Sports Target Tracking Based on Discriminant Correlation Filter and Convolutional Residual Network

Yuan Linlin¹ and Yao Liu²

¹College of Physical Education, Huaqiao University, Xiamen, China
²College of Science and Engineering, National Quemoy University, Kinmen, Taiwan

Correspondence should be addressed to Yao Liu; yowk0529@gmail.com

Received 9 February 2022; Accepted 14 March 2022; Published 4 April 2022

Academic Editor: Xin Ning

Copyright © 2022 Yuan Linlin and Yao Liu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

During the sports tracking process, a moving target often encounters sophisticated scenarios such as fast motion and occlusion. During this period, erroneous tracking information will be generated and delivered to the next frame for updating; the information will seriously deteriorate the overall tracking model. To address the problem mentioned above, in this paper, we propose a convolution residual network model based on a discriminative correlation filter. The proposed tracking method uses discriminative correlation filters as basic convolutional layers in convolutional neural networks and then integrates feature extraction, response graph generation, and model updates into end-to-end convolutional neural networks for model training and prediction. Meanwhile, the introduction of residual learning responds to the model failure due to changes in the target appearance during the tracking process. Finally, multiple features are integrated such as HOG (histogram of oriented gradient), CN (color names), and histogram of local intensities for comprehensive feature representation, which further improve the tracking performance. We evaluate the performance of the proposed tracker on MultiSports datasets; the experimental results demonstrate that the proposed tracker performs favorably against most state-of-the-art discriminative correlation filter-based trackers, and the effectiveness of the feature extraction of the convolutional residual network is verified.

1. Introduction

Target tracking is a characteristic problem in the field of visual technology research, which is widely used in intelligent monitoring, autonomous driving, robot visual perception, and other scenarios [1–3]. In recent years, with the rapid development of the sports industry, the tracking of the goals in the complex sports scenarios represented by basketball and football has gradually attracted attention. Target tracking problems in sports scenarios are more challenging than the most widely studied pedestrian monitoring scenarios, such as more severe occlusion and appearance interference, more dramatic posture changes, and more complex forms of movement. In the face of complex and changeable external environment changes and motor target deformation, computers can often cause target drift and target loss. Therefore, the research direction still faces great challenges and space for progress.

At present, the conventional visual tracking algorithm can be simply divided into two methods, including the correlation filter (CF-) based [4] and deep learning-based [5–7] methods. Both of these methods can improve the algorithm performance from different aspects. Since the correlation filter tracking algorithm can effectively solve the problem of ridge regression in the Fourier domain, it can greatly relieve the computing burden and then realize the real-time tracking. However, most CF-based approaches employ limited hand-crafted features such as HOG [8, 9], CN [10], or a combination of these features [11, 12] for feature representation. However, those approaches achieve sub-optimal performance in comparison to these deep features. The deep learning-based tracking approaches usually employ the pretrained deep high-dimensional features extracted from convolutional neural networks (CNNs) for feature representation. Although desirable tracking results are achieved, a large amount of computational complexity...
is added in the feature extraction process and will severely affect real-time performance.

In the tracking process, the updating process of the tracking model is crucial to decide whether or not tracking is accurate. At present, the common algorithm [8, 9] mostly updates the tracking model from frame to frame. This updating method has some shortcomings. When the target encounters some significant appearance change such as illumination changes, occlusion, and out of view, this will generate some erroneous tracking information. The information will be delivered to the next frame, after accumulating for a long time, and will increase the risk of tracking drifting. It is difficult to recover to the original status and will eventually cause tracking failure.

In this paper, we consider the problems mentioned above and focus on how to design a robust strategy for updating the model to further improve the quality of the model. The correct tracking information in the previous frames can be delivered to the next frame in time, and the model can be correctly updated when the target objects undergo sophisticated scenarios such as occlusion and illumination variation, which keep the model of subsequent video sequences from being deteriorated to some extent. The main contributions of this work can be summarized as follows:

1. A convolution residual network model based on a discriminative correlation filter is proposed that follows the tracking-by-detection paradigm, which can efficiently prevent the model from being deteriorated due to the delivery of erroneous tracking information.

2. The proposed algorithm integrates multiple powerful discriminative hand-crafted features such as HOG, CN, and histogram of local intensities to achieve preferable feature representation and further enhance the overall performance of the algorithm by taking advantage of various features.

3. In addition, an accurate scale estimation approach is integrated into our algorithm for improving scale adaptability, and an accurately self-adaptive way to update the scale model and the translation model is proposed. The algorithm performance is evaluated on MultiSports datasets. The tracking results demonstrate that the proposed method performs favorably against existing state-of-the-art discriminative CF-based trackers.

2. Related Work

2.1. Discriminative Correlation Filter-Based Tracking Algorithms. Among most existing tracking-by-detection methods, discriminative correlation filters have recently achieved excellent predominant tracking results on multiple standard tracking benchmarks [12, 13] and have been widely used in visual tracking algorithms due to their efficient computing performance. The complicated convolution calculation in the frequency domain can be converted into a simple element-wise multiplication operation in the Fourier domain, by employing the properties of a circular matrix and the fast Fourier transform (FFT), which enables trackers to quickly learn and detect in the frequency domain. A large number of correlation filter-based methods have been proposed one after another. Bolme et al. [14] first introduced the theory of correlation filters into the visual tracking and learned a kind of minimum output sum of squared error (MOSS) filter based on the grayscale image samples. In addition, they utilized PSR as the model update criterion, by analyzing the strength of response map peak to evaluate the tracking performance. Based on MOSS, Henriques et al. [4] utilized the kernel circular structure to track and utilize the circular matrix to sample densely. The training and detection of classifiers are performed in the Fourier domain, which can realize high-speed tracking. The disadvantage of this method is that the boundary effects are caused by the circular shift, and only the grayscale features are utilized.

2.2. Deep Learning in Tracking Algorithms. Recently, most of the tracking algorithms have begun to use the deep learning method to replace the traditional hand-crafted features [8–10] with information-rich deep features [5–7]. A large number of methods based on deep learning have been proposed. Wang and Yeung [5] were the first to apply deep networks to single-object tracking and forward the formulation of “offline pretraining and online fine-tuning,” which largely solved the problem of insufficient training samples in tracking. Ma et al. [6] utilized the hierarchical convolution features for appearance representation and then obtained the hierarchical layer response map by correlation filter and integrated all of the hierarchical layer response map linearly, which greatly improved the accuracy of the algorithm. Song et al. [7] redefined discriminative correlation filters as a layer of a convolutional neural network, and they integrated features extraction, response map generation, and model updating into the convolutional neural network for end-to-end training, which effectively improved the accuracy of the tracking performance. Zhu et al. [15] proposed a joint convolutional tracking method, which regarded the process of feature extraction and tracking as convolution operation and trains them simultaneously. In addition, they introduced a peak-versus-noise ratio (PNR) criterion as the model updating method to avoid tracking drift, which achieves desirable tracking performance.

3. Tracking Component

3.1. The Discriminative Correlation Filter-Based Framework. In this section, we first introduce some a priori knowledge about the context-aware correlation filter framework used in our algorithm, and then we introduce information concerning the fundamental theory about how the scale correlation filter integrates into our algorithm [1].

The most common discriminative correlation filter-based trackers always tend to ignore the surrounding contextual information. Nevertheless, the surrounding contextual region around the target location plays an important
role in tracking performance. The context-aware framework based on discriminative correlation filter is proposed by Mueller et al. [16], which incorporates global context information into the learned filter. The goal is to train a filter that has a high response to the target image patch and a near-zero response to the context region. In our work, we choose the context-aware correlation filter (CACF) as our translation filter; for a more detailed derivation, please see [16].

In the CACF framework, the main objective is to train the optimal correlation filter \( w \), for all of the training samples \( D_0 \) (\( D_0 \) contains all circular shifts of the vectorized image patch \( d_0 \) generated by the circular shift operator and the desirable regression target \( y \) (\( y \) is a vectorized image of a 2D Gaussian). The following ridge regression problem will be optimized efficiently by

\[
\min_w \|D_0 w - y\|_2^2 + \lambda_1 \|w\|_2^2.
\]

According to the property of the circulant matrix in the Fourier domain, we will train a filter \( w \), which has a high confidence response for the target image patch and a near-zero confidence response for the context patches. By adding the context patches as a regularization term to standard formulation (see Equation (1))

\[
\min_w \|D_0 w - y\|_2^2 + \lambda_1 \|w\|_2^2 + \lambda_2 \sum_{k=1}^k \|D_k w\|_2^2,
\]

\( \lambda_1 \) and \( \lambda_2 \) are regularization weight parameters, where \( \lambda_2 \) is applied to control the context patches regressed to zeros. \( D_0 \in \mathbb{R}^{n \times n} \) and \( D_k \in \mathbb{R}^{n \times n} \) are the corresponding circulant matrices. Since the target image patches contain many context image patches and form a new data matrix \( B \in \mathbb{R}^{(k+1) \times n} \), the main objective function (2) can be rewritten as follows:

\[
f_y(W, B) = \|Bw - \tilde{y}\|_2^2 + \lambda_1 \|w\|_2^2,
\]

where

\[
B = \begin{bmatrix}
A_0 \\
\sqrt{\lambda_2} A_1 \\
\vdots \\
\sqrt{\lambda_2} A_k
\end{bmatrix}
\]

and

\[
\tilde{y} = \begin{bmatrix}
y \\
0 \\
\vdots \\
0
\end{bmatrix}
\]

which can be minimized by derivation operation as follows:

\[
w = (B^TB + \lambda I)^{-1} B^T \tilde{y}.
\]

According to the property of the circulant matrix, the closed-form solution of the standard formulation (see Equation (6)) can be solved efficiently in the Fourier domain:

\[
\hat{w} = \frac{\hat{a}_0 \circ \tilde{y}}{\hat{a}_0 \circ \hat{a}_0 + \lambda_1 + \lambda_2 \sum_{k=1}^k \hat{a}_k \circ \hat{a}_k}.
\]

The translation estimation of the target objects is performed by the learned filter \( \hat{w} \) convolving with the image patch \( z \) (search window) in the next frame. The location of the maximum response of all training sample response vectors \( y_r(z,w) \) is the predicted location of the target. For a given single image patch \( z \), the output response is given by

\[
f(z) = \mathcal{F}^{-1}(z \circ w) = \tilde{z} \circ \hat{w},
\]

where \( \mathcal{F}^{-1} \) denotes the inverse Fourier transformation and \( \circ \) denotes the convolution operation. Then, we update the filter model by employing the following equations:

\[
\hat{w}_i = \eta w_i + (1 - \eta) \hat{w}_{i-1},
\]

\[
\hat{x}_i = (1 - \eta) \hat{x}_{i-1} + \eta \tilde{x}_i,
\]

where the subscript \( i \) denotes the sequence number of the current frame, \( \eta \) is the learning rate parameter, and \( \tilde{x}_i \) denotes the appearance model of the target object.

### 3.2. The Basic Convolutional Layer.

Song et al. [3] formalize the single-feature discriminant correlation filtering (DCF) algorithm into the basic convolution layer. Inspired by this, this paper formalize multiple features into the basic convolution layer and use the neural network backpropagation algorithm to train a single convolutional layer to equivalently replace the training mode of the traditional correlation filtering algorithm. For the input \( x \), the corresponding output Gaussian response is \( y \), and then training the discriminative correlation filter \( w \) can be converted to solving the following minimization problem.

\[
w^* = \arg\min_w \|w * x - y\|_2^2 + \lambda \|w\|_2^2,
\]

where \( \lambda \) denotes regularization parameter. The loss function is expressed as

\[
L(w) = \frac{1}{N} \sum_{i=1}^N L_w(x^{(i)}) + \lambda r(w),
\]

where \( N \) denotes sample size, \( L_w(x^{(i)}) (i \in N) \) denotes loss function for the \( i \)th sample, and \( r(w) \) denotes the weight decay. When \( N = 1 \), \( r(w) \) is formalized to the L2 norm.
And formula (12) is expressed as
\[
L(w) = L_w(x) + \lambda \|w\|^2, \quad (13)
\]

When \( L_w(x) = \|F(x) - Y\|^2 \), \( F(x) \) denotes neural network output, \( Y \) denotes truth value, and \( L_w(x) \) is equivalent to the L2 loss between \( F(x) \) and \( Y \). The loss function (see Equation (13)) is equivalent to the discrimination correlation filter (see Equation (11)).

3.3. The Convolutional Residue Learning. Figure 1 shows the structure of the basic convolutional layers and the residual learning layers, if the ideal response map for the input frame \( K \) is represented as \( F_B(K) \); then, the residual learning \( F_R(K) \) can be expressed as
\[
F_R(K) = FH(K) - F_B(K). \quad (14)
\]

Therefore, after adding residual learning, the final actual response result \( F_R(K) \) of frame \( K \) can be represented as
\[
F(K) = F_B(K) + F_R(K) = F_B(K, \{w_B\}) + F_R(K, \{w_R\}). \quad (15)
\]

The temporal residuals are used to capture the differences when the spatial residues are invalid, and its network structure is similar to the spatial residual structure. The temporal residual input is extracted from the first frame containing the initial object appearance. If \( K_t \) represents frame \( t \) of the input \( K \), the spatial-temporal residuals can be expressed as
\[
F(K_t) = F_B(K_t) + F_{SR}(K_t) + F_{TR}(K_t), \quad (16)
\]

where \( F_B(K_t) \) denotes the basic convolutional layer response value, \( F_{SR}(K_t) \) denotes the spatial residual layer response value, and \( F_{TR}(K_t) \) denotes the temporal residual layer response value. When the appearance of the target object experiences small changes, the basic convolutional layer output varies very little from the real value, and the residual layer has a weak effect on the final response results. When the appearance of the target object experiences large changes, such as rapid motion, it is difficult to distinguish the target from the background, and the residual layer compensates for the difference between the basic convolution layer output and the true soft tag (the maximum response target predicted by the previous frame serves as the true soft tag).

4. The Proposed Algorithm

In this section, we mainly introduce the principle of our proposed algorithm. First, we introduce the general model updating methods of visual tracking and were inspired by these methods. Then, we introduce our proposed model updating strategy. Finally, we introduce the integration of multiple hand-crafted features for representation in our algorithm. Figure 2 illustrates the rough flowchart of the proposed algorithm.

4.1. The Model Initialization and Online Detection. In this paper, the network containing parameters has a basic convolutional layer, three temporal residual layers, and three spatial residual layers, updated objects for each layer of weight parameter \( w \) and bias parameter \( b \). First, the underlying convolutional layer and the residual layer parameters are randomly initialized according to the zero-mean Gaussian distribution. Then, given the first frame image with the target position, the training block is extracted around the target center position. The training block is introduced to the feature extraction module (pretraining VGG-16 network) for feature extraction. The iterative training stops when the loss function \( f(w, b) \) (see Equation (15)) is less than the set threshold of 0.02; the model parameter initialization is complete. Forward propagation and backpropagation are used for the update parameters by employing the following equations:
\[
\begin{align}
Z^l & = w^l A^l - 1 + b \\
A^l & = g^l(Z^l) \\
w^l_i & = w^l_i - \eta \frac{\partial}{\partial w^l_i} f(w, b) \\
b^l_i & = b^l_i - \eta \frac{\partial}{\partial b^l_i} f(w, b),
\end{align} \quad (17)
\]

where \( l = (0, 1, 2) \) represents the basic convolutional layers, temporal layers, and spatial residual layers, respectively. For forward propagation, the network \( l = (1, 2, 3, \cdots) \) layer objective function value is \( Z^l \), and the current layer output \( A^l \) is acquired by activation function \( g^l(Z^l) \) (ReLU function). During backpropagation, the current \( w^l_i \) minus the product of the loss function on the \( w^l_i \) partial derivative and the learning rate \( \eta \) is the updated weight. The current \( b^l_i \) minus the product of loss function on the \( b^l_i \) and the learning rate \( \eta \) is the updated bias value.

In online detection, when a new frame appears, the search block of the same size as the training block is extracted based on the predicted target center location of the previous frame, and the search block is input into the model to generate a response map in the model. The maximum response value (see Equation (16)) is used as the positioning target.

4.2. The Scale Estimation and Self-Adaptive Model Updating. Motivated by the above method [15, 17], in this paper, we regarded the PSR of the response map that was pointed out in [14], and its response map peaks as our tracking quality evaluation [1]. The expression equation is given by
\[
PSR = \frac{R_{\text{max}}(x) - \mu_{s_1}(R(x))}{\sigma_{s_1}(R(x))}, \quad (19)
\]

where \( R_{\text{max}}(x) \) is the maximum scores of the response map, \( R \), \( s_1 \) is the peak side lobe region around the peak, which is
15% of the response map area in this paper, and $\mu_s$ and $\sigma_s$ are the mean value and standard deviation of the side lobe area. A desirable response map should have only one sharp peak, and the peak area shows that it is relatively prominent centered around the target location, while the other areas show a relatively unapparent target location [17]. In this paper, the maximum response score is taken as a dynamic threshold, due to the corresponding response map peak of each video sequence varies from sequence to sequence. Whether the model is updated or not depends on $PSR$ and corresponding response peak score at each frame. Thus, the model can adaptively judge the results according to a maximum response score of each different video sequence. Only when the certain condition is met where the $PSR$ is greater than its maximum response peak score $R_{\text{max}}(x)$, the translation filter model (see Equation (8)) and the convolutional residual model (see Equation (17)) will be updated online with a learning rate parameter $\eta$ (see Equations (9) and (18)). The tracking result is regarded as accurate in the current frame. It is effectively preventing the incorrect update information from being transmitted to subsequent frames that lead to tracking drift, to create a self-adaptive way of updating the scale model and the translation model. The overall flowchart of the proposed algorithm is as shown in Figure 2.

Figure 3 illustrates the effect of the proposed accurate update strategy. We can see from the figure that the target encounters a sophisticated appearance change such as occlusion, from #187 to #190 (point B and C). At this time, the $PSR$ score obviously decreases—that is, the $PSR$ score decreases from 7.088 to 6.672. When the $PSR$ score is just lower than the given dynamic threshold (the maximum response score $R_{\text{max}}$, the score is approximately 7.28 in sequence Basketball1), the accurate proposed update method will choose not to update the model in the current frame. To avoid erroneous tracking information, it will deteriorate the tracking model. When the target leaves out of the occluded area at #194 (point D), corresponding $PSR$ scores rise significantly—that is, the $PSR$ score rises from 6.672 to 10.09. At this time, the tracking model updating should be considered, and the tracking result is considered to have high confidence at this frame.
4.3. The Integration of Multiple Hand-Crafted Features for Representation. The visual feature representation is an important part of the visual tracking framework. In general, existing trackers employ features which include hand-crafted features and deep features. Our works mainly focus on hand-crafted features. The common hand-crafted features include HOG features and CN features [1]; both of them have their own advantages and disadvantages. The HOG feature is widely employed for most existing trackers [8, 9, 18] and object detection [19]. It maintains superior invariance to both geometric and optical deformation by calculating and completing a statistical histogram of the gradient direction in a spatial grid of cells of the image patch to generate the feature. The CN feature [10] descriptor utilized the PCA (principal component analysis) technique [11] for dimensionality reduction, which has been successfully employed for tracking due to its preferable invariance for objective shape and scale. The histogram of local intensities (HOI) [18] is complementary of HOG features by computing the histogram of local intensities.

We consider integrating both features of the HOG, CN, and HOI analyzed above in the stage of feature extraction for the sake of achieving superior feature representation and complementing each of their advantages. In this paper, the proposed framework combines multiple features such as HOG and CN and uses another pixel intensity histogram feature of intensity in a $6 \times 6$ local window with 8 bins (the same setting as [18]) for feature representation and then employs the response map generated by the integrated multifeature for translation estimation.

Figure 4 illustrates the tracking results of the proposed algorithm with different hand-crafted features on different hand-crafted features on the Basketball1 sequence.
MultiSports datasets [20]. We know that the proposed algorithm, with the integration of multiple hand-crafted features, outperforms the algorithm with only the HOG feature. This finding further demonstrates the effectiveness of our methods regarding feature representation.

5. Experiments

5.1. Experiment Implementation Details. The hardware configuration used for the experiment is as follows: Intel(R) Xeon(R)-E-2124G 3.40 GHz CPU, RTX 2080Ti memory for 16 GB. The proposed tracker is implemented in MATLAB2018a, TensorFlow is selected for the deep learning framework, and VGG-16 is selected for the feature extraction network. In the training phase, the Adam optimizer is used to calculate the update coefficient iteratively. We use HOG features in a $4 \times 4$ local window with 31 bins. The regularization parameters $\lambda_1$ and $\lambda_2$ in Equation (7) are set to $1e-4$ and 0.4, respectively. The size of the search window is set to 2.2 times the target size. The spatial bandwidth is set to $1/10$. The learning rate $\eta$ in Equations (9) and (18) is set to 0.025. The number of scale space $S = 33$, and the scale factor $a$ is set to 1.02. The PSR$_0$ (the PSR initial value) is set to 1.8. We use the same parameter values for all of the sequences.

5.2. Overall Tracking Results on MultiSports Datasets. To validate the performance of our proposed tracker, we evaluate our tracker on MultiSports datasets and compare it with 10 state-of-the-art discriminative correlation filter-based trackers including SRDCF [21], LMCF [17], LCT [18], KCF [8], CSK [4], SAMF [22], DSST [9], DCF_CA [16], SAMF_CA [16], and MOSSE_CA [16]. We use three metrics provided by the MultiSports datasets [20] to evaluate the 10 trackers. MultiSports datasets (https://github.com/MCG-NJU/MultiSports/) contain 66 fine-grained action categories from four different sports, selected from 247 competition records. The records are manually cut into 800 clips per sport to keep the balance of data size between sports, where

![Figure 5: Distance precision and overlap success plots comparing the proposed tracker with state-of-the-art discriminative correlation filter-based trackers on MultiSports datasets using one-pass evaluation (OPE).](image)

![Table 1: The tracking results of the distance precision and the overlap success plots comparisons with state-of-the-art trackers perform on the MultiSports datasets.](table)
Figure 6: Attribute-based evaluation of the distance precision and overlap success plots over four challenges of basketball, football, volleyball, and gymnastics. The number of sequences for each attribute is shown in brackets.
we discard intervals with only background scenes, such as an award, and select the highlights of competitions as video clips for target tracking. The distance precision (DP) is defined as the percentage of frames whose predicted location is within the given threshold distance of the ground truth, and the threshold is generally specified as 20 pixels. The overlap successful plot (OP) is defined as the percentage of frames whose overlap rate surpasses a certain threshold; the threshold is generally specified as 0.5. The center location error (CLE) shows the average Euclidean distance between the ground truth and the predicted target center location. We report the tracking results in a one-pass evaluation (OPE) using a distance precision plot and overlap success plot, as shown in Figure 5, with comparison to the aforementioned trackers. We use the distance precision plot and the area-under-the-curve (AUC) of success plot as a criterion to rank the trackers (see Table 1).

Figure 5 and Table 1 illustrate the distance precision and overlap success plots of eleven trackers on MultiSports datasets. As seen in Table 1, the proposed tracker performs favorably against existing trackers in distance precision (DP) and overlap success (OS). The proposed tracker achieves desirable results with an average DP of 79.0, which outperformed SRDCF (75.4), LMCF (72.8), and LCT (76.7). The overlap success plots maintain similar accuracy (0.537) to SRDCF (0.552) and outperform LMCF (0.530). In the speed aspect, our tracker mainly employs the computation efficiency of the CFs in the frequency domain and multiple hand-crafted features for tracking. The tracking speed outperforms SRDCF (5.3), LCT (20.5), and DSST (27.6) trackers, which obtained a real-time speed of 46.8 fps. These results demonstrate the effectiveness of the ability of the tracking model to update adaptively and combine multiple features efficiently for feature representation.

5.3. The Attribute-Based Tracking Results. We evaluate the attribute-based evaluation on MultiSports datasets [20], which include 40 video sequences. All of these videos in the datasets are annotated by 4 attributes containing different challenging scenarios as the following: basketball, football, volleyball, and gymnastics. We report the results of partial attributes as shown in Figure 6, and the number...
shown on the heading indicates the number of datasets with this challenge attribute. We can see that our tracker obtained superior tracking performance in almost all of the displayed attributes.

Figure 6 illustrates that the proposed algorithm performs well with distance precision and overlap success plots in four attribute challenges, and it shows that the proposed method achieved superior DP in attributes of basketball (74.8%), football (79.2%), volleyball (77.5%), and gymnastics (73.8%). The overlap success plots achieve the second-best performance in all of the displayed attributes. These results demonstrate that an effective model update method will improve tracking accuracy to some extent, especially in the football attribute.

5.4. Qualitative Evaluation. Figure 7 illustrates the qualitative comparisons of the proposed tracker with eight discriminative correlation filter-based trackers perform on MultiSports datasets [20], including LMCF [17], SRDCF [21], STAPLE [12], LCT [18], SAMF [22], KCF [8], DSST [9], and CSK [4]. We can see that the proposed tracker performs well in scenarios with occlusion (Basketball2), fast motion (Basketball3, Football, and Volleyball), and motion blur (Gymnastics).

In the Basketball2 sequence, we know that the target undergoes an occlusion challenge from #150 to #158; our tracker can still locate the target object accurately. We did not choose the updated tracking model in case of serious occlusion. The other trackers such as KCF cause tracking drift after serious occlusion, which demonstrates the effectiveness of the proposed methods.

In the Basketball3, Football, and Volleyball sequences, the target object undergoes challenges such as fast motion. The KCF, CSK, LCT, and LMCF trackers fail to track the object. Our tracker can track the object from the beginning to the end of sequences successfully because we integrate multiple powerful discriminative features and the updating strategy in our tracker.

In the Gymnastics sequence, the target mainly undergoes challenges such as motion blur. Most trackers such as LMCF and KCF lose the object and cause drift; our tracker can track the object correctly; it demonstrates that the proposed method is robust to motion blur.

6. Conclusion
In this paper, a convolution residual network model based on a discriminative correlation filter is used. The response peak score generated by the discriminative correlation filter is utilized as a dynamic threshold, with comparisons with the peak side lobe ratio of the response map at each frame; afferent convolutional neural networks serve as basic convolutional layers, and the comparative result is then used as the differentiated conditions for updating the translation filter model and the scale filter model; the introduction of residual learning responds to the model failure due to changes in the target appearance during the tracking process, which achieve a self-adaptive updating method. At the feature extraction stage, multiple powerful discriminative hand-crafted features are integrated such as HOG, CN, and a pixel intensity histogram for comprehensive feature representation. The experimental results demonstrate that the proposed tracker performs well against most state-of-the-art discriminative correlation filter-based trackers and performs well in sophisticated scenarios of Basketball, Football, Volleyball, and Gymnastics.

Data Availability
The data included in this paper are available upon request to the corresponding author without any restriction.

Conflicts of Interest
The authors declare that they have no conflicts of interest to report regarding the present study.

Acknowledgments
This work was supported by the Fujian Provincial Science and Technology Major Project (No. 2020HZ02014) and Quanzhou Science and Technology Major Project (No. 2021GZ1).

References
[1] Z. Chen, P. Liu, Y. Du, Y. Luo, and J. M. Guo, “Robust visual tracking using self-adaptive strategy,” Multimedia Tools and Applications, vol. 79, no. 1, pp. 141–162, 2020.
[2] M. Wang, T. Sun, K. Song, S. Li, J. Jiang, and L. Sun, “An efficient sparse pruning method for human pose estimation,” Connection Science, vol. 34, no. 1, pp. 960–974, 2022.
[3] Y. B. Song, C. Ma, L. J. Gong et al., “CREST: convolutional residual learning for visual tracking,” in IEEE International Conference on Computer Vision, pp. 2555–2564, Venice, October 2017.
[4] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting the circulant structure of tracking-by-detection with kernels,” European Conference on Computer Vision, vol. 2012, pp. 702–715, 2012.
[5] N. Wang and D.-Y. Yeung, “Learning a deep compact image representation for visual tracking,” Advances in Neural Information Processing Systems, vol. 71, pp. 809–817, 2013.
[6] C. Ma, J. B. Huang, X. Yang, and M. H. Yang, “Hierarchical convolutional features for visual tracking,” in IEEE International Conference on Computer Vision, pp. 3074–3082, Santiago, Chile, 2015.
[7] S. Yibing, C. Ma, G. Lijun, Z. Jiawei, W. H. L. Rynson, and Y. Ming-Hsuan, “CREST: convolutional residual learning for visual tracking,” in IEEE International Conference on Computer Vision, pp. 2574–2583, Venice, Italy, 2017.
[8] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 583–596, 2015.
[9] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, “Accurate scale estimation for robust visual tracking,” British Machine Vision Conference, Nottingham, 2014, Bmva Press, September 2014, 2014.
[10] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive color attributes for real-time visual tracking," Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1090–1097, Columbus, OH, USA, 2014.

[11] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1-3, pp. 37–52, 1987.

[12] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: complementary learners for real-time tracking," IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1401–1409, Las Vegas, NV, USA, 2016.

[13] T. C. Lu, “CNN convolutional layer optimisation based on quantum evolutionary algorithm,” Connection Science, vol. 33, no. 3, pp. 482–494, 2021.

[14] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 2544–2550, San Francisco, CA, 2010.

[15] Z. Zhu, G. Huang, W. Zou, D. Du, and C. Huang, "UCT: learning unified convolutional networks for real-time visual tracking," IEEE International Conference on Computer Vision, 2017, pp. 1973–1982, Venice, Italy, 2017.

[16] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1396–1404, Honolulu, HI, USA, 2017.

[17] M. Wang, Y. Liu, and Z. Huang, "Large margin object tracking with circulant feature maps," IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4800–4808, Honolulu, HI, USA, 2017.

[18] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5388–5396, Boston, MA, USA, 2015.

[19] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627–1645, 2010.

[20] Y. Li, "MultiSports: a multi-person video dataset of spatio-temporally localized sports actions," IEEE International Conference on Computer Vision, 2021, pp. 13536–13545, Nashville, TN, USA, 2021.

[21] M. Danelljan, G. Hager, and F. S. Khan, "Learning spatially regularized correlation filters for visual tracking," International Conference on Computer Vision, 2015, pp. 4310–4318, Santiago, Chile, 2015.

[22] Y. Li and J. Zhu, “A scale adaptive kernel correlation filter tracker with feature integration,” in European Conference on Computer Vision, pp. 254–265, Springer, Cham, 2014.