A Detection Aided Multi-Filter Target Tracking Algorithm

YUE MIN, ZHENZHONG WEI, AND KE TAN

Key Laboratory of Precision Opto-mechatronics Technology, Ministry of Education, Beihang University, Beijing 100191, China

Corresponding author: Zhenzhong Wei (zhenzhongwei@buaa.edu.cn)

This work was supported in part by the Leading Talents in Science and Technology Innovation under the Ten Thousands Plan, in part by the National Science Fund for Distinguished Young Scholars of China under Grant 51625501, and in part by the Co-Construction Project of Scientific Research and Postgraduate Cultivation—Achievements Transformation and Industrialization Program—On-line dynamic detection system of train pantograph-catenary operating conditions provided by the Beijing Municipal Education Commission.

ABSTRACT When the target is partially occluded or undergoes drastic appearance change, drift is prone to occur in traditional correlation filter-based trackers and ultimately leads to tracking failure. This paper proposes a robust tracking algorithm, which can be applied to practical engineering. The proposed algorithm divides all training samples into different training sets based on their similarities. An independent filter is trained for every training set. In each frame of the tracking process, the similarity between the candidate region and any training set’s training samples is calculated to select the most matched filter for target location. As a newly introduced training sample, a feature map within the current frame’s estimated target bounding box is assigned to the most similar training set to update the most matched filter. The updating of each filter is relatively independent. Each filter corresponds to one kind of target’s typical appearance, and thus, the proposed tracking algorithm can memorize a variety of typical target appearances that appeared in the past and adapt to target’s discontinuous appearance change. The detector improves tracking accuracy and assists occlusion judgment. After occlusion, the target may reappear with a different posture. As long as similar postures have been captured in the previous tracking process, the target can be retrieved accurately while not being confused with other objects of the target’s category. We evaluated our method on massive videos shot in actual engineering sites, and the result demonstrates that the proposed tracking algorithm can handle occlusion and target’s posture changes very well and significantly outperforms other tracking algorithms.

INDEX TERMS Tracking, correlation filter, anti-occlusion, multi-filter, detector.

I. INTRODUCTION

Object tracking is one of the research hotspots in computer vision. It predicts the trajectory of the target in a continuous image sequence given initial target position in the first frame. Object tracking plays an important role in many real-time visual applications, such as intelligent surveillance system, automatic driving, intelligent traffic control and human-machine interaction.

In recent years, correlation filter-based target tracking algorithms have made great progress, its speed and accuracy have been effectively improved [1]–[3]. The correlation filter algorithm adopts a discriminant appearance model. Dispense with identifying specific characteristics of the target, it trains a filter which can distinguish the target from the background. The target’s probability of existence in any location of the candidate region can be calculated by convoluting the candidate region’s feature map and the trained filter.

MOSSE (Minimum Output Sum of Squared Error) tracking method [4] convolutes the original gray image and the filter, and greatly improves speed and accuracy of correlation filter-based tracking algorithm. The kernel function can convert the inner product operation of high dimensional space into low dimensional input space operation. João F. Henriques uses the kernel function to train the high dimensional filter [5]. Different kernel functions and multi-channel feature maps are then introduced in the kernelized correlation filters tracking algorithm(KCF) [6].

The KCF algorithm is improved in many aspects. Danelljan M designs the desired color feature input to
improve the speed [7], adapts target’s scale change [8] and reduces the boundary effect [9]–[11]. Danelljan M discards the kernel function and turns different scale features maps into continuous feature maps [12]. They are fused to train a universal continuous filter which is capable of grasping target’s different scale features maps, but the training of a continuous filter must use the conjugate gradient method for cyclic iterative optimization and is so time-consuming.

Although the ECO (Efficient Convolution Operators) algorithm minimizes the redundancy of filters and training samples [13], the filter calculation is still complex and has a low speed.

Regression models of different updating rates are used to deal with target’s fast deformation and conservative scale estimation separately [14]. Reliable weights of the training samples are learned during tracking process [15] and the correct training samples are selected automatically [16]. Mueller M incorporates global context within trackers [17]. Lukezic A adjusts the filter to track the object’s suitable part [18]. Zhang T complements KCF with a particle filter [19].

These algorithms tend to improve their performance by complex algorithms by increasing number of parameters, which may lead to overfitting and low speed thus cannot be applied to practical engineering. Besides, typically adopted target’s continuous appearance model is not robust to factors such as drastic target appearance changes or occlusion. It’s likely to be overfitted by newly introduced training samples and cannot memorize target’s typical appearances shown in the long past. Short time occlusion may completely pollute the filter. The contaminated filter cannot identify the target even when the target reappears in the candidate region.

There are many studies working on how to handle target’s deformation and occlusion. Occlusion can be captured by target’s trajectory information [20], [21] or background patches’ interaction with the target [22], [23]. The occluded tracking target corresponds to the combination of target templates and sparse representation where non-zero element appear in the small square of the image [24]. Highly deformed or partially occluded target can also be denoted by combination of local tracked patches and the global layer [25]. Target’s most suitable parts are selected to track the target in every frame [26], [27]. Motions and color cues of the target can be used to recover target labels after occlusion [28]. However, the speed and accuracy of these algorithms need to be improved.

Tracking and detection often work together to complement each other. Tracking results can be used to train the detector, and the detector detects the target when tracking fails [29], [30]. With the popularity of deep neural networks, many neural network-based tracking algorithms emerge. The deep attentional network is used to select a subset of filters according to the dynamic properties of the tracking target [31]. It can also be used to select the sequential actions which iteratively move the initial bounding box to the target bounding box in each frame [32]. The Correlation Filter learner as a differentiable layer in a deep neural network enables learning deep features that are tightly coupled to the Correlation Filter [33]. Recurrent neural network (RNN) is utilized to model object structure and improve its robustness in presence of similar distractors [34].

The neural network model trained by online obtained target bounding boxes often faces the problem of insufficient training samples [35]. Adversarial learning is used to capture rich appearance variations to overcome the insufficiency of training samples [36]. However, if the type of tracking target is known, the prior information can be used, that is, the trained detector can be used to assist the tracking algorithm [37].

There are many high quality offline training images databases such as Pascal VOC (pattern analysis, statistical modeling and computational learning, visual object classes) [38], ILSVRC2012(ImageNet Large Scale visual Recognition Challenge 2012) [39] and COCO (Common Objects in Context) [40]. These databases can be used to train very high precision detectors [41], [42], such as SSD (the single shot multi-box detector).

In this paper, a stable multi-filter tracking algorithm is proposed. It is based on detection model SSD and the KCF tracking algorithm which has an obvious speed advantage. Compared with KCF, the distinction and advantages of our proposed method are:

1. When the target’s appearance changes dramatically, KCF and other single filter-based trackers can only store the relatively new target appearance. When the target suddenly changes into a typical appearance shown in long time ago, the filter cannot identify the target. However, the proposed algorithm divides all training images into several sets according to their similarity and trains a filter for each set. Every filter corresponds to one typical target appearance shown in history. The newly introduced training sample will only update its corresponding filter and won’t interfere with other filters. So, it can memory a variety of typical target appearances in history and adapt to discontinuous appearance model. Even if the wrong training sample is introduced, it does not pollute target’s appearance information stored in other filters and can be redeemed.

2. The traditional KCF tracking algorithm cannot recognize occlusion. Once occlusion occurs, plenty of wrong training samples are introduced to pollute the existing filter. The proposed algorithm combines the detector with various typical target appearances stored in filters to judge occlusion. It can recognize occlusion accurately to avoid the introduction of wrong training samples. After verifying occlusion, the target may reappear in the camera’s field of view with a different posture. As long as similar postures have been captured in former tracking process, the target can be retrieved accurately by the proposed algorithm while not being confused with other objects of target’s category.

3. In the KCF tracking process, similar background or poor-quality image caused by air flow may lead to tracking failure even when the calculated target’s probability of existence is high all the time. In the proposed algorithm, the detector is introduced to correct target’s
bounding box. Compared with KCF or the more recent tracking algorithm ECO, the accuracy of the proposed algorithm has been significantly improved. This algorithm also remains the speed advantage of the KCF and can still achieve 25 frames per second.

The rest of this paper is organized as follows. Section II briefly reviews the KCF tracking algorithm and analyzes its advantages and disadvantages. Section III describes the improvement of the proposed algorithm. Experimental results are reported in Section IV, demonstrating the improved algorithm’s accuracy and speed. Section V concludes this work.

II. BASELINE APPROACH: KERNELIZED CORRELATION FILTERS

The original feature map within the estimated target bounding box in current frame is the newly introduced training sample \(x\). \(x\) is of \(C\) dimensions with \(m\) pixels length and \(n\) pixels width.

To generate more high dimensional training samples and full exploit their information. \(x\) is cyclically shifted to obtain \(m^2n\) same sized images \(x_i, i = 1, 2, \ldots, m^2n\). By nonlinear transformation, they are then transformed into large sized training samples \(X_i, i = 1, 2, \ldots, m^2n\).

Through the discrete Fourier inverse transformation, new matrixes are reported in Section IV, demonstrating the improved algorithm’s advantages and disadvantages. Section III describes the detection aiding multi-filter tracking algorithm ECO, the accuracy of the proposed algorithm to determine the plane’s specific position from the control computer. The control computer uses the proposed tracking algorithm to take pictures of the airplane and the data is transmitted to the control computer. Every station is composed of a visible-light zoom and a two-axis servo system. The visible-light zoom imaging system and a two-axis servo system. The visible-light imaging system consists of an industrial camera and a visible-light zoom lens. They are installed on outside mobile platform to get a wider view and protected by a protective cover as shown in Fig. 1.

At work, the protective cover opens and the camera begins to take pictures of the airplane and the data is transmitted to the control computer. The control computer uses the proposed tracking algorithm to determine the plane’s specific position.

\[ k^{\alpha*} = \exp\left(-\frac{1}{\sigma^2}(\|x\|^2 + \|x\|^2 - 2 \sum_c F^{-1}(\hat{x}_c \circ \hat{x}_c))\right) \]  

\[ \lambda \] is the regular term coefficient. The objective function to obtain \(\alpha\) is:

\[ \min_{\alpha} \left(\|(K\alpha - y)^2 + \lambda \|\alpha\|^2\right) \]

The practical application device of the proposed tracking algorithm to determine the plane’s specific position from the control computer. The control computer uses the proposed tracking algorithm to take pictures of the airplane and the data is transmitted to the control computer. Every station is composed of a visible-light zoom imaging system and a two-axis servo system. The visible-light imaging system consists of an industrial camera and a visible-light zoom lens. They are installed on outside mobile platform to get a wider view and protected by a protective cover as shown in Fig. 1.

At work, the protective cover opens and the camera begins to take pictures of the airplane and the data is transmitted to the control computer. The control computer uses the proposed tracking algorithm to determine the plane’s specific position.
in current frame, adjusts the lens focus according to the proportion of the target in the whole picture, and controls the two-axis servo system by the deviation between the target center and the image center. Through above methods, the two-axis servo system is controlled to keep the target in the center of camera’s field of view.

Because of the shelter of nearby buildings, there is a visual blind spot in a single station. Once the target is lost, one station doesn’t know how to control its two-axis servo system to retrieve the lost target. So, if only one station is adopted, it can’t handle occlusion of the target airplane shown in Fig. 2.

While in the dual-station target cooperative tracking system, another station whose target is not lost can roughly estimate the three-dimensional space position of the airplane based on the size of the target in the image, the actual size of the target and the current focal length of the zoom lens. Combining the location relationship between the two stations, it can guide the movement of the two-axis servo system of the lost target station. The overall working diagram of the two stations is shown in Fig. 3.

After target’s longtime loss, the flying attitude of the airplane may change when it reappears in the lost target station’s field of view. The change of airplane’s flying attitude leads to the change of the target’s appearance. When the target reappears, it may not be the appearance before occlusion, instead, it may be target’s appearance that has appeared a long time ago. In order to locate the target accurately after it reappears, the tracking algorithm applied in the dual-station system must memorize various target’s appearances that appeared in the past. The engineering needs requires the tracker store multiple typical target appearances while still remains a high speed.

If the project site has multiple tracking targets, multiple stations can be set up, and a network of multiple cameras can be used for target detection and tracking. Each site has a fixed location, the location of the nearby buildings is also fixed, therefore the observation blind area of each station can be measured in advance. When the target is about to enter one station’s blind area, another station is adjusted to capture and track the target, while original tracking station will be reset to view the area where the next target appears and track it. When there are a large number of tracking targets, alternate of the tracking station can also be in two station’s view overlap range, so as to achieve multi-target tracking.

B. FUNCTION AND UPDATING MODE OF MULTI-FILTER ALGORITHM

The traditional single filter-based trackers cannot store multiple target’s typical appearances, it can only store the relatively new target appearance thus cannot memorize target appearances that shown in the long past. As shown in Fig. 4, it can’t retrieve the target if target reappears in a posture appeared long time ago. Besides, once one wrong training sample is introduced, it will contaminate all the reliable information stored in the filter.

Therefore, an ideal tracking algorithm is required to record a variety of target’s typical appearances which are photographed from different angles. It’s also supposed to deal with occlusion and retrieve the target accurately when target reappears in the field of vision.

Based on KCF, the proposed algorithm has an improved function and updating mode. It divides all training samples into multiple training sets according to their similarity, each
training set corresponds to one typical target appearance shown in the past. Training samples in one training set are used to train one independent filter. The similarity between candidate region and any training set’s training samples is calculated to select the most matched filter for target location in current frame. Then, the feature map within the estimated target bounding box is used to update this filter.

When determining the target location, candidate regions of different scales are extracted according to the target position in previous frame. The most matched filter is selected for them respectively and target’s probability of existence in each location of different-scale candidate region is calculated. Different filters correspond to different training sets. If one training set has the most similar training samples to the candidate region, the filter corresponding to that set is the most matched filter. Function and updating mode of the proposed algorithm can be seen in Fig. 5.

One filter is trained by all training samples in one training set. The filter’s information is stored by a feature map, a weight coefficient vector and a filter weight. The filter is a linear combination of all the images obtained from cyclic shift and nonlinear transformation of the feature map. The weight coefficient vector contains weight information of the linear combination, and the filter weight reflects the confidence of that filter to a certain extent. The sum of every filters’ filter weight is 1.

Assuming the feature map of filter \( f_i \) is \( I_{f_i} \). One candidate region’s feature map is \( I \). To choose the most matched filter of \( I \), the similarities between \( I \) and every filter’s corresponding feature map are calculated. The following similarity evaluation principle is adopted:

\[
S = ||I - I_{f_i}||^2 \quad i = 1, 2 \ldots N
\]  

(5)

The best matched filter is the filter whose feature map can achieve the minimum value of \( S \). Candidate region of different scales are matched with filters and their maximum response score are calculated respectively. If the maximum response score is smaller than the threshold value, it might be occlusion and will be judged further. If the maximum response score is greater than the threshold, the target is located, and the corresponding scale and filter are selected. In this way, the proposed algorithm can adapt to target’s discontinuous appearance model and switch freely among different typical target appearances. Even if some wrong training samples are introduced during certain period, they will be attributed to some certain training sets corresponding to certain filters and will not pollute the reliable information stored in other filters, which will facilitate the subsequent correction of the target position and retrieval of the target.

When the filter is updated, feature map within the estimated target bounding box in current frame is used to update the most matched filter. The filter’s weight is increased and the others are decreased.

The filter weight determines the way of filter updating when the target is rediscovered after occlusion. If one filter’s weight is lower than the threshold, the filter is updated by the retrieved target bounding box’s feature map \( x \), and if there doesn’t exist filter whose weight is below the threshold, the most matched filter is to be updated. After the occluded target is retrieved, the updating method of the training feature map and elements of the weight coefficient vector are:

\[
\begin{align*}
I_{f_\alpha} &= \theta * I_{f_{\alpha}} + (1 - \theta) * x, \quad \alpha f_{\alpha} \\
I_{f_\beta} &= \theta * I_{f_{\beta}} + (1 - \theta) * x, \quad \alpha f_{\beta} < \tau \\
I_{f_\beta} &= \theta * I_{f_{\beta}} + (1 - \theta) * x, \quad \alpha f_{\beta} \geq \tau
\end{align*}
\]  

(6)
The updating method of filters’ filter weights are:

\[
\begin{align*}
\omega_i &= 0.5/(1 - \omega_{fb})^\alpha, \omega_{fb} = 0.5 \quad \text{if } \tau \\
\omega_i &= 0.9^\alpha, \omega_{fb} = 1 - \sum_{i \neq fb} \omega_i \quad \text{else}
\end{align*}
\]

(7)

\(\alpha\) is the weight coefficient vector trained by the current training feature map \(x\). \(\omega_{fb}, I_{fb}, \alpha_{fb}\) are filter weight, feature map and weight coefficient vector of the filter \(f_{fb}\) whose filter weight is minimum. \(\omega_{fb}, I_{fb}, \alpha_{fb}\) are filter weight, feature map and weight coefficient vector of the most matched filter \(f_{fb}\). \(\omega_i\) represent filter weight of other filters. \(\tau\) is the interpolation factor. \(\tau\) is the threshold. These similar wrong training samples will be attributed to update certain filters, thus will not pollute or interfere with other filters’ information. Weights of these certain filters will gradually decline when wrong training samples cease to be introduced. When their weights are less than the threshold. These polluted filters will be covered by the newly trained filter after the occluded target is retrieved.

During tracking process, drift may happen while the calculated target’s probability of existence is very high all the time. Reason for that situation is that the filter is slowly being polluted. To prevent the target’s bounding box from containing too much similar background, the proposed algorithm sends the candidate region into the detector every M frame, and outputs the maximum detection score of the target’s class. If the score of the target’s class is less than the threshold, it enters the occlusion state.

C. OCCLUSION JUDGMENT AND LOST TARGET RETRIEVAL

Because the historical target appearance information is overlaid by relatively new learned target appearance, KCF can actually recognize limited kinds of target appearances. Therefore, without prior knowledge, it is hard to determine whether it’s dramatic target appearance changes or occlusion when the calculated target’s probability of existence score is low. KCF will take it all as drastic target appearance change and introduce plenty of wrong training samples during occlusion. All target appearance information stored in the filter will be polluted quickly.

The proposed algorithm combines the SSD detector and various target appearances stored in filters to judge occlusion even occlusion of other objects of target’s category. It can avoid the introduction of wrong training samples and reduce the detection range in the process of target retrieving (only needs to detect the original target candidate region and its surrounding areas during short time occlusion). Interference of other objects of target’s category can be excluded and target of any historical typical appearance can be retrieved.

As said before, when the maximum response score is smaller than the threshold value, to further judge whether it’s occlusion or not, SSD is called to detect all objects of the target’s category in the detection area. The detection area is expanded as N times of the target bounding box in the previous frame. If all detection scores \(d\) are lower than pre-defined threshold \(v_1\), occlusion is verified. Otherwise, the detected target bounding box is obtained. In circumstances where there are many similar objects’ interferences, the detected target bounding box should be re-matched to the filter to calculate the maximum filter response score to ensure it’s not other objects of target’s category. In circumstances without the interference of other objects of target’s category, the process of rematch can be omitted. Flow chart of the multi filter tracking algorithm is shown in Fig. 6. It can be applied to engineering field where occlusion and similar objects’ interference are rare. Flow chart of the robust enhanced multi filter tracking algorithm is shown in Fig. 7. It can be applied to situations where there are frequent occlusion and similar objects’ interference.

After the detected target bounding box is obtained by SSD, candidate regions of different scales are extracted according to the detection boxes whose detection scores \(d\) exceed pre-defined threshold \(v_1\). Candidate regions’ matched filters and maximum filter response scores are calculated respectively. Their maximum filter response scores are \(p_1, p_2, p_3 \ldots\), the maximum of which is \(p_{max}\). If \(p_{max}\) is less than the threshold \(v_2\) occlusion is affirmed. If \(p_{max}\) is greater than the
threshold \( v/2 \), it is identified as drastic change of the target’s appearance. The occlusion mark is \( \chi \), above can be expressed as:

\[
\begin{align*}
\chi &= 0; \quad d >= v1 & \land & p_{\text{max}} > = v2 \\
\chi &= 1; \quad \text{else}
\end{align*}
\]

(8)

When occlusion is confirmed, filter updating is stopped immediately to prevent the introduction of wrong training samples. The target retrieval process is similar to the occlusion judgment process. Candidate region’s position remains the same. The candidate region and its vicinity are detected to find the target. Regions of different scales are extracted based on the detection box whose detection score is greater than the threshold \( v1 \). Different scale regions’ most matched filters are calculated and the maximum filter response is calculated respectively. When both the detection score \( d \) and its corresponding maximum filter response score \( p_{\text{max}} \) are greater than the threshold value, the target is retrieved. If several detection boxes satisfy these requests, \( \sqrt{p_{\text{max}}^2 + d^2} \) is used as a measurement to select the most matched detection box. When occlusion continues for 20 frames and the target object still cannot be retrieved, it enters a state of emergency. In the emergency state, the detection range is enlarged, the \( N \) value is increased and target detection continues.

IV. EXPERIMENTS

A. EXPERIMENTAL ENVIRONMENT AND PARAMETERS

The platform for KCF and multi-filter tracking algorithm is ubuntu16.04. All experiments are completed in the computer of Intel Core i7 GPU, 2.81Ghz main frequency, 8 GB memory, NVIDIA GeForce GTX 1050 Ti video card. Other parameters in the experiment are default of the original code.

B. DATA AND EVALUATION INDICATORS

The important evaluation indicators for tracking are Frames Per Second (FPS), center location error (CLE) and average overlap rate (OR). FPS is the number of frames processed per second, which is related to the size and the resolution of the frame. CLE means the average deviation value between the center coordinate of the bounding box and the center coordinate of the ground truth box. Average OR of the bounding box (a) and the ground truth box (b) is defined as: \( \text{OR} = |a \cap b|/|a\cup b| \).

C. EXPERIMENTAL RESULTS AND ANALYSIS

Different tracking algorithms are tested in video sequences photographed in engineering sites. The ECO with depth feature, KCF of hog feature and the proposed multi-filter algorithm are selected to track 10 airplane videos segments where airplane undergoes drastic appearance changes without occlusion. Parameters of the algorithm are as follows: the interpolation factor \( \theta = 0.005 \), threshold \( v1 = 0.3 \), threshold \( v2 = 0.46 \), threshold \( r = 0.03 \), expansion factor \( N = 3 \), interval frames of Checking \( M = 20 \). Screenshots of the tracking process are shown in Fig. 8. The first, second and third column of the figure shows the estimated target bounding boxes of ECO, KCF and the multi-filter algorithm in the tracking process respectively. Results show clearly that KCF is not stable. When the airplane has not taken off, quality of the picture is not so good because of turbulence in the field. Target bounding box of the KCF has obviously offset. ECO was a bit more stable, but after the airplane took off, the airplane’s appearance undergoes significant changes and ECO was unable to adapt to such intense deformation so that the target bounding box gradually deviated from the target. However, the proposed method can track the airplane very well without drift of the bounding box.

Quantitative evaluation results of different tracking algorithms in Experiment 1 are shown in Table 1.

Success plot of our proposed algorithm and some other compared algorithms are shown in Fig. 9. The precision plot of our proposed algorithm and some other compared algorithms are shown in Fig. 10. It can be seen that our algorithm performs better than other tracking algorithms. In 1920\(^*\)1080 field image sequences, the distance deviation of the central position of more than 70% of the images is less than 30pixel, and the distance precision (DP) is nearly doubled.

The 10 videos contain images of the same size with resolution of 1920\(^*\)1080 pixels. However, the target proportions in different images are different. The proportion determines the computation complexity and tracking speed to some extent. KCF trains a filter for one target bounding box at one time then uses that filter to update the existing filter. The filter is a weighted combination of all images obtained from cyclic shift and nonlinear transformation of the original training
feature map. The weight coefficient can be calculated directly, so the speed of KCF is the fastest on almost all videos. Only frame rates above 25 frames per second can fulfill the real-time request. KCF can satisfy or even exceed the real-time request, but its accuracy is not guaranteed. Experimental results of KCF show that the average OR is always around 0.7, and the drift is obvious when the target appearance changes drastically. The average CLE even achieves 100 pixels on the launch3 video.

ECO uses all previous target bounding boxes’ feature maps to train a continuous filter, and it is necessary for it to use the time-consuming conjugate gradient method during filter training, so it is the slowest tracking algorithm in all videos and cannot meet the real-time requirement. The average OR is around 0.75. The drifts is obvious when the target appearance changes drastically, and the average CLE even achieves 130 pixels in land3 video.

Speed of the multi-filter algorithm varies in different videos, because it is a combination of SSD and KCF, and the speed of SSD and KCF are not consistent. If there are more changes in appearance of the airplane in the video, the detector will be used more frequently, then the running time increases. The reason that the proposed algorithm sometimes approximates or even exceeds KCF in speed is that the target bounding box in KCF tends to contain extra background parts that do not belong to the target which will increase amounts of calculation. Speed of the proposed algorithm is improved because the proposed algorithm is always more

**TABLE 1. Comparison of ECO, KCF and the proposed algorithm evaluation indicators.**

<table>
<thead>
<tr>
<th>Video Name</th>
<th>Length (frame)</th>
<th>FPS (ECO)</th>
<th>FPS (KCF)</th>
<th>FPS (proposed)</th>
<th>CLE (ECO)</th>
<th>CLE (KCF)</th>
<th>CLE (proposed)</th>
<th>Average OR (ECO)</th>
<th>Average OR (KCF)</th>
<th>Average OR (proposed)</th>
<th>Average value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land1</td>
<td>4800</td>
<td>&lt;10</td>
<td>55.0</td>
<td>48.0</td>
<td>90.5</td>
<td>55.0</td>
<td>26.4</td>
<td>0.76</td>
<td>0.70</td>
<td>0.92</td>
<td>4231</td>
</tr>
<tr>
<td>Land2</td>
<td>7200</td>
<td>&lt;10</td>
<td>70.0</td>
<td>34.0</td>
<td>74.9</td>
<td>37.0</td>
<td>23.0</td>
<td>0.72</td>
<td>0.77</td>
<td>0.92</td>
<td>61.1</td>
</tr>
<tr>
<td>Land3</td>
<td>1997</td>
<td>&lt;10</td>
<td>57.0</td>
<td>69.0</td>
<td>131.6</td>
<td>94.0</td>
<td>26.0</td>
<td>0.65</td>
<td>0.73</td>
<td>0.88</td>
<td>49.4</td>
</tr>
<tr>
<td>Land4</td>
<td>2760</td>
<td>&lt;10</td>
<td>47.0</td>
<td>74.6</td>
<td>94.5</td>
<td>48.5</td>
<td>34.6</td>
<td>0.80</td>
<td>0.73</td>
<td>0.87</td>
<td>73.1</td>
</tr>
<tr>
<td>Land5</td>
<td>5101</td>
<td>&lt;10</td>
<td>64.0</td>
<td>62.0</td>
<td>81.0</td>
<td>45.0</td>
<td>30.5</td>
<td>0.75</td>
<td>0.78</td>
<td>0.83</td>
<td>59.3</td>
</tr>
<tr>
<td>Launch1</td>
<td>4800</td>
<td>&lt;10</td>
<td>61.0</td>
<td>33.6</td>
<td>51.7</td>
<td>60.0</td>
<td>15.0</td>
<td>0.76</td>
<td>0.73</td>
<td>0.94</td>
<td>46.3</td>
</tr>
<tr>
<td>Launch2</td>
<td>3600</td>
<td>&lt;10</td>
<td>62.0</td>
<td>50.0</td>
<td>43.0</td>
<td>74.0</td>
<td>25.0</td>
<td>0.82</td>
<td>0.78</td>
<td>0.90</td>
<td>46.3</td>
</tr>
<tr>
<td>Launch3</td>
<td>3449</td>
<td>&lt;10</td>
<td>34.0</td>
<td>34.0</td>
<td>81.8</td>
<td>105.0</td>
<td>17.6</td>
<td>0.63</td>
<td>0.71</td>
<td>0.93</td>
<td>46.3</td>
</tr>
<tr>
<td>Launch4</td>
<td>3960</td>
<td>&lt;10</td>
<td>73.0</td>
<td>43.5</td>
<td>36.0</td>
<td>105.0</td>
<td>16.7</td>
<td>0.82</td>
<td>0.73</td>
<td>0.93</td>
<td>46.3</td>
</tr>
<tr>
<td>Launch5</td>
<td>4640</td>
<td>&lt;10</td>
<td>87.5</td>
<td>45.0</td>
<td>46.3</td>
<td>30.5</td>
<td>31.8</td>
<td>0.75</td>
<td>0.68</td>
<td>0.81</td>
<td>46.3</td>
</tr>
</tbody>
</table>

**FIGURE 9. Success plot of our proposed algorithm and some other contrast algorithms.**

**FIGURE 10. Precision plot of our proposed algorithm and some other contrast algorithms.**
accurate in bounding the target and the calculation amounts are reduced. Experimental results show that the FPS of the proposed algorithm are at least 25, enough to meet the real-time requirements, and the accuracy is improved prominently. The average OR is more than 0.8, even reaches 0.9, and the average CLE is within 35 pixels.

Experiment 2 tests the multi-filter tracking algorithm when there is interference from other objects of target’s category. All the 3624 images in the test video are 640*480 pixels, 24 bits depth JPG images. The target pedestrian was occluded 9 times and the total frame number of occlusion images was 1379. Results show that the robust enhanced multi filter tracking algorithm can judge occlusion accurately and retrieve the target without confusing with other pedestrians once the target reappears after occlusion. In Fig. 11, each row shows a process of the target being occluded and then be retrieved. Continuous detection during target occlusion lengthened the overall consuming time. The average FPS is about 25. Several challenging videos in the vot2017 database are used to test the proposed tracking algorithm. The girl, bmx, crossing, gymnastics1, iceskater2 are adopted since the target in these sequences undergoes frequent occlusion and interferences. As shown in Fig. 12, the propose tracking algorithm can recognize occlusion and interference from
similar objects and handle them very well. Even if the target undergoes drastic appearance change, the tracker can adapt very well and locate the target correctly.

If the target changes dramatically, the multi-filter tracking algorithm can rely on multiple filters to remember its different typical appearances. When the target is lost due to occlusion or other reasons, the algorithm uses the detector to search for the target near the lost location. Then if the target is not found for a long time, it will search in the whole map. Since a pretrained detector can only locate objects of target’s category, it may mistakenly locate other similar objects. Through a variety of typical target appearances stored in filters to validate detector location results can further confirm the target and avoid confusion, so in the experiments, other pedestrians’ occlusion does not affect tracking of the target pedestrian, the duet partner’s blocking did not affect the target dancer’s normal tracking. Without occlusion happens, detector can also play the role of correcting target bounding boxes’ position. General tracking algorithm in the case of target’s dramatic appearance changes is often very inaccurate or even lost the target. Inaccurate location will increase the computational load and affect the speed. The tracking algorithm aided by detector can greatly reduce the background interference and computation burden. The experimental results show that the multi-filter tracking algorithm can track and locate the target accurately even when the target changes violently.

V. CONCLUSION

All kinds of tracking algorithms try to use the maximum target appearance information captured in history, but only increasing complexity and parameter numbers to realize single filter’s fitting of various target appearances will often lead to redundancy and low speed. Multiple filters can record multiple typical target appearances shown in different periods. Target location and filter updating are directed to the filter corresponding to current target appearance, so time passing and filter updating won’t deteriorate the preserved historical typical target appearance. When the previous typical target appearance reappears, the tracker can quickly identify it with the corresponding filter. Besides, target appearance information recorded by multiple filters is independent of each other, and even the introduction of wrong training samples won’t pollute all historically preserved target appearance information. If the target appearance changes rapidly during the tracking process, the first appeared target typical appearance won’t be captured correctly, and deviation of the bounding box from the ground truth box will affect the following tracking process. But for a known category target, the off-line training model can be introduced to capture the target’s new appearance accurately. Although the detection process is more time-consuming, the precision and speed can be balanced by tuning the threshold of turning to the detector, and accurate detector can remove the extra background parts from the target bounding box and improve the tracking speed to a certain extent. Combining the stored typical target appearances with detection is of great significance to distinguish the target from other similar objects of target’s category.

REFERENCES


