Improved C-COT based on feature channels confidence for visual tracking

Shuwei SHEN*, Shengjing TIAN*, Leyuan WANG*, Aihong SHEN*,** and Xiuping LIU*

* Department of Mathematical Science, Dalian University of Technology
  Dalian, Ganjingzi, Liaoning 116024, China
  E-mail: xpliu@dlut.edu.cn

** Department of Basic Courses, Criminal Investigation Police University of China
  Shenyang, huangguqu, Liaoning, 110854, China

Abstract
In the field of visual tracking, the methods of Discriminative Correlation Filters (DCF) have showed excellent performance, which rely heavily on the choice of feature descriptors. The Continuous Convolution Operator Tracker (C-COT) is a novel correlation filter to track the target position in the continuous domain, which achieved significant effects. However, as for various visual scenes, different feature descriptors are suitable to different environments. If each feature channel is given the same confidence during the tracking phase, it would limit the performance of some good features. To address this problem, this paper proposes an improved C-COT algorithm that can adaptively perform feature channel weighting. The Average Peak Correlation Energy (APCE) is used to evaluate the corresponding response map of each feature channel, guiding the target appearance model to give different weights to different features. Then, we can obtain the final weighted feature response map whose peak value is applied to locate the target. In addition, the C-COT updates the appearance model rigorously in every frame, which may lead to over-fitting and increase computational complexity. Therefore, in order to reduce the redundancy of the online training sample and avoid similar background interference, we adopt the method of Peak Side Lobe Ratio (PSLR) to update the model. We perform comprehensive experiments on OTB50 and OTB100. The results show that the improved tracker achieves better accuracy, especially in some specific video scenes. In addition, speed has also improved.

Keywords: Visual tracking, Feature channel confidence, Continuous filtering, Confidence update, Feature descriptors

1. Introduction

Visual tracking is one of the most important research topics in the field of computer vision. It has been widely used in intelligent security video surveillance, autonomous vehicle navigation, robot vision system, intelligent human-computer interaction, intelligent transportation, etc. (Fortune, 1987), and receives extensive attention from researchers. During the actual tracking process, in order to adapt to the complex and various environments, robustness and real-time are required. In recent years, although the target tracking has made great progress and development, there are still many challenging issues that need to be solved, such as non-rigid deformation of the target, occlusion, fast motion, and background clutter.

The visual tracking algorithm is mainly divided into two ways, one is generative model and another is discriminative model (Wang et al., 2015). Recently, the method based on discriminative correlation filtering has performed well in some public databases (Wu et al., 2013; Wu et al., 2015) with its higher success rate and robustness. The method of correlation filtering that takes advantage of the property of the circulant matrix trains a linear ridge regression classifier and uses the fast fourier transform to solve the closed solution, which allows the filter to have a rich training sample set and a high frame rate. The Minimum Output Sum of Squared Error filter tracker MOSSE (Bolme et al., 2010) applies the idea of correlation filtering to visual tracking for the first time. It uses the single gray features to obtain a very high frame rate, but
its tracking results are not prominent. Based on MOSSE, the Circulant Structure of Tracking-by-detection with Kernels tracker CSK (João et al., 2012) introduces the kernel function technique. Subsequently, Henriques et al. (2015) and Danelljan et al. (2014a) propose multi-channel HOG features and multi-channel Color Names features (CN) respectively based on the CSK, which greatly improved the accuracy of tracking algorithm. However, the correlation filtering is not ideal for fast motion and has a boundary effect. In order to alleviate this problem, Danelljan et al. proposed a spatial regularization correlation filter to suppress the filter coefficient values near the boundary, solved by the Gauss-Seidel method. In addition, the scale estimation is also a common problem in visual tracking. The Scale Adaptive with Multiple Features tracker SAMF proposed by Li et al. (2014), the Discriminative scale space tracker DSST and fast DSST (fDSST) proposed by Danelljan et al. (2014; 2017) improve the scale estimation performance of tracker.

With the appearance of deep learning, in addition to HOG features, color features and so on, deep convolution features are commonly used in image representations. Considering the important role of features in visual tracking, Danelljan et al. (2016) propose a multi-resolution continuous spatial domain filtering learning method for object tracking based on Color Names features and deep convolution features. The C-COT integrates multi-resolution feature maps into the continuous spatial domain through an interpolation model and then uses the conjugate gradient method to train the continuous domain filter, which is an important evolutionary algorithm of DCF. Previous works (Bolme et al., 2010; João et al., 2012; Li et al., 2014; Danelljan et al., 2014, 2016, 2017) have been to directly sum the feature channels. Such results tend to make the contribution of each feature channel the same. However, this strategy may limit the performance of some features, which plays an significant role for object recognition. Therefore, we evaluate the response map of each feature by the APCE and use it to guide the appearance model through giving different weights to the filters. Finally, we can obtain the final weighted summation response map, which helps us to locate the target.

The updating strategy plays an important role in tracking problem. So, how to automatically adapt to the changes of the target and simultaneously avoid the tracking failure caused by the model drift is the key to the problem. The C-COT uses a continuous learning strategy and it updates the model rigorously in every frame. This over-update strategy can lead to over-fitting and it also reduces the speed of tracking. Therefore, in order to improve the quality of the training samples, we apply the improved method PSLR to update the model.

In view of the deficiencies of the original C-COT algorithm, the main improvement works of this paper are summarized as follows:

1. We propose an adaptive feature block weighting method based on C-COT. The weight of each filtering is calculated by the APCE of the corresponding feature block response map, and finally the target position is determined by the weighted feature response map.

2. Using PSLR as a basis for model updating helps prevent similar background interference and overfitting phenomena, which can improve accuracy and speed at the same time.

3. In OTB100, compared with the original C-COT tracker, our algorithm significantly improves the tracking accuracy in occlusion (OCC), motion blur (MB), deformation (DEF) and low resolution (LR).

Fig. 1 shows an overview of our algorithm. The channel reliability scores are used for weighting the per-channel filter responses in localization. It can also be used in most modern correlation filters.

2. Related Work

Recently, a group of correlation filter (CF) based trackers (Danelljan et al., 2014a, 2014b; Henriques et al., 2015; Bertinetto et al., 2016; Zhang et al., 2014; Bolme et al., 2010) have sparked a lot of interests due to their significant computational efficiency and robustness. CF enables training and detection with densely-sampled examples and high dimensional features in real time by using the fast Fourier transform (FFT). Bolme et al. (2010) propose a minimum output sum of squared error (MOSSE) filter, modeling target appearance by learning an adaptive correlation filter which is optimized by minimizing the output sum of squared error. Henriques et al. (2016) exploit the circulant structure of shifted image patches in a kernel space and propose the CSK method based on intensity features, and extend it to the KCF approach (Henriques et al., 2015) with the HOG descriptors. In addition, to deal with the scale problem, SAMF (Li et al., 2014), DSST (Danelljan et al., 2014) and an improved Kernelized Correlation Filter (Henriques et al., 2015) have been proposed subsequently and achieved state-of-the-art performance. The winner of the challenge, DSST (Danelljan et al., 2014), incorporate a multi-scale template for Discriminative Scale-Space Tracking using a 1D Correlation Filter. The imperfect training example issue is addressed in the studies (Danelljan et al., 2015, 2016) by applying a spatial regularization on the correlation filter to increase the search range. One deficiency of Correlation Filters is that they are
constrained to learn from all circular shifts. Several recent works (Danelljan et al., 2014, 2016; Kala et al., 2012) have sought to resolve this issue, and the Spatially Regularised formulation in particular has demonstrated excellent tracking results. However, this was achieved at the cost of real-time operation. Recently, in (Ma et al., 2016; Yao et al., 2016), correlation filters are learned independently for each type of feature. Choi et al. (2016, 2017) propose the integrated tracking system to handle the various type of correlation filters with attentional mechanism. C-COT is another important evolutionary algorithm of DCF, which has achieved good results on VOT2016.

With the advent of deep CNNs, fully connected layers of the network have been commonly employed for image representation (Oquab et al., 2014; Simonyan et al., 2014). Recently, the last (deep) convolutional layers are shown to be more beneficial for image classification (Cimpoi et al., 2015; Liu et al., 2015). To overcome the insufficient representation of the hand-crafted features, deep convolutional features are utilized in the correlation filters (Danelljan et al., 2015, 2016) which have achieved the state-of-the-art performance. The first deep learning tracker is deep learning tracker (DLT) (Wang et al., 2013), which using a multi-layer autoencoder network. Kristan et al. (2015) construct multiple CNN classifiers on different instances of target objects to rule out noisy samples during model update. Hong et al. (2015) learn target-specific saliency map using a pre-trained CNN. Then, a novel fully-convolutional Siamese network (Bertinetto et al., 2016) achieve dense and efficient sliding-window evaluation. Based on the Siamese framework of Bertinetto et.al., the CFNet (Valmadre et al., 2017) enables ultra-lightweight networks of a few thousand parameters to achieve state-of-the-art performance on multiple benchmarks while running at high frame rates. Danelljan et al. (2016) also propose a novel correlation filter to find the target position in the continuous domain, while incorporating features of various resolutions. Their framework shows state-of-the-art performance with deep convolution features. C-COT uses the deep neural network VGG-net to extract the features, interpolates the feature maps with different resolutions into the continuous space domain by cubic interpolation, and then uses the Hessian matrix to find the target position of the sub-pixel precision.

Our main contribution is to put feature confidence into the learning framework, using APCE metrics to weight different feature channels in order to improve accuracy. Finally, the model is updated by the Peak Side Lobe Ratio (PSLR), through this way the accuracy and speed can both improve.

3. Baseline Approach: C-COT

The baseline in this paper is Continuous Convolution Operators For Visual Tracking (C-COT) (Danelljan et al., 2016). C-COT achieved excellent performance in the competition VOT2016 and the OTB100. Compared with traditional
DCF, this method allows multi-channel fusion of different resolution features. In addition, the response map of this method can be expressed as a continuous function and has higher accuracy in locating the target position. The following briefly introduces the C-COT algorithm.

3.1. Continuous space domain convolution operator

Let \( \{x_j\}_{j=1}^M \subset \chi \) be a collection of \( M \) training samples. Each sample has \( D \) feature channels, which are \( x^d_j \) \( d = 1, \cdots, D \). Unlike traditional DCF, the dimension of each feature channel \( x^d_j \in \mathbb{R}^{N_d} \) is \( N_d \); the sample space \( \chi \) is \( \mathbb{R}^{N_1} \times \cdots \times \mathbb{R}^{N_D} \). The algorithm realizes the learning of the continuous spatial domain by interpolating training samples. Set the interval \( [0, T) \subset \mathbb{R} \) to be the local support of the map, for each feature channel \( d \), define the interpolation operator as \( J_d : \mathbb{R}^{N_d} \to L^2(T) \), the specific form is as follows,

\[
J_d(x^d)(t) = \sum_{n=0}^{N_d-1} x^d[n]b_d(t - \frac{T}{N_d}n),
\]

where \( b_d \) is an interpolation function with period \( T > 0 \) and \( n \in \{0, \cdots, N_d - 1\} \). Use \( J \{x\} \) to represent the feature map of all channels in the sample after interpolation, where \( J \{x\} (t) \in \mathbb{R}^D \).

The target of the C-COT is to learn multi-channel convolution operators \( S_f : \chi \to L^2(T)^D \). The operator \( S_f \) is parameterized by a set of convolution filters \( f = (f^1, \cdots, f^D) \in L^2(T)^D \). Here \( f^d \in L^2(T) \) is a continuous filter of the characteristic channel \( d \). Therefore, the circular convolution operator is defined as,

\[
S_f \{x\} = f * J \{x\} = \sum_{d=1}^{D} f^d * J_d \{x_d\}.
\]

The maximum value of the final response map \( S_f \{x\} (t) \) corresponds to the location of the target, where \( t \in [0, T) \).

3.2. Filter training

Given \( M \) sample pairs \( \{(x_j, y_j)\}_{j=1}^M \subset \chi \times L^2(T) \), the labeled detection scores \( y_j(t) \) corresponding to sample \( x_j \) is taken by a periodic Gaussian function. Similar to the traditional DCF, the C-COT learns the filters by minimizing the following objective,

\[
E(f) = \sum_{j=1}^{M} \alpha_j \left[ \|S_f \{x_j\} - y_j\|^2_{L^2} + \sum_{d=1}^{D} \|w f^d\|^2_{L^2} \right],
\]

where, \( \alpha_j > 0 \) is the weight of the sample \( x_j \), and \( w \) is the spatial regularization coefficient that satisfies the Gaussian distribution to suppress the boundary effect.

Equation (3) can be transformed into a more manageable optimization problem through Fast Fourier Transform. Applying Parseval’s formula we can get,

\[
E(f) = \sum_{j=1}^{M} \alpha_j \left[ \|S_f \{x_j\} - \tilde{y}_j\|^2_{L^2} + \sum_{d=1}^{D} \|\tilde{w} \ast \tilde{f}^d\|^2_{L^2} \right],
\]

where, \( \wedge \) represents the Fourier transform of the function. The Fourier transform of the target response of Equation (2) is given by the formula \( S_f \{x_j\} = \sum_{d=1}^{D} \tilde{f}^d \wedge x^d \wedge \), where \( \wedge x^d \) is the discrete Fourier Transform of \( x^d \).

In practice, we let the Fourier transform of the continuous filter \( f_d \) be a finite non-zero element \( \{\tilde{f}^d[k]\}_{K_d} \), where \( K_d = \left\lfloor \frac{N_d}{2} \right\rfloor \). Equation (4) can convert to a quadratic optimization problem, optimized by solving the normal equations,

\[
(A^H \Gamma A + W^H W) \tilde{f} = A^H \Gamma \tilde{y}.
\]

Where \( \tilde{f} \) and \( \tilde{y} \) are respectively vectorized representations of \( f^d \) and \( y_j \) in the Fourier coefficients. The diagonal blocks of matrix \( A \) contain elements of \( x^d_j[k] b_d[k] \). In addition, \( \Gamma \) is a diagonal matrix with weights \( \alpha_j \), and \( W \) is a convolution matrix related to \( w[k] \). When the number of non-zero elements in \( \tilde{w}[k] \) is very small, the linear system is sparse, and it can be solved iteratively by the conjugate gradient method.

4. Proposed Method

4.1. Adaptive feature weighting

Considering each feature map has different description capabilities for the current frame image, we should entrust the feature that possesses more discriminative representation ability. In other words, if each feature map is given the same
degree of confidence, the performance of some features may be limited. Fig. 2 shows that the response maps of different feature channels has different influences in the same frame. Therefore, in order to improve the reliability of the tracking target, we propose an adaptive feature-weighted improved C-COT algorithm.

We introduce the confidence score (Alan et al., 2017) to evaluate the quality of the learned filter corresponding to each feature channel. During online tracking stage, the confidence score reflects the reliability of the feature which would be treated as the weight of the filters. Finally the response map will be obtained by weighted sum of filters corresponding to feature channels.

When the tracked target is matched well, the desired response map should only have one sharp peak surrounded by smooth region. Otherwise, the entire response map will fluctuate dramatically. The sharper the correlation peak, the better the positioning accuracy. Thus a straight-forward measure of channel learning reliability is the maximum response of a learned channel filter. However, the maximum response is not enough to represent the degree of oscillation of the response map, which is the key factor to measure the quality of the response map. so how to calculate the weight is particularly important. Here, the average peak correlation energy (APCE) (Wang et al., 2017) is used to evaluate the quality of the response map for each feature channel. It is defined as follows,

\[
APCE = \frac{|F_{\text{max}} - F_{\text{min}}|^2}{\text{mean}(\sum_{w,h}(F_{w,h} - F_{\text{min}})^2)},
\]

where \(F_{\text{max}}\) and \(F_{\text{min}}\) are the maximum value and minimum value of the response map \(F\) respectively. The APCE indicates the degree of fluctuation of the response map. According to the formula, we can find that when the response map only has single peak surrounded by the other smooth areas, the success rate of capturing target is higher, and the APCE value is large. Conversely, if the APCE value is small, the target may be lost. The response map corresponding to this feature channel will affect the precision of tracking result. Based on this phenomenon, the feature that possesses a large APCE will be assigned higher weight. A small average peak correlation energy indicates the feature channel has no significant effect on the target tracking.

In practice, we divide feature channels into feature blocks for reducing computational redundancy. Each feature block corresponds to a kind of feature representation like HOG. Supposing there are \(N\) feature blocks, each feature block has \(D^{(n)}\) channels. Let \(\beta^{(n)}\) be the APCE of the feature block and obviously there exist \(\eta_{d}^{(n)}\) that satisfy \(\beta^{(n)} = \sum_{d=1}^{D^{(n)}} \eta_{d}^{(n)} \beta_{d}^{(n)}\). Finally, we set the feature block weight as \(\lambda^{(n)}\), whose specific formula is as follows,

\[
\lambda^{(n)} = \frac{\beta^{(n)}}{\sum_{n=1}^{N} \beta^{(n)}},
\]

When there are multiple similar objects in the vicinity of the target, it will also bring about the oscillation of response map, futher leading to a small APCE value, even if the peak is the desired position. At this time, if the APCE values are considered as the criterion of reliability, it will lead to incorrect weighting. Therefore, in order to prevent improper punishment, the ratio is clamped by 0.5. The weight of feature block changes into \(\lambda^{(n)} = \max(\lambda^{(n)}, 1/2)\).

Since the feature channels in each feature block have the same weight, i.e., \(\lambda^{(n)} = \lambda_{d}^{(n)}\), the convolution operator is redefined as follows,

\[
S_{f} \{x\} = f * J \{x\} = \sum_{n=1}^{N} \sum_{d=1}^{D^{(n)}} \lambda_{d}^{(n)} f^{d} * J_{d} \{x^{d}\}.
\]
4.2. Model updating strategy

Most existing trackers update the appearance model every frame, regardless of whether the detection result is accurate or not and whether the current frame is reliable or not. In practice, once the target is drift or severely occluded, online updating of the appearance model will acquire an incorrect classifier, eventually resulting in the tracking of failure. In addition, updating frame by frame also limits the tracking efficiency, see the case in Fig. 3. In view of this, we consider using the feedback from the tracking results to determine the necessity of model updating. Specifically, according to the response map, we should apply an appropriate criteria which can reveal whether the tracking results are drift and occluded or not to determine the model updating in the current frame. Here, we take advantage of the peak side lobe ratio (PSLR) as a confidence criteria. It is defined as the ratio of the peak intensity of the main lobe to the peak intensity of the strongest side lobe. The greater the peak side lobe ratio value, the higher the reliability of the model updating. It can be writen as,

$$PSLR = \frac{g_{\text{max}} - \mu_{s1}}{\sigma_{s1}}.$$  

(9)

Where, $g_{\text{max}}$ is the maximum value of the response map, $\mu_{s1}$ and $\sigma_{s1}$ is the mean and variance respectively. When the peak side lobe ratio reaches a certain threshold, the model is updated, otherwise, do not update. This not only can improve the tracking accuracy, but also can effectively avoid the model updating errors caused by the target drift. Moreover, it will reduce the number of updating and enhance the tracking speed. Here, the threshold is determined by the experiment.

5. Experimental results and analysis

We have implemented relevant experiments of the proposed algorithm FCW-CCOT to verify the performance of our method.

We perform experiments on OTB50 (Wu et al., 2013) and OTB100 (Wu et al., 2015). All these sequences are annotated with 11 attributes which cover a variety of challenging factors, including scale variation (SV), occlusion (OCC), illumination variation (IV), motion blur (MB), deformation (DEF), fast motion (FM), out-of-plane rotation (OPR), background clutters (BC), out-of-view (OV), in-plane rotation (IPR) and low resolution (LR). We evaluate through one-pass evaluation (OPE). Two performance measures are used. One is Overlap precision (OP), which is defined as the percentage of frames in a video where the intersection-over-union overlap exceeds a threshold of 0.5. Another is Area-under-the-curve (AUC), which is computed from the success plot, where the mean OP over all videos is plotted over the range of threshold [0,1].

Firstly, we introduce the details of experiments in Section 5.1. Then, in Section 5.2 , we introduces the self-contrast experiment of this algorithm. Next, we analyze the comparison experiment with baseline in Section 5.3. And finally, in Section 5.4, we analyze the comparison experiment with some current mainstream algorithms on OTB2013 and OTB2015.
5.1. Experiment details

Our tracker is implemented in Matlab on a 4.00GHz CPU PC. We use the same feature representation as the C-COT, that is, the combination of the first (Conv-1), the forth (Conv-4) and the fifteenth (Conv-15) convolutional layer in the VGG-m network. In order to select a suitable threshold for our model during updates, a threshold test based on the peak side lobe ratio is performed. First, we set the threshold in (Wu et al., 2013; Guo et al., 2014) with 1 as the step length. From the experiment result as shown in Fig.4, it was found that at the threshold 6, the precision and success rate reach the highest value, which are 0.665 and 0.887 respectively. Other specific parameters are shown in Table 1. Please note that all parameters in the experiment remain fixed.

| Parameter                              | Value |
|----------------------------------------|-------|
| Learning rate                          | 0.0075|
| The number of scale parameters         | 5     |
| Scale factor                           | 1.02  |
| Model update threshold                 | 6     |

Fig. 4 The precision and success rate of threshold experiments.

5.2. Self-contrast experiment

In order to demonstrate the effect of the proposed high confidence update strategy, we first test our algorithm with different versions on OTB100. We denote the tracker of this paper without high confidence updating strategy as FCW-CCOT-NU, and the tracker with high confidence updating strategy is represented as FCW-CCOT. Fig.5 shows the tracking results of FCW-CCOT on the OTB100.

As shown in Fig.6, FCW-CCOT shows the best tracking success rate and the highest precision in the OPE evaluation metrics. Without a weight update strategy, FCW-CCOT-NU gets a poor performance because of it updates the tracking model in each frame, the incorrect results are likely leading to unexpected updates. Also, this will result in the fact that operating efficiency is lower than FCW-CCOT and thus the tracking speed is significantly reduced to about half of the FCW-CCOT.

As shown in table 2, according to the experimental results, using a high confidence update strategy can significantly improve the tracking performance both in success rate and tracking speed. It is mainly due to more reliable updating of the model and no need to update each frame.

5.3. Comparison experiment with C-COT

In order to verify the effectiveness of our tracker and show its improved performance, we compare it with the base-
Fig. 5  Precision plot and success rate plot.

Fig. 6  Comparison of updates and no updates. The red line represents the updated tracker and the green dotted line represents no update.

line algorithm C-COT. We conducted experiments on 100 videos in OTB100. The OTB100 dataset created 100 video sequences to evaluate the performance of the trackers. The initial bounding box is sampled in time and space to evaluate the tracker’s robustness and other performance. Performance is evaluated both in terms of precision and success rate.

We compare the visual attributes of FCW-CCOT and C-COT as shown in Fig.7. They are respectively low resolution, out of view, deformation, motion blur, occlusion. The orange line is the baseline tracker C-COT, and the blue line is the tracker of this paper FCW-CCOT. From the radar chart, the results of our algorithm have better performance in these visual attributes. And especially under the motion blur attribute and the occlusion attribute, the success rate of FCW-CCOT is 2.3% and 2.1% higher than C-COT. And as shown in figure 8, the average success rate and precision of this algorithm are 1.0% and 1.0% higher than those of C-COT, respectively, have been improved. Fig.9 is the captured result images from the image sequences. The red box is the tracking result of our tracker, and the green box is the tracking result of C-COT. The image sequence of Bird1 has the properties of deformation and fast motion. From the tracking results, can find that our tracker can still track the target correctly when the bird passes through the cloud, but the C-COT cannot track the target correctly because each frame is updated which resulting in the incorrect results. The image sequences of Bike and Panda have dimensional changes, occlusion, out-of-plane rotation, and low-resolution video properties. It can be find that

Table 2  The frame rate and tracking results of FCW-CCOT-NU and FCW-CCOT.

| Algorithm    | Frame rate | Success rate | Precision |
|--------------|------------|--------------|-----------|
| FCW-CCOT-NU  | 0.7129     | 0.658        | 0.878     |
| FCW-CCOT     | 1.4307     | 0.665        | 0.887     |
Fig. 7  Expected success performance on different visual attributes on the OTB100 benchmark. The proposed FCW-CCOT and C-COT are shown. The orange line represents C-COT and the blue line represents the FCW-CCOT.

Fig. 8  Comparison of FCW-CCOT and C-COT. The red line represents FCW-CCOT and the green dotted line represents C-COT.
for this kind of video, our tracker tracks better than C-COT and does not appear to be lost.

Fig. 9 The tracking results of bird1, bike and panda. The red bounding box indicates the tracking result of FCW-CCOT and the green one indicates the tracking result of C-COT.

5.4. Comparison with other algorithms

We compare our tracker with other trackers and perform experiments on the OTB50 and OTB100 datasets, respectively. In order to ensure the fairness of comparison, we have unified experiments in our computer and for the trackers involved in the comparison, we use the same parameters as the original.

Fig. 10 and Fig. 11 has shown the success plot and precision plot of all of the ten trackers. We summarize the experimental results in Table 3, which are the success rate and precision of the ten trackers on OTB50 and OTB100. By analyzing the experimental results, our tracker achieved the best performance in comparing with the other nine trackers. It is worth mentioning that ECO is also an improved tracking device based on C-COT. Compared with the best performing ECO in 2017, although the speed is difficult to compare favorable with, the success rate and precision have exceeded, the tracking accuracy has also been improved.
Fig. 11  Precision plots and success rate plots for OPE on the OTB100 benchmark.

Table 3  Comparison of various algorithms.

| algorithm | FCW-CCOT | ECO | C-COT | ADNET | MCPF | SiamFC | Staple | CXT | CSK | DFT |
|-----------|----------|-----|-------|-------|------|--------|--------|-----|-----|-----|
| Success rate (OTB50) | 0.621 | 0.614 | 0.593 | 0.595 | 0.583 | 0.516 | 0.507 | 0.323 | 0.309 | 0.273 |
| Success rate (OTB100) | 0.665 | 0.663 | 0.655 | 0.646 | 0.628 | 0.582 | 0.578 | 0.412 | 0.382 | 0.328 |
| Accuracy (OTB50) | 0.858 | 0.777 | 0.819 | 0.729 | 0.767 | 0.639 | 0.636 | 0.419 | 0.400 | 0.341 |
| Accuracy (OTB100) | 0.887 | 0.804 | 0.877 | 0.797 | 0.795 | 0.717 | 0.702 | 0.515 | 0.493 | 0.406 |
In Fig. 12, we show the tracking results of our tracker and other nine trackers on the eight video sequences. Among them, the red box represents our tracker. We can see that our tracker performs particularly well with occlusion properties and dramatic deformation properties, and it is not easy to drift. This is exactly what the model update strategy effects.

5.5. Limitations and future work

While we are able to track most of the challenging videos more accurately, but because we only use the CNN features, the video with dramatic deformation and strong illumination variation will not achieve good performance as show in figure 13. At the same time, even if our model update strategy improves the accuracy of video with occlusion properties, we can’t avoid the tracking failure of some challenging cases. In addition, although the speed has been improved, it is still impossible to achieve real-time tracking. In view of the above problems, in future work, except CNN features, we can consider combining Color Names features and HOG features. Mean while, it is possible to improve the tracking speed by judging the response map. We first use the low features, if the response map is good enough to determine the target, we do not need to using the deep features.
6. Conclusion

Based on C-COT, we propose an improved C-COT algorithm FCW-CCOT that adaptively performs feature channels weighting for each frame of image. Firstly, the average peak correlation energy is used to evaluate the response map corresponding to each feature block, and this guides the appearance model to give different weights to different filters, finally obtains the final weighted feature response map, from which we can locate the target by the peak value. Secondly, by applying the peak side lobe ratio (PSLR) as a basis for model updating, it is beneficial to prevent similar background interference and over-fitting, and at the same time, it can improve accuracy and speed. Compared with the most advanced trackers, it shows that our method has improved the performance of C-COT, and has an excellent results among most of the trackers.

References

Alan Lukežič, Tomáš Vojíř, Luka Čehovín, Jiří Matas and Kristian, M. , Discriminative Correlation Filter with Channel and Spatial Reliability (2017), pp.4847-4856, IEEE Conference on Computer Vision and Pattern Recognition.

Bertinetto, L. , Valmadre, J. , Henriques, João F. , Vedaldi, A. and Torr, P. H. S. , Fully-convolutional siamese networks for object tracking (2016), pp.850-865, IEEE Conference on Computer Vision and Pattern Recognition.

Bertinetto, L. , Valmadre, J. , Golodetz, S. , Miksik, O. and Torr, P. H. S. , Staple: Complementary Learners for Real-Time Tracking (2016), pp.2,5,7, International Conference on Computer Vision and Pattern Recognition (CVPR).

Bolme, D. S. , Beveridge, J. R. , Draper, B. A. and Lui, Y. M. , Visual object tracking using adaptive correlation filters (2010), pp.2544-2550, The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition.

Choi, J. , Chang, H. J. , Jeong, J. , Demiris, Y. and Choi, J. Y. , Visual Tracking Using Attention-Modulated Disintegration and Integration (2016), pp.4321-4330, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Choi, J. , Chang, H. J. , Yun, S. , Fischer, T. and Choi, J. Y. , Attentional Correlation Filter Network for Adaptive Visual Tracking (2017), pp.4828-4837, IEEE Conference on Computer Vision and Pattern Recognition.

Cimpoi, M. , Maji, S. and Vedaldi, A. , Deep filter banks for texture recognition and segmentation (2015), pp.3828-3836, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Danelljan, M. , Häger, G. and F.S. Khan. , Accurate scale estimation for robust visual tracking (2014b), pp.65.1-65.11, British Machine Vision Conference.

Danelljan, M. , Häger, Gustav, Khan, F. S. and Felsberg, M. , Discriminative scale space tracking (2016), pp.1561-1575, IEEE Transactions on Pattern Analysis and Machine Intelligence.

Danelljan, M. , Häger, G. , Khan, F. S. and Felsberg, M. , Convolutional Features for Correlation Filter Based Visual Tracking (2015), pp.58-66, 2015 IEEE International Conference on Computer Vision Workshop (ICCVW).

Danelljan, M. , Khan, F. S. , Felsberg, M. and Weijer, J. V. D. , Adaptive Color Attributes for Real-Time Visual Tracking (2014a), pp.1090-1097, 2014 IEEE Conference on Computer Vision and Pattern Recognition.

Danelljan, M. , Robinson, A. , Khan, F. S. and Felsberg, M. , Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking (2016), pp.472-488, European Conference on Computer Vision.

Fortune, S. , A sweepline algorithm for voronoi diagrams (1987), pp.153–174, Algorithmica.

Galoogahi, H. K. , Sim, T. and Lucey, S. , Correlation filters with limited boundaries (2014), pp.4630-4638, Computer Vision and Pattern Recognition.

Guo, Y. , Chen, Y. , Tang, F. , Li, A. , Luo, W. and Liu, M. , Object tracking using learned feature manifolds (2014), pp.128-139, Computer Vision and Image Understanding.

Henriques, J. F. , Caseiro, R. , Martins, P. and Batista, J. , High-speed tracking with kernelized correlation filters (2015), pp.583-596, IEEE Transactions on Pattern Analysis and Machine Intelligence.

Hong, Z. , Chen, N. Z. , Wang, C. , Mei, X. , Prokhorov, D. and Tao, D. , Multi-Store Tracker (MUSTer): A cognitive psychology inspired approach to object tracking (2015), pp.749-758, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

João F. Henriques, Caseiro, R. , Martins, P. and Batista, J. , Exploiting the Circulant Structure of Tracking-by-Detection with Kernels (2012), pp.702-715, Proceedings of the 12th European conference on Computer Vision - Volume Part IV.

Kalal, Z. , Mikolajczyk, K. and Matas, J. , Tracking-learning-detection (2011), pp.1409-1422, IEEE Transactions on Software Engineering.
Kristan, M. et al., The Visual Object Tracking VOT2015 Challenge Results (2015), pp.564-586, ICCV Workshops.
Liu, L., Shen, C. and Hengel, A. V. D., The treasure beneath convolutional layers: cross-convolutional-layer pooling for image classification (2014), pp.4749-4757, IEEE Conference on Computer Vision and Pattern Recognition.
Li, Y., and Zhu, J., A scale adaptive kernel correlation filter tracker with feature integration (2014), pp.254-265, Computer Vision - ECCV 2014 Workshops.
Ma, C., Huang, J. B., Yang, X. and Yang, M. H., Hierarchical Convolutional Features for Visual Tracking (2015), pp.3074-3082, 2015 IEEE International Conference on Computer Vision (ICCV).
Nam, H. and Han, B., Learning multi-domain convolutional neural networks for visual tracking (2015), pp.4293C4302, IEEE Conference on Computer Vision and Pattern Recognition.
Oquab, M., Bottou, L., Laptev, I. and Sivic, J., Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks (2014), pp.1717-1724, 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Simonyan, K. and Zisserman, A., Very deep convolutional networks for large-scale image recognition (2014), Computer Science.
Valmadre, J., Bertinetto, L., Henriques, João F., Vedaldi, A. and Torr, P. H. S., End-to-end representation learning for correlation filter based tracking (2017), pp.5000-5008, IEEE Conference on Computer Vision and Pattern Recognition.
Wang, N. and Yeung, D. Y., Learning a Deep Compact Image Representation for Visual Tracking (2013), pp.809-817, International Conference on Neural Information Processing Systems. Curran Associates Inc.
Wang, N., Shi, J., Yeung, D. Y. and Jia, J., Understanding and diagnosing visual tracking systems (2015), pp.3101-3109, IEEE International Conference on Computer Vision.
Wang, M., Liu, Y. and Huang, Z., Large margin object tracking with circulant feature maps (2017), pp.4800-4808, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Wu, Y., Lim, J. and Yang, M. H., Online Object Tracking: A Benchmark (2013), pp.2411-2418, 2013 IEEE Conference on Computer Vision and Pattern Recognition.
Wu, Y., Lim, J. and Yang, M. H., Object Tracking Benchmark (2015), pp.1834-1848, IEEE Transactions on Pattern Analysis and Machine Intelligence.
Yao, H., Lim, J., Qin, L., Yang, M.H., Huang, Q., Zhang, S. and Qi, Y., Hedged deep tracking (2016), pp.4303-4311, Computer Vision and Pattern Recognition.
Zhang, K., Zhang, L., Liu, Q. and Yang, M. H., Fast visual tracking via dense spatio-temporal context learning (2014), pp.127-141, European Conference on Computer Vision, ECCV2014.