of aerial infrared thermal image sequences. (a) Aerial infrared thermal images. (b) VOT-TIR2016 and AMCOM infrared images. (c) Infrared simulation images.

All the tracking methods are evaluated by two metrics. The first one is precision plot, which shows the percentage of image frames whose tracked location is within the given threshold distance of ground truth. The second metric is success plot, which shows the degree of overlap between the predicted target area and the real target area. The overlap score is defined as $S = \left|B_t \cap B_{gt}\right| / \left|B_t \cup B_{gt}\right|$, where B_t is the tracking bounding box, B_{gt} is the ground truth bounding box, and $\cap$ and $\cup$ denote the intersection and union operators. To verify the real-time performance of the algorithm, we introduce the FPS (frames per second) with the following equation:

$$FPS = \frac{\sum_{i=1}^{n} N_i}{\sum_{i=1}^{n} t_i} \tag{12}$$

where N_i represent the length of the image sequence, t_i denotes the cost time, and i denotes the frame index of the image sequence.

A. EXPERIMENT SETUP

We implemented the proposed tracker using MATLAB and Darknet on an Intel i7-6700 CPU (2,80 GHz) PC with 4 GB memory. The re-detector YOLOv3 is performed with Darknet and the others are implemented in MATLAB. The two networks communicated with each other via a simple TCP-IP socket. The re-detector is based on YOLOv3, which was trained with a training dataset consisting of 1500 infrared labeled images, which is selected from each sequence. We selected one for a training dataset every ten frames. The iteration times is 5000.

We use the same parameter configurations for all implementations as described in [9]. The regularization parameter of (1) is still set to $\lambda = 10^{-4}$. The used in Gaussian function is set to 0.8. However, there are some changes in our proposed method. The sizes of the search window for DTB-CF tracker and UTA-CF tracker are set to 1.8 times and 1.2 times of the target size respectively. The learning rates η and ξ for DTB-CF and UTA-CF trackers are set to 0.01 and 0.008 respectively. We use the scaling pool S ={1 0.975 0.98 0.985 0.99 0.995 1.005 1.01 1.015 1.02 1.025 1.03 1.035}. All parameters are same for all following experiments.

B. ABLATION STUDY

We evaluate the performance of our four implementations trackers, which is made up of Adaptive Gauss Weighted tracker (AGW), Optimal Position and Scale tracker (OPS), Re-detector (LCT-YOLOv3) and the proposed integrated algorithm. We compare these trackers with the LCT tracker. Firstly, the introduction of the Gaussian window in AGW results in increased precision and success rates over the LCT tracker. Secondly, we implement a tracker which can locate optimal position and scale simultaneously. The idea of inspiration comes from SAMF trackers, which is used in the LCT tracker with multi-feature fusion. In addition, the re-detector YOLOv3 is used in the LCT tracker, it is essential for the tracker to re-detect the target when the algorithm fails to track targets successfully. We report the results on aerial infrared thermal image sequences by using the precision plot at a threshold of 20 pixels, for we can easily see the contributions of each component to the whole algorithm. The difference of five trackers is summarized in Table 2.

As shown in Figure 8, comparing to LCT, the AGW tracker outperforms the LCT by the distance precision of 5% due to the use of the adjustive Gaussian window. The OPS tracker can locate optimal position and scale simultaneously

TABLE 1. Comparisons with state-of-the-art trackers on aerial infrared image sequences. Our approach performs favorably against existing methods in distance precision (DP) at a threshold of 20 pixels, overlap success (OA) rate at an overlap threshold 0.5.

	OURS	KCF	SAMF	CSK	TLD	DSST	LCT	SRDCF	MOSSE	FDSST	BACF	C-COT	ECO-HC	ECO
DP (%)	100	19.9	23.9	39	49.3	56.4	75.6	76.9	85	88.2	98.2	98.6	98.8	98.9
OA (%)	65	9	12	18.1	14	2	37.1	29.1	55.8	57.8	22.2	40.4	57.4	63.5
FPS	22.4	101.8	41.66	927	12.4	114	40.1	2.21	1509	192.87	62.64	0.62	71.32	2.63

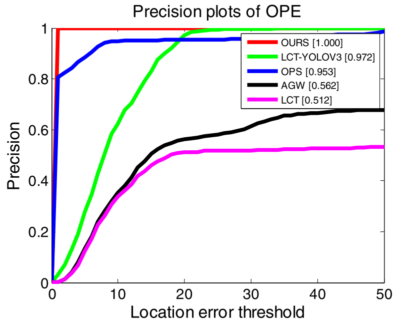


FIGURE 8. Precision plot for five implementations.

TABLE 2. The difference among five trackers.

Index	Adjustable Gauss function	Optimal Position and Scale	Re-detector	Baseline tracker
LCT	No	No	No	Yes
AGW	Yes	No	No	Yes
OPS	No	Yes	No	Yes
LCT-YOLOv3	No	No	Yes	Yes
Integrated algorithm	Yes	Yes	Yes	Yes

which can get accurate tracking results with distance precision of 95.3%. The LCT-YOLOv3 tracker significantly outperforms the OPS method due to the effectiveness of the target re-detection scheme in case of tracking failure. The proposed integrated approach (equipped with all the components) performs favorably against the other three alternative implementations, meanwhile these improved methods all outperform LCT tracker.

C. OVERALL PERFORMANCE

We evaluate the proposed algorithm with 13 state-of-the-art trackers, including TLD [17], MOSSE [2], CSK [18], KCF [3], DSST [5], SAMF [6], BACF [20], FDSST [4], SRDCF [7], LCT [9], C-COT [23], ECO-HC [24], and ECO [24]. For fair evaluations, we compare all the methods on our aerial infrared image sequences. These methods contain deep trackers, correlation filtering method and machine learning method. The tracking results are reported in one-pass evaluation (OPE) which uses the distance precision plot and overlap success plot in Figure 9.

In addition, we present the quantitative comparisons of distance precision plot at 20 pixels, overlap success plot at 0.5, and tracking speed in Table 1. Our approach is optimal in both distance precision and overlap success, and the FPS is moderate.

In order to test the algorithm further, one experiment was done on two public databases named VOT-TIR2016 and AMCOM. For VOT-TIR2016, there are 25 infrared image sequences. The paper mainly studies rigid body objects, and 7 of those can meet the requirement. The available sequences are boat1, boat2, car1, car2, quadrocopter, quadrocopter2, and ragged. In a similar way, for AMCOM, there are 5 infrared image sequences, and they are all available. We compare all the methods on the 12 available sequences. The tracking results are reported in one-pass evaluation (OPE) which uses the distance precision plot and

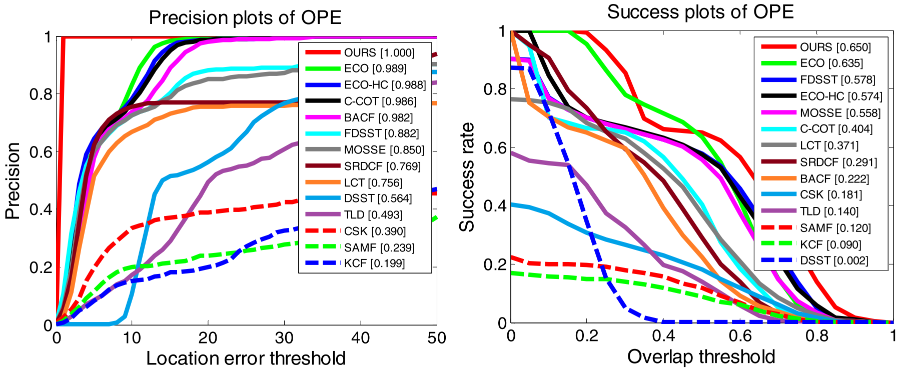


FIGURE 9. Precision plot and success plot over infrared image sequences about aerial target.

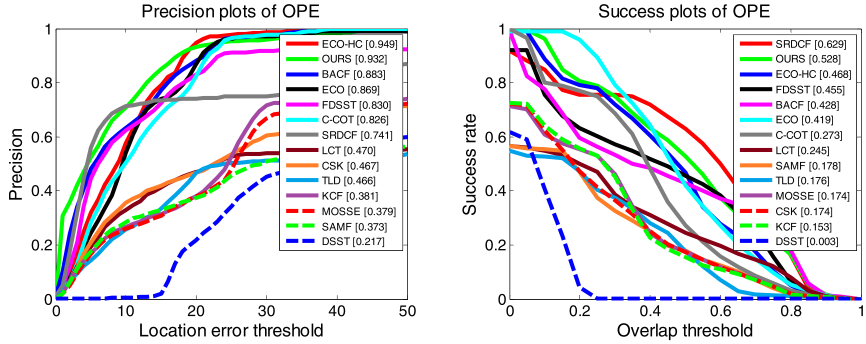


FIGURE 10. Precision plot and success plot over the choosing sequences of VOT-TIR2016.

TABLE 3. Comparisons with state-of-the-art trackers on aerial infrared image sequences. Our approach performs favorably against existing methods in distance precision (DP) at a threshold of 20 pixels, overlap success (OA) rate at an overlap threshold 0.5.

	OURS	KCF	SAMF	CSK	TLD	DSST	LCT	SRDCF	MOSSE	FDSST	BACF	C-COT	ECO-HC	ECO
DP (%)	93.2	38.1	37.3	46.7	46.6	21.7	47	74.1	37.9	83	88.3	82.6	94.9	86.9
OA (%)	81.8	15.3	17.8	17.4	17.6	0.3	24.5	62.9	17.4	45.5	42.8	27.3	46.8	41.9
FPS	20.7	111.4	44.22	1142	16.48	106	30.3	1.98	1897	168.39	55.37	0.68	59.83	2.19

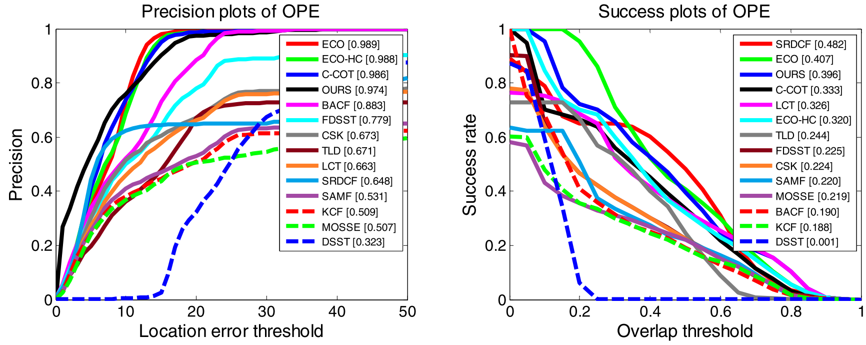


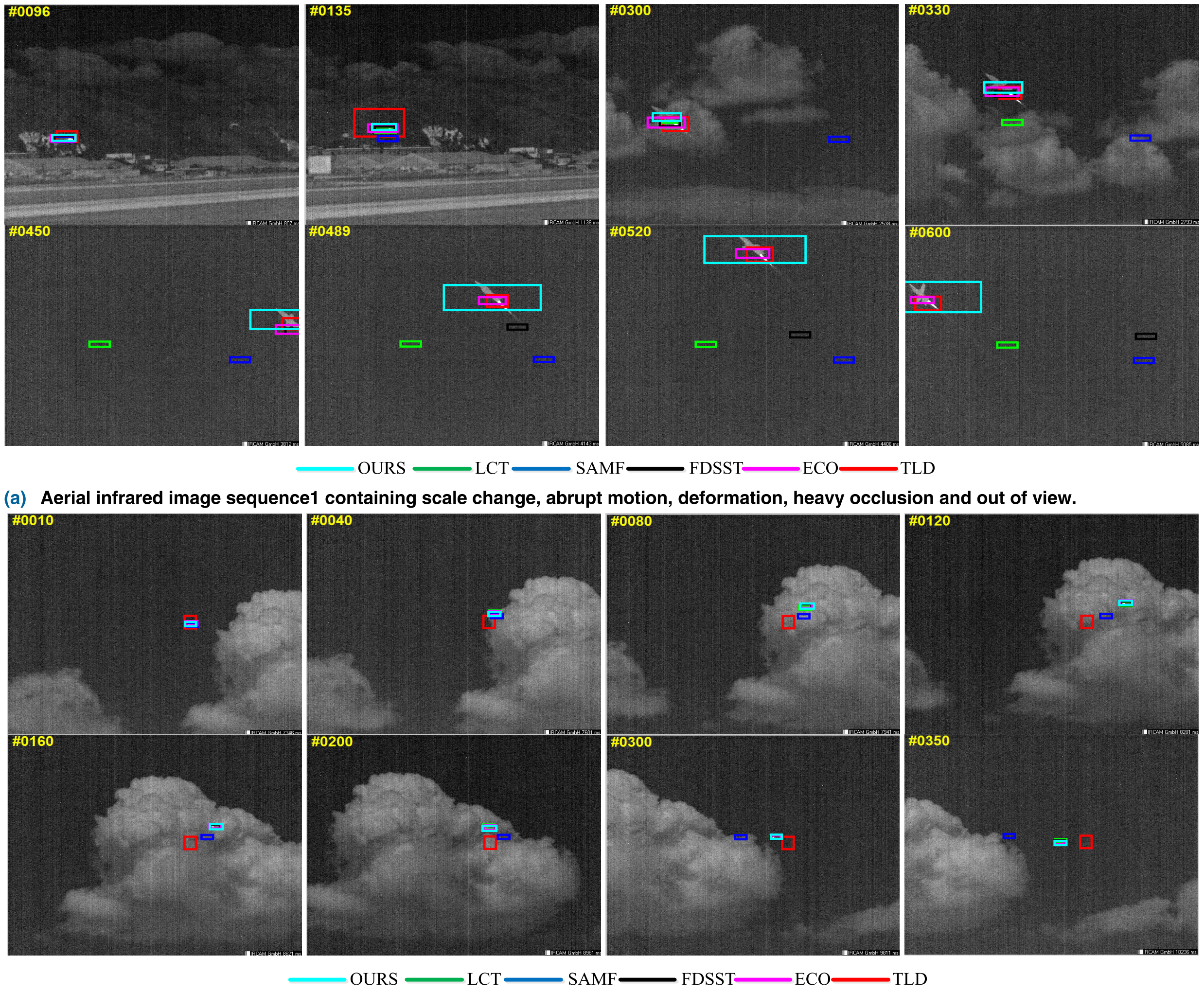
FIGURE 11. Precision plot and success plot over infrared image sequences about aerial target.

overlap success plot in Figure 10. We also present the quantitative comparisons of distance precision plot at 20 pixels, overlap success plot at 0.5, and tracking speed in Table 3. Our approach is inferior to the ECO-HC tracker in tracking precision. However, the ECO-HC tracker has a poorer tracking overlap success than that of our approach. Though SRDCF tracker has a higher overlap success than that of our approach, it has a poor real-time performance. Our approach is also good in both distance precision and overlap success, and the FPS is moderate.

Nowadays, the use of public databases is the most common tool to evaluate the performance of object tracking algorithms. However, algorithm testing is performed over a limited set of scenarios for aerial infrared target, which will affect the testing validity. Aiming at this problem, we use 1000 groups of infrared simulation image sequences to test algorithm, and each image sequence contains 600 to 1000 image frames. The infrared simulation image sequences are obtained by simulation platform. The simulation platform uses Microsoft Visual Studio as the development environment and uses common modeling software to create the scene model. In the end, the platform uses OSG to complete the final rendering. The tracking results are reported in one-pass evaluation (OPE) which uses the distance precision plot and overlap success plot in Figure 11. We also present the quantitative comparisons of distance precision plot at 20 pixels, overlap success plot at 0.5, and tracking speed in Table 4. Our approach is inferior to the ECO tracker, ECO-HC tracker, and C-COT tracker in tracking precision. However, both the C-COT tracker and ECO tracker are deep learning trackers which have poor real-time performance. The ECO-HC tracker has a poor tracking overlap success. Though SRDCF tracker has a higher overlap success than that of our approach, it has

TABLE 4. Comparisons with state-of-the-art trackers on aerial infrared image sequences. Our approach performs favorably against existing methods in distance precision (DP) at a threshold of 20 pixels, overlap success (OA) rate at an overlap threshold 0.5.

	OURS	KCF	SAMF	CSK	TLD	DSST	LCT	SRDCF	MOSSE	FDSST	BACF	C-COT	ECO-HC	ECO
DP (%)	97.4	50.9	53.1	67.3	67.1	32.3	66.3	64.8	50.7	77.9	88.3	98.6	98.8	98.9
OA (%)	39.6	18.8	22	22.4	24.4	0.1	32.6	48.2	21.9	22.5	19	33.3	32	40.7
FPS	23.4	106.8	46.87	1008	18.4	118	35.46	2.19	1659	175.6	58.4	0.65	63.57	2.45



(a) Aerial infrared image sequence1 containing scale change, abrupt motion, deformation, heavy occlusion and out of view.

(b) Aerial infrared image sequence2 containing abrupt motion, deformation, and heavy occlusion.

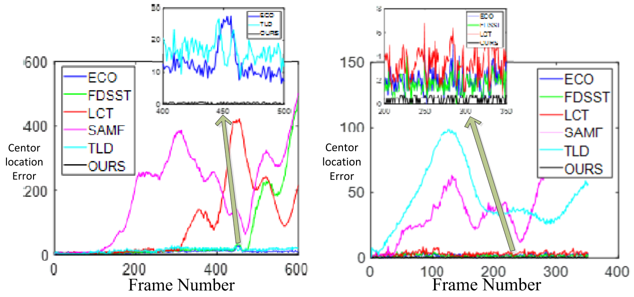
FIGURE 12. Tracking results of our algorithm, TLD, LCT, SAMF, FDSST and ECO methods on two representative aerial infrared image sequences.

a poor real-time performance. Our algorithm has a good performance in both distance precision and overlap success, and the FPS is moderate.

D. QUALITATIVE EVALUATION

In the experiment, we made a qualitative evaluation between our algorithm and other trackers. Figure 12 summarized qualitative comparisons of our method with five representative state-of-the-art trackers (e.g., TLD [17], LCT [9], SAMF [6], FDSST [4], and ECO [24]), which can solve different problems, such as long-term tracking, multi-feature fusion, scale change and deep feature learning. Compared with other trackers, our approach could track the aerial infrared target reliably. Likewise, in the event of multiple challenges (e.g., deformation, abrupt motion, heavy occlusion and out of view), our algorithm can track the center of target accurately. As shown in Figure 12(a), the SAMF tracker is based on a correlation filter learned from multi-feature fusion (i.e., HOG, CN and gray features). The SAMF tracker performs well in handling significant deformation and fast motion due to the robust representation of fusion features.

However, it drifts when target object is surrounded by complex background and does not re-detect targets in the case of tracking failure. LCT cannot adapt to scale change well. In the case of serious scale change, the tracking precision would decrease. And LCT tracker cannot get optimal position and scale simultaneously, which easily result in bad tracking result. The FDSST tracker can deal with scale change very well for its feature pyramid construction. However, when the target is out of view, it cannot evaluate the scale accurately and the target will be lost. As shown in Figure 12(b), although the TLD tracker is able to re-detect target objects in the case of tracking failure, it does not fully exploit the temporal motion cues and therefore does not follow targets undergoing significant deformation and fast motion. Moreover, when the target is small, the TLD method cannot perform well. Our approach and ECO tracker can achieve the tracking successfully in the event of multiple challenges. But our algorithm is superior to ECO tracker in both accuracy and speed.



(a). The center location error results of infrared image sequence1
(b). The center location error results of infrared image sequence2

FIGURE 13. Frame-by-frame comparison of center location errors (in pixels) of TLD, LCT, SAMF, FDSST, ECO and our methods on two representative aerial infrared image sequences.

In addition, we compare the center location error frame-by-frame on the two representative aerial infrared image sequences in Figure 13, which shows that our method performs well against existing trackers. Our method is able to track targets accurately and stably. As shown in Figure 13(a), the aerial infrared target begins to do leap maneuver, change scale seriously and be out of view in the 120th, 309th and 450th frames respectively, the SAMF, LCT and FDSST trackers fail to track target one after another. The deep tracker, ECO, benefits from the expressive power of CNN features, and achieves good tracking effect. The TLD tracker can manage to re-detect the target when trackers fail to locate the target. Our approach takes advantage of both deep features and re-detector, achieving the best tracking results. In a similar way, as shown in Figure 13(b), when the target is small, the TLD and SAMF methods cannot perform well. Although the ECO, LCT, FDSST and our trackers all get better performance, our algorithm works better among these trackers.

Overall, our proposed approach performs well in long-term tracking on these challenging aerial infrared target image sequences. The four Components, DTB-CF, UTA-CF, YOLOv3 and APCE criterion are related to each other to achieve real-time and efficient infrared target long-term tracking, all playing an important role in the algorithm.

V. CONCLUSION

This paper undertakes an in-depth study over twenty used object tracking algorithms, including deep trackers, correlation filtering tracker and machine learning tracker. The lack of training data and real-time requirement limit the application of deep tracker in engineering. We focus on our research on correlation filtering tracker and machine learning tracker. Firstly, aiming at the problem of universal tracking, many correlation filtering algorithms only can address one situation. For example, MOSSE, DSST, and SRDCF trackers can solve the problem of limited samples, scale estimation, and boundary effect respectively. However, an LCT tracker achieves near-perfect solution which can address various cases by decomposing the task of tracking into translation and scale estimation of objects. The translation and scale estimation of objects is a good idea to improve tracking performance for infrared object tracking. Secondly, aiming at the significant defects of infrared images, many useful features need to be integrated. For example, the SAMF tracker integrates the powerful features including HOG and color-naming to further boost the overall tracking performance. Therefore, achieving the best multi-feature fusion is useful for improving tracking performance for infrared object tracking. Last but not least, it is inevitable to re-detect targets in engineering. Only a few existing algorithms consider this problem, such as TLD, LCT, and CTAD trackers. Re-detecting targets in the case of tracking failure for infrared object tracking is important. On this basis, aiming at the problem of universal tracking and significant defects of infrared images, correlation filtering tracker and re-detection are combined to achieve long-term tracking for aerial infrared target.

Aiming at stability-plasticity dilemma, our algorithm effectively alleviates this dilemma by using two regression model based on correlation filters, namely UTA-CF and DTB-CF, to update the target appearance and discriminate the target from background. Aiming at the problem of tracking failure, the re-detection YOLOv3 is performed to obtain the correction position for successful tracking. In addition, we further improve the APCE criterion to determine whether the UTA-CF tracker needs to be updated or the target position should be initialized by re-detector YOLOv3. Our algorithm is suitable for long-term tracking because it is outstanding for solving the problem of long-term tracking. We prove by experiments that our approach is more valid against state-of-the-art methods in terms of efficiency, accuracy, and robustness.

In the future, we will improve the aerial infrared image database to better assess the performance of the object tracking algorithm. On this basis, we will use deep learning algorithms to extract the effective features and put forward a new correlation filtering framework based on deep features with better performance.

REFERENCES

[1] Q. Liu, X. Lu, Z. He, C. Zhang, and W. S. Chen, "Deep convolutional neural networks for thermal infrared object tracking," *Knowl.-Based Syst.*, vol. 134, pp. 189–198, Oct. 2017.

[2] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2016, pp. 850–865.

[3] X. Li, Q. Liu, N. Fan, Z. He, and H. Wang, "Hierarchical spatial-aware siamese network for thermal infrared object tracking," *Knowl.-Based Syst.*, vol. 166, pp. 71–81, Feb. 2019.

[4] 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 Trans. Image process.*, vol. 28, no. 4, pp. 1837–1850, Apr. 2019.

[5] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2015, pp. 3074–3082.

[6] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4293–4302.

[7] H. Nam, M. Baek, and B. Han, "Modeling and propagating CNNS in a tree structure for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016.

[8] B. Han, J. Sim, and H. Adam, "BranchOut: Regularization for Online ensemble tracking with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3356–3365.

[9] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2015, pp. 3119–3127.

[10] X. K. Lu, C. Ma, B. Ni, X. Yang, I. Reid, and M.-H. Yang, "Deep regression tracking with shrinkage loss," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 353–369.

[11] E. Gundogdu, H. Ozkan, H. S. Demir, H. Ergezer, A. Erdem, and S. K. Pakin, "Comparison of infrared and visible imagery for object tracking: Toward trackers with superior IR performance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[12] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.

[13] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[14] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2014, pp. 1–11.

[15] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2016.

[16] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 254–265.

[17] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4310–4318.

[18] N. Wang, J. Shi, D. Yeung, and J. Jia, "Understanding and diagnosing visual tracking systems," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3101–3109.

[19] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5388–5396.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[21] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 7263–7271.

[22] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 11–18.

[23] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[24] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.

[25] M. Kristan, J. Matas, A. Leonardis, T. Vojír, R. Pflugfelder, G. Fernández, G. Nebehay, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2137–2155, Nov. 2016.

[26] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. S. Torr, "Struck: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, Oct. 2016.

[27] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[28] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 702–715.

[29] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1090–1097.

[30] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2017, pp. 1144–1152.

[31] A. S. Montero, J. Lang, and R. Laganière, "Scalable kernel correlation filter with sparse feature integration," in *Proc. IEEE Conf. Comput. Vis.*, Dec. 2015, pp. 587–594.

[32] C. Fu, F. Lin, Y. Li, and G. Chen, "Correlation filter-based visual tracking for UAV with Online multi-feature learning," *Remote Sens.*, vol. 11, no. 5, p. 549, Jul. 2019.

[33] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 472–488.

[34] M. Wang, Y. Liu, and Z. Huang, "Large margin object tracking with circulant feature maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4021–4029.

[35] T. Li, S. Zhao, Q. Meng, Y. F. Chen, and J. B. Shen, "A stable long-term object tracking method with re-detection strategy," *Pattern Recognit. Lett.*, Jun. 2018, pp. 1–9.

[36] Y. G. Hu, M. Q. Xiao, K. Zhang, and X. T. Wang, "Aerial infrared target tracking in complex background based on combined tracking and detecting," *Math. Probl. Eng.*, vol. 2019, Mar. 2019, Art. no. 2419579.

[37] X. Sheng, Y. Liu, H. Liang, F. Li, and Y. Man, "Robust visual tracking via an improved background aware correlation filter," *IEEE Access*, vol. 7, pp. 24877–24888, 2019.

[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[39] R. Girshick, "Fast R-CNN," in *Proc. IEEE Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

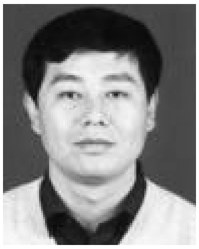
[41] C. Zheng and Z. Wei, "Real-time tracking based on keypoints and discriminative correlation filters," *IEEE Access*, vol. 7, pp. 32745–32753, 2019.

XIAOTIAN WANG received the B.S. degree from the North China Institute of Aerospace Engineering, in 2013, and the M.S. degree from the School of Electronics and Information, Northwestern Polytechnical University, in 2016, where he is pursuing the Ph.D. degree with the School of Astronautics. He has presented papers in national and international journals and conferences. His research interests include object detection, object tracking, and image quality evaluation.

KAI ZHANG received the Ph.D. degree from the School of Astronautics, Northwestern Polytechnical University, in 2009, where he is currently an Associate Professor. His research interests include guidance, navigation, and control.

SHAOYI LI received the bachelor's degree from Southwest Jiaotong University, in 2008, and the master's and Ph.D. degrees from Northwestern Polytechnical University, in 2011 and 2015, respectively, where he is currently an Assistant Researcher. His current research interests include artificial intelligence and image processing.

YANGGUANG HU received the M.E. degree from Air Force Engineering University, Xi'an, China, in 2016, where he is currently pursuing the Ph.D. degree with the School of Aeronautics Engineering. His research interests include IR object tracking, image processing, and computer version.



JIE YAN was born in 1960. He received the Ph.D. degree from the School of Astronautics, Northwestern Polytechnical University, in 1988, where he is currently a Professor and a Ph.D. Candidate Supervisor. His research interests include flight control, guidance, system simulation, and aircraft design.