International Journal of Advanced Computer Technology (IJACT)
ISSN: 2319-7900 Volume VIII, Issue II, April, 2019

KCF TRACKING ALGORITHM BASED ON VGG16 DEPTH FRAMEWORK

Sida Zhang[1] , Information Engineering College of HAUST, Luoyang 471003, China;
Wei Li , Information Engineering College of HAUST, Luoyang 471003, China;
Renfang Wang ,College of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo 315100,China;
[1]starzhang0621@gmail.com

Abstract

In order to solve the problem that the KCF tracking algorithm has occlusion or deformation and the disturbance factors such as similar objects cause tracking failure; this paper proposes an improved algorithm combining VGG-16 neural network. Firstly, the VGG-16 network's powerful feature extraction capability is used to extract features that are more robust to deformation and occlusion from different layers and different operations. Then, using the cyclic shift matrix of KCF algorithm, a large number of sample training classifiers are generated, and then new images are calculated. The filtering response of the block predicts the target position; in order to improve the real-time performance of the algorithm, the model and the new strategy for the KCF algorithm reduce the computational complexity by updating the model with a fixed frame interval. Compared with the traditional KCF algorithm, this method can effectively deal with the interference factors such as deformation and occlusion, and can achieve target tracking more quickly while ensuring accuracy.

Key words : Video tracking; KCF algorithm; VGG-16 neural network; equal interval frame update

Introduction

Target tracking analyzes the video sequence and calculates the position, size and speed of the target in each frame of the image. It is of great significance and wide application in the fields of video surveillance, human-computer interaction, visual effects, weapon guidance, air traffic, etc.

Since the introduction of AlexNet in 2012, deep learning has been rapidly applied to all areas of computer vision[1]. It is also rapidly applied in the target tracking field. The more classic networks include VOT2015 champion MDNet, VOT2016 champion TCNN, real-time SiamFC and GOTURN. These algorithms have two drawbacks. First, the speed is slow and real-time is difficult to guarantee. Second, it is easy to overfit and perform well in the test set, but it is difficult to achieve the nominal effect in practical applications. The advantages are also very obvious. The deep neural network has powerful feature extraction ability[2], which can effectively cope with various interference factors such as illumination, deformation and occlusion; there is no boundary effect, no need to add cosine window; it naturally supports GPU acceleration.

The related filtering class tracking algorithm shines on the OTB100 test in 2014. Among them, the KCF algorithm [3] and the DCF algorithm greatly lead the Struck algorithm with the best effect on the OTB50 in accuracy and speed, and the related filtering algorithm. The MOOSE algorithm [4], although only 43% accurate, has reached the speed of 615 FPS. Therefore, correlation filtering has quickly become the most interesting algorithm in the target tracking field.

The KCF algorithm was proposed by João F. Henriques et al. at Oxford University [3], and the core part is the ridge regression and multi-channel HOG features. Then, Martin Danelljan of Linköping University improved to Multi-channel color feature (Color Names CN) [5], and reduced the color attribute by principal component analysis, but neither of them had scale update, and the target scale deformation easily led to drift and failure. YangLi of Zhejiang University, based on the KCF algorithm [6], simultaneously extracts HOG features and CN features to complement each other, and then combines the translation filter to target tracking on multi-scale scaled images. Martin Danelljan returns to HOG features. Combined with DCF for translational position detection, special training MOSSE-related filters are used to detect scale changes [7]. This type of approach greatly increases the robustness of correlation filtering in terms of scale variation interference.

In addition to the target scale deformation, the boundary effect is also a problem that is difficult to solve by the related filtering algorithm. The strategy adopted by the KCF algorithm is to add a cosine window, but this will cause some pixels to be filtered out, and the response is globally maximum, resulting in tracking failure. Martin Danelljan solves the marginal effect problem by ignoring the boundary part pixels of all displacement samples, or limiting the filter coefficients near the boundary, and adopting a larger detection area, adding a spatial regularization method, but the speed is only 5fps [8]; Hamed Kiani It is proposed to solve the boundary effect by using a larger size detection image block and a smaller size filter to improve the ratio of the true sample, but the accuracy is insufficient [9].

With the deepening of the research, pure CNN methods and related filtering methods combined with depth features have appeared in large numbers, and the results are excellent. Among them, C-COT combined with multi-layer depth features ranks first in VOT2016 [10], and its idea is a continuous convolution operator of multi-resolution depth feature mapping. Scholars have come to realize that the lack of correlation factors such as scale deformation, occlusion, fast motion and boundary effects can be solved by using different layers of convolutional neural networks, and the speed advantage of related filtering methods is maintained. Based on this, Wang used the different layer features proposed by VGG-16 neural network for target tracking[11], and carried out in-depth analysis. The features extracted by different layers can cope with different interference factors, and the resolution of high-level features is low. It has stronger semantic information and is insensitive to the deformation of the target appearance. The low-level features have high resolution, which is suitable for expressing the local detail information of the target object, and has better robustness against interference such as occlusion and similar object interference. By simultaneously extracting the features of different convolutional layers and

weighting the response output, the robustness of the tracker to various interference factors can be improved.

In view of the shortcomings of KCF algorithm in dealing with scale deformation, occlusion, similar object interference, this paper proposes a target tracking algorithm combining VGG-16 neural network and KCF algorithm, and uses VGG-16 network to extract different layer features and calculate weighted filtering response. The target object is tracked in real time using an equal interval frame update strategy. The experimental results show that the robustness of the algorithm to interference factors is improved under the premise of guaranteeing speed and accuracy. The KCF algorithm can be divided into feature extraction, sample generation, training classifier, calculation of filter response, and updating of classifier parameters.

1. Improved classifier

It can be seen from the above analysis that the performance of the classifier directly affects the success rate of the tracker. When different influencing factors are applied to the target object, the performance of the classifier is improved, and the response can be better under various interference factors, which can effectively improve the tracking algorithm. Great, improve tracking success rate. In this paper, multiple classifiers are trained in different layers of the VG Classifier training can be divided into the following steps: feature extraction, sample generation, and training.G-16 neural network.

1.1 Feature extraction

In order to solve the tracking failure caused by scale deformation, occlusion, similar object interference and other factors in KCF algorithm, this paper uses VGG-16 neural network as the basic network to extract different layer features and analyze the influence of different layer features on tracking effect and speed. The layer is used to train the classifier to improve the robustness of the KCF algorithm.

The VGG-16 network model consists of 8 Conv3 convolutional layers, 5 pooling layers, and 3 fully connected layers. The structure is shown in Figure 1:

It can be seen from Fig. 1 that the upper level expresses the overall information of the target, and the resolution is low; the low level expresses the target detailed information, and the local remains intact, and can cope with interference such as occlusion. Based on this, this paper proposes to use different layer feature combinations for KCF tracking algorithm. Using low-level high-resolution features, local details maintain good characteristics, improve the robustness of KCF algorithm to occlusion, similar object interference, etc., use high-level network low resolution, but express the characteristics of the target as a whole, improve the KCF algorithm to cope with scale changes The robustness, high-level features complement each other, training the classifier separately, and improving the robustness of the KCF algorithm to the interference factors such as occlusion, scale variation and similar object interference. The different layers extracted by the VGG-16 network model have different semantics, and the visualization results are shown in Figure 2 below. It can be seen from the above figure

that the upper level expresses the overall information of the target with low resolution; the low level expresses the target detailed information, and the local remains intact, and can cope with interference such as occlusion. Based on this, this paper proposes to use different layer feature combinations for KCF tracking algorithm. Using low-level high-resolution features, local details maintain good characteristics, improve the robustness of KCF algorithm to occlusion, similar object interference, etc., use high-level network low resolution, but express the characteristics of the target as a whole, improve the KCF algorithm to cope with scale changes The robustness, high-level features complement each other, training the classifier separately, and improving the robustness of the KCF algorithm to the interference factors such as occlusion, scale variation and similar object interference.

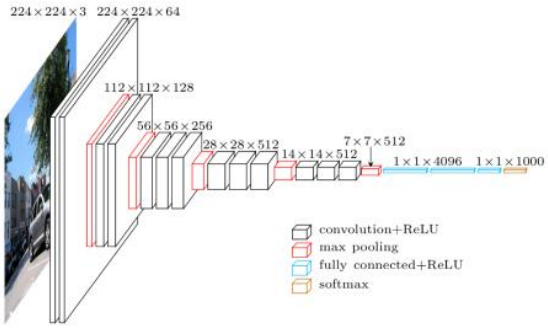

Figure 1. VGG network

Figure 2 VGG network model visualization

Table 1.Effect of different layer feature combinations on tracking performance

Feature	Average accuracy	Speed
C3_3	72.1%	74.2FPS
C4_3	72.%8	73.5FPS
C5_3	73.4%	71.7FPS
P3	76.1%	69.8FPS
C3_3+C4_3+C5_3	80.2%	51.3FPS
C3_3+C4_3+P3	79.3%	66.4FPS

In the VGG-16 network model is used to extract different features[12]. After analyzing the characteristics of different layers, the author uses the 3rd, 4th, and 5th pooling layers to extract features for tracking, and assigns weights to the three layers of 0.5, 1, 0.02. Robustness to illumination changes has been achieved. This paper extracts the features of each layer in the VGG network and uses the bilinear interpolation to uniformly scale to 224*224 size. It is used in the KCF algorithm for experiments. The experiment uses the average precision, that is, the prediction center position and the true center position error are smaller than this experimental design. The number of frames of the threshold 30 is a percentage of the total

number of frames of the video. The results are organized in table 1. It can be seen from the above table that the results obtained by the combination of C3_3, C4_3 and P3 layers are not only high precision, but also faster. Therefore, the C3_3, C4_3 and P3 layer combinations are used for feature extraction, and the classifiers are trained separately.

1.2 Sample generation

In the KCF algorithm, a large number of samples are generated by using a cyclic displacement matrix, and the feature matrix is gradually shifted to obtain different samples.

Extracted for the l networks VGG-16 d -dimensional feature layer x_l^d , the cyclic matrix generated by the sample, the nature of the diagonalized by Fourier transform, i.e. according to formula (1):

$$X = Fdiag(\hat{x})F^H \quad (1)$$

In the formula, F denotes a Fourier transform matrix, x is obtained by x after DFT, and H denotes a symbol of conjugate transpose.

1.3 Training classifier

In the KCF algorithm, sample training is considered as a ridge regression problem, and a linear regression function is introduced:

$$f(x_i) = \omega^T x_i \quad (2)$$

Because we extract the characteristics of different operations of different layers and layers, we need to train different classifiers separately. For each classifier, the residual function of sample set $\{x_l\}$ and label set $\{y_l\}$ is minimized [13], and calculation is considered. Complexity and over-fitting, for each classifier, the training purpose can be expressed by Equation 3.

$$\min_{w_l} \sum_{l=1}^{N} (f(x_l) - y_l)^2 + \lambda \| \omega \|^2 \quad (3)$$

Where ω_l represents a layer l classifier, N represents the total number of samples, and λ is used for parameter normalization.

Equation 4-7 consists of the only optimal solution:

$$\omega_l = (X_l^H X + \lambda I)^{-1} X_l^H y \quad (4)$$

Where I is the identity matrix, transform the feature into the Fourier domain, and note that the transformed feature is X , which can avoid matrix inversion and matrix product, reduce computational complexity, from $O(n^3)$ to $O(n \log n)$, so the correlation filtering algorithm It is much faster than the deep learning based tracking algorithm.

In order to deal with nonlinear features, KCF algorithm uses the kernel function to transform the linear model into a nonlinear model based on the ridge regression to improve the performance of the classifier. Because this is not the focus of this article, it is no longer exhaustive.

2. Forecasting and updating

For the trained classifier, the tracking process searches for the highest response area, and as the video frame progresses, the classifier is updated to deal with the deformation, occlusion, illumination, and other issues of the target object.

2.1 Target forecast

For the newly input image block z , calculate its convolution feature ω_l and the correlation output f_l of the classifier:

$$f_l = F^{-1}(W_l * \overline{Z}_l) \quad (5)$$

Where F denotes the inverse Fourier transform. According to the weights of the layers determined in the previous section, the final filter response output f is:

$$f = \sum_l \alpha_l f_l \quad (6)$$

The location with the highest response value is the target location.

2.2 Model update strategy

The algorithm belongs to the discriminant online tracking model. Firstly, the tracking model is trained by using the annotation information given by the first frame of the video, and then the model is updated according to the changes of the target and the background in the subsequent frames, which has strong flexibility and Adaptability.

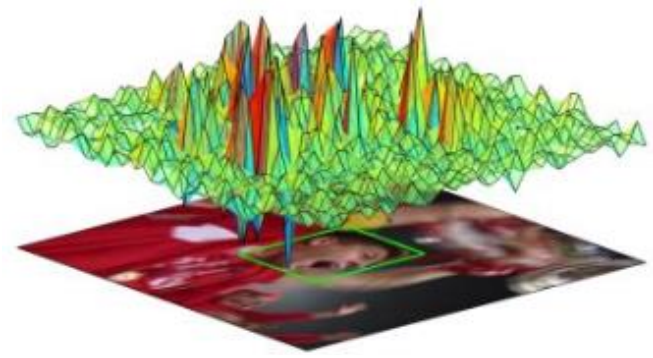

Figure 3 Schematic diagram of response values

Table 2. Model update experimental data

Frame interval	Average DP	Speed
1	0.793	66.4fps
2	0.789	66.9fps
3	0.782	67.3fps
4	0.779	67.7fps
5	0.772	68.0fps
6	0.768	68.4fps

At present, the application of more update strategies is mainly for each frame to update, more parameters are updated, the

algorithm is lagging in real time, and the model update every frame also easily leads to target drift. In order to improve the real-time performance of the algorithm, this paper uses a sparse update strategy, which is to select a moderate, low-frequency model and a more detailed strategy. This paper selects different frame intervals to update the model, and analyzes the tracking effect at different intervals. After updating the model every 5 frames, the tracking speed is improved and the tracking accuracy is not affected.

3. Results and Analysis

This experiment platform uses Windows10 64-bit operating system, I9-9900K CPU, 16GB DDR4 3000MHz memory, RTX2080Ti graphics card, using Tensorflow depth framework, python programming language, test data set is OTB-100, use Bolt2 data to test target scale change interference, select Soccer test occlusion interference, using the Football data to test similar object interference.

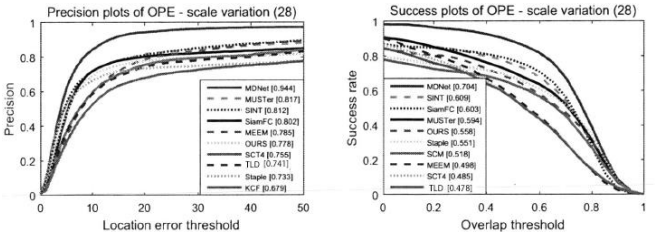

Figure 4.Scale change

The data set uses the one-pass-evaluation (OPE) evaluation method to determine the accuracy rate and success rate of different visual tracking algorithms under the disturbance factors such as scale change (SV), occlusion (OCC), and deformation (DEF). Rate) for quantitative analysis. The precision plot refers to the ratio of the number of frames in the video sequence where the tracker's predicted position and the given label position are less than a given threshold to the total number of frames, and 20 pixels are selected as the threshold. Success rate plot (sucessplot) refers to plotting the success rate of each frame of a video sequence based on the coverage of the tracker's predicted position and label position. The coverage score [14] formula is:

$$S = \frac{|\Upsilon_t \cap \Upsilon_a|}{|\Upsilon_t \cap \Upsilon_a|} \qquad (5)$$

Where Υ_t, Υ_a represent the predicted position and the real tag position information, respectively. $|\bullet|$ represents the number of pixels in the area.

Based on this, three sets of experiments were carried out to verify the effectiveness of this paper in SV, OCC and DEF, and compared with mainstream algorithms. It can be seen from the above experiment. In this paper, we have achieved excellent results compared with the original KCF algorithm in dealing with scale deformation, occlusion and similar object interference. Compared with high-performance algorithms such as MDNet, the gap is relatively small, but thanks to the KCF algorithm, the processing speed is very fast, and the algorithm is real-time. It is much better than the 1fps of algorithms such as MDNet.

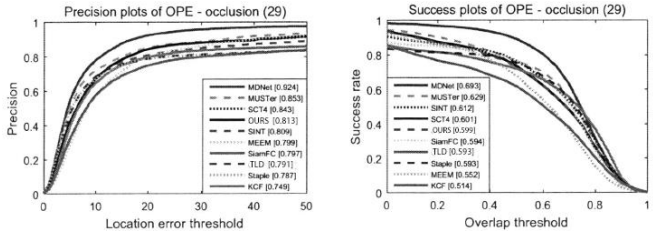

Figure 5.Occlusion interference

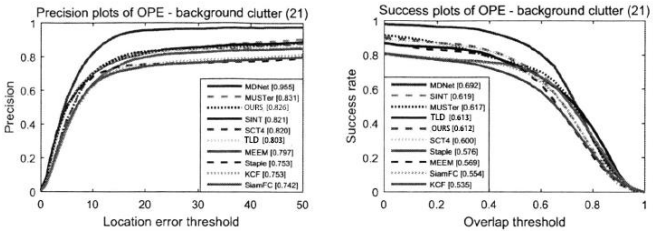

Figure6. Similar interference

Figure 7. occlusion test

The 79 frames, 135 frames, and 155 frames of the Soccer video of the OTB-100 data set are selected. The actual observation algorithm is anti-jamming to the occlusion factor. The result is as follows:

4. Conclusion

Based on the powerful feature extraction ability of VGG-16 neural network, this paper extracts different layer features and combines the classifiers used to train KCF algorithm, which can improve the KCF algorithm to achieve better tracking performance in occlusion, deformation and scale change. The equal-interval frame model update strategy is adopted to make the algorithm further improve the real-time performance under the premise of guaranteeing the success rate. However, the algorithm is sensitive to resolution. When the resolution is reduced, the performance of the algorithm will drop sharply. At the same time, the algorithm cannot improve the tracking failure of noises such as fast motion, lighting drastic changes and motion blur.

References

[1] Wu Y, Lim J, Yang M H. Object Tracking Benchmark[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9):1834-1848.

[2] Yang L, Wu T, Zhu S C. Online Object Tracking, Learning, and Parsing with And-Or Graphs[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2014.

[3] Henriques J F , Caseiro R , Martins P , et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3):583-596.

[4] Bolme D S , Beveridge J R , Draper B A , et al. Visual object tracking using adaptive correlation filters[C]// The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010. IEEE, 2010.

[5] Danelljan M , Khan F S , Felsberg M , et al. Adaptive Color Attributes for Real-Time Visual Tracking[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014.

[6] Li Y , Zhu J . A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration[J]. 2014.

[7] Danelljan M, Hager G, Khan F S, et al. Discriminative Scale Space Tracking[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(8):1561-1575.

[8] Danelljan M , Häger, Gustav, Khan F S , et al. Learning Spatially Regularized Correlation Filters for Visual Tracking[J]. 2016.

[9] Galoogahi H K , Sim T , Lucey S . Correlation Filters with Limited Boundaries[J]. 2014.

[10] Danelljan M , Robinson A , Khan F S , et al. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking[C]// European Conference on Computer Vision. Springer International Publishing, 2016.

In the reference list, after numbering the bracket, each reference will start with an indentation of 0.375".

[11] Wang L , Ouyang W , Wang X , et al. Visual Tracking with Fully Convolutional Networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2016.

[12] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2016.

[13] Changzhen X , Manqiang C , Runling W , et al. Real-time visual tracking algorithm based on correlation filters and sparse convolutional features[J]. Journal of Computer Applications, 2018.

[14] Wu Y , Lim J , Yang M H . Online Object Tracking: A Benchmark[C]// 2013 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2013.