Research on face tracking Algorithm Based on Detection and Supervision Tracking

Gang Yuan¹, Fengkun Lv¹, Zhihao Xu¹*, Lv Luo¹, Yi Rong¹, Yanan Zhang¹, Zheng Tian¹

¹ School of Mechanical and Electrical Engineering, Nanchang Institute of Technology Nanchang, China

*Corresponding author’s e-mail: zhxuhi@whu.edu.cn

Abstract. Abstract: Accurate face detection and tracking is widely used in many social life scenes. However, the uncontrollable background noise and random illumination change in the application scene will reduce the detection accuracy of the tracked target, and the rotation, occlusion and overlap of the tracked target will also affect the accuracy and success rate of the tracking algorithm. In order to solve the above problems, this paper proposes a method that uses deep learning detection method to supervise the correlation filter tracking algorithm to improve the success rate of de-tection and tracking. First of all, in the first frame of the picture, we use the deep learning SSD (Single Shot MultiBox Detector) algorithm to detect the face, and take the detected face as the tracking target, and use the correlation filtering algorithm DSST (Discriminative Scale Space Tracker) Tracking in the process of tracking, face detection is continuously carried out on the tracking target, and the detection results are used to monitor the tracking results, so as to reduce the target drift caused by the boundary effect, thus improving the accuracy of the tracking algorithm. The algorithm is tested and verified on OTB100 data set. The final experimental results show that the accuracy of this algorithm is obviously better than the mainstream classical algorithm, and the frame rate meets the real-time re-quirements.

1. Introduction

Face tracking is a classic problem in computer vision, which is applied in many fields, such as human-computer interaction, intelligent video surveillance and autonomous driving, etc. In recent years, face tracking has been very popular in the field of computer vision. Face tracking is to predict the position and scale of a given face target in the following frames. Generally speaking, the tracking framework consists of four modules, which are feature extraction, appearance modeling, motion estimation and target location. Feature extraction is the process of mapping the k-dimensional feature subset of the target to a new feature set. It can also be expressed as the method and process of extracting characteristic information from images by using computers. Appearance modeling is a process of detecting target features and constructing a mathematical model with the ability to discriminate. Motion estimation is the process of inferring and tracking the position of the target in subsequent frames. At this time, the predicted position may be multiple. Target location is the process of determining target location by non-maximum suppression algorithm or greedy algorithm. Face tracking is faced with several difficulties as well as target tracking, such as the deformation of target appearance, environmental lighting, motion blur and jitter, high speed movement and background interference. The commonly used face tracking methods include generation model method and discriminant model method. The generation model method is to use the current frame to model the target region, and the next frame will look for the region
with the largest model response according to the model. Common methods include Kalman filter algorithm[1], particle filter algorithm[2] and means-shift algorithm[3]. The discriminant model method is mainly to extract image features and then perform machine learning on the features. Common methods include TLD algorithm[4] and Struct algorithm[5] based on online learning, KCF[6] and DSST algorithm based on correlation filtering, and SiamFC algorithm[7] based on the deep learning.

Correlation filtering algorithms have incomparable advantages in speed, especially MOSSE (Visual Object Tracking using Adaptive Cor-relation Filters) algorithm[8]. In the OTB50 data set, the FPS of MOSSE algorithm reaches 615, but the average accuracy only reaches 43.1%[9]. This is the first time that correlation filtering method has been applied in the field of target tracking, showing the amazing potential of correlation filtering algorithm in response speed. MOSSE is a correlation filtering based on single channel gray features. On the basis of MOSSE, Henriques Joao F et al proposed CSK (Circulant Structure of Tracking-by-detection with Kernel) algorithm[10] and KCF ((high-speed Tracking with Kernelized Correlation Filters) algorithm, based on MOSSE, adds dense sampling, cyclic displacement and closed solution of computational kernel function. KFC algorithm adds HOG feature of multi-channel gradient on the basis of CSK algorithm. Although this reduces THE FPS of KCF to 292, the average accuracy reaches 72.8%, which is higher than most tracking algorithms at that time. Martin Danelljan et al extended CSK with multi-channel color features to obtain the adaptive Color Attributes for real-time Visual Tracking (CN) algorithm[11]. Yang Li from Zhejiang University proposed a SCALe-Adaptive and Multi Feature Integration Tracker (SAMF) algorithm based on KCF, which used HOG+CN Feature to search targets with position filters on multi-scale images. The position at the maximum response output and its scale are taken as the target of the current frame[12]. SAMF can be regarded as the first Scale adaptive target Tracking algorithm, but Martin Danelljan proposed DSST (Discriminative Scale Space Tracking) algorithm to carry forward the target Tracking algorithm, DSST algorithm only uses HOG feature. DCF is used for translational filtering, and scale filter is introduced to build 33 layers of pyramid feature vector to respond to the target scale. Although correlation filtering can solve the problem of scale change of targets and improve the accuracy compared with traditional methods, it is not very ideal when tracking fast targets and fast deformation targets.

Firstly, the Deep Learning Tracker (DLT) algorithm proposed by Naiyan Wang et al.[13] combines Deep Learning with target tracking. This algorithm first uses the stacked denoting autoload, SDAE) to obtain the object representation ability, and then the sigmoid part was trained online to get the classifier, and then the particle filter was used to extract a batch of candidate regions, and the candidate regions were input into the classification network to get the prediction target with the highest confidence. With the development of deep learning, MA et al proposed the HCFT (Hierarchical Convolutional Features for Visual Tracking) algorithm[14], which uses VGGNET-19 as the feature extraction device. After the feature extractor, three feature layers are obtained. After training the three feature graphs, responses are obtained through filters. Finally, the three responses are added with different weights to obtain the final response graph. Bertinetto et al proposed the target tracking method SiamFC (Fully Convolutional Siamese Networks)[15], which uses the network model of offline training to obtain a faster speed, but the positioning accuracy is lower. Although most deep learn-based algorithms improve the tracking accuracy compared with the correlation filtering method, most of them fail to meet the requirement of real time in terms of speed[16].

In the method of correlation filtering, DSST algorithm can obtain fast and relatively accurate effect, but when the fast movement of the target produces boundary effect, it will lead to the failure of tracking. On the basis of literature[17], this paper firstly extracts the multidimensional HOG feature of the target region and retains more target information. By introducing the weight coefficient matrix, it makes a centralized response to the central position of the target region and obtains the coarse positioning of the target. At the same time, the image is input into the SSD network to obtain the supervised position of the target, and the target template of the filter is updated by comparing the Euclidean distance between the coarse positioning position and the supervised position and the threshold value. The algorithm in this paper refers to the long time tracking algorithm and the tracking algorithm based on deep learning, and adds a neural network detector, which reduces the tracking failure and drift.
caused by the boundary effect, and realizes accurate positioning. Experimental results show that the proposed algorithm can significantly improve the tracking accuracy on the basis of real-time performance.

2. Detection and monitoring tracking algorithm
The algorithm idea in this paper is Detection and supervised Tracking, namely Detection SU-Pervise Tracking, which is abbreviated as DST algorithm in this paper. First of all, the input image sequence is judged, if it is the first frame, directly into the detection module to detect the face, the detected face as the target to create a filter. If it is not the first frame, it is filtered and detected. In the filtering part, the HOG feature of the frame is extracted, and the feature input filter is filtered to get the response of position and scale, and then the non-maximum suppression algorithm is applied to the response to get the optimal solution of position and scale. Detection part to an initialization of the input image, adjust its 300 x300x3 three-dimensional matrix, according to the model, in turn, to extract the image feature of the network structure, the characteristics of the layer to test for the input, get more target detection box, the use of the maximum suppression algorithm to filter results, get the optimal output.

After the above process to get the detection and tracking box, calculate the center of the center and tracking test box, both subtraction, get the relative error of the detection and tracking, the error compared with the adaptive threshold, if it is greater than the threshold, will detect the target image as the input updating filter, namely to detect the target as a new target tracking. If the value is lower than the threshold, the tracking result is accurate and is displayed. The detector uses the relatively advanced SSD algorithm, and the tracker uses the classical correlation filter tracking algorithm DSST algorithm.

3. Face detection based on SSD

3.1. Formatting the title
The target detection algorithm of SSD (Single Shot MultiBox Detector)[18] was proposed by Wei Liu et al. VGG was used as the basic network. The full connection layer of VGG was changed into the convolution layer and several additional convolution layers were added.

SSD network structure: (1) The network structure of SSD300 is shown in Figure 1. The original input image is scaled to a fixed size as the input of the convolutional network model, whose base network uses VGG-16. (2) SSD to VGG16 FC6 layer, FC7 layer full connection mode to convolution layer, and then continue to add a number of convolution layer to Conv10_2, the last full connection layer. (3) Save the feature map obtained after convolution. SSD adopts the feature pyramid structure for regression detection of objects, that is, feature maps generated by multiple different convolution layers are used in detection. (4) Extract the feature layer as the input of the detection layer. Generate multiple Prior boxes...
on feature maps of different sizes, and perform classification and order regression on the features generated by these feature maps. (5) Non-maximum suppression algorithm is used to filter the final results.

### 3.2. SSD network loss function

The loss function of SSD is defined as the weighted sum of position error and confidence error:

\[
L(x, c, l, g) = \frac{1}{N} (L_{\text{conf}}(x, c) + \alpha L_{\text{loc}}(x, l, g))
\]

Where \( N \) is the number of positive samples of the prior box, \( L_{\text{conf}} \) is the confidence error, and \( L_{\text{loc}} \) is the position error. For position error, Smooth L1 loss is adopted, which is defined as follows:

\[
L_{\text{loc}}(x, l, g) = \sum_{k} x_{k}^{\text{smooth}} (l_{k} - g_{k}^{n})
\]

\[
\text{smooth}_{L1}(x) = \begin{cases} 
  0.5x^2 & \text{if } |x| < 1 \\
  |x| - 0.5 & \text{otherwise}
\end{cases}
\]

For confidence error, surtax Loss is adopted:

\[
L_{\text{conf}}(x, c) = -\sum_{i} \sum_{p} \log(c_{p}^{i}) - \sum_{i} \log(c_{p}^{i})
\]

Here, \( X_{p} \in \{1, 0\} \) (this symbol should correspond to Formula 1) is an indicator parameter. When \( X_{p} = 1 \), it means that the in prior box matches the JTH real target, and the category of the real target is \( p \). \( c \) is the predicted value of class confidence, \( L \) is the predicted value of the corresponding boundary box of the prior box, and \( g \) is the positional parameter of the real target[1]. Due to the existence of \( X_{p} \), the position error is calculated only for positive samples. The weight factor \( \alpha \) was set to 1 by crossing validation.

### 4. Face tracking based on DSST

DSST algorithm is an efficient correlation filter tracking algorithm, including position filter and scale filter. The position filter can locate the target of the current frame, and the scale filter can estimate the target scale of the current frame. The two filters are relatively independent. So different feature types and eigenvalue calculation methods can be selected for the design.

#### 4.1. Position filter

Assuming the size of the image block \( P \) where the target is located is \( N \times N \), the feature of the target \( P \) is extracted to obtain the feature \( f \) with the size of \( N \times N \times d \), where the dimension of the feature is \( d \) dimension. According to the Gaussian function, the response \( g \) is constructed with the size of \( N \times N \), and the intermediate response value of the response is the largest, which decreases successively in all directions.

Take the two-dimensional Fourier transform of the features of each dimension of \( f \) to get \( F \), and take the two-dimensional Fourier transform of \( g \) to get \( G \). Set the regular term as \( \gamma \), and take the minimum value of \( \varepsilon \), then the optimal filter \( H \) should satisfy the following formula:

\[
\varepsilon = \left\| \sum_{l=1}^{d} h^{l} \star f^{l} - g \right\|^2 + \lambda \sum_{l=1}^{d} \| h^{l} \|^2
\]

Where the asterisk represents cyclic correlation, \( h^{l} \) and \( f^{l} \) are the \( l \)-th dimension of \( f \) and \( l \) ranges from 1 to \( d \). Minimize the above equation to obtain the formula (6), where the overline represents the complex conjugate.

\[
H^{l} = \frac{\overline{G}F^{l}}{\sum_{k=1}^{d} F^{k} \overline{F}^{k} + \lambda}
\]
Divide $H$ into numerator $A$ and denominator $B$, and carry out iterative update respectively. The update method is as follows\[2\]:

$$A_i = (1 - \eta)A_{i-1} + \eta\hat{G}_iF_i, \quad B_i = (1 - \eta)B_{i-1} + \eta\sum_{k=1}^{d}F^k_i$$

(7)

Where, $\eta$ is the learning rate, and the new image, its feature is $z$. Take the two-dimensional Fourier transform of each dimension to get $Z$. The corresponding response $y$ can be obtained according to the following formula:

$$y = \mathcal{F}^{-1}\left\{\frac{\sum_{i=1}^{d}A_iZ^i}{B + \lambda}\right\}$$

(8)

The position of the maximum value of $y$ is the position of the new target center.

4.2. Scaling filter

After the position of the target is obtained through the position filter, the scale information is determined by the one-dimensional scale filter. The target sample scale selection principle for scale evaluation is as follows:

$$a^nP \times d^nR \quad n \in \left\{\frac{S - 1}{2}, \ldots, \frac{S - 1}{2}\right\}$$

(9)

Where, $P$ and $R$ are the width and height of the target in the previous frame, $a = 1.02$ is the scale factor, and $S = 33$ is the total scale series\[2\]

Since the scale factor is not a linear function, the growth of each scale is not the same. For the scale larger than the current scale, a larger stride length is adopted, and for the scale smaller than the current scale, a smaller step length is adopted, so as to achieve coarse detection of large targets and fine detection of small targets, so as to achieve higher accuracy.

Each sample was normalized to a unified fixed size, and 31-dimensional flow features were extracted respectively. All frog features of each sample were connected in series into a feature vector to form a 33-layer pyramid feature, which was multiplied by an one-dimensional hand window as input $Z$.

The corresponding output response $y$ can be obtained by substituting the input $Z$ into the equation below:

$$y = \mathcal{F}^{-1}\left\{\frac{\sum_{i=1}^{d}A_iZ^i}{B + \lambda}\right\}$$

(10)

The scale of the maximum value of $y$ is the final scale estimation result of the current frame.

5. The experimental results

The experimental environment in this paper is Ubuntu18.04, the CPU is Intel Core I5 7300H, the GPU is NVIDIA GTX 1050, the regularization parameter is set as 0.01, the scale factor $A = 1.02$, and the total weight scale coefficient $= 0.75$. In this paper, DST algorithm is compared with TLD algorithm, CT algorithm, CXT algorithm, MS-V algorithm, KCF algorithm and DSST algorithm. The video used in the experiment is selected from the FaceOcc2 video of OTB100 data set, including the background environment of lighting change, occlusion, rotation inside and outside the plane, etc.

This paper compares the above seven algorithms from three aspects: success rate, accuracy and frame rate. Where, the Euclidean distance between the center position of the target box output by the tracking algorithm and the center position of the actual target is compared with the threshold value, and the ratio of the number of frames less than the threshold value to the total number of frames of the video is the accuracy of the video at this distance. Overlap degree between the output box of the tracking algorithm and the actual box is calculated, and the overlap degree is compared with the threshold value. The ratio
between the number of frames greater than the threshold value and the total number of frames is called the success rate of the video under the overlap rate.

Figure 2 for face recognition and tracking test under different condition video capture, figure 3 to test the success rate of video figure, figure 4 to test the precision of the video, the red represents the average test results of the algorithm in this paper the success rate and the average accuracy of algorithm is superior to the rest to table 1 for seven kinds of algorithm is the average success rate, average precision and frame rate of the comparison table.

| The test algorithm | Average success rate | The average accuracy |
|--------------------|----------------------|---------------------|
| DST                | 0.633                | 0.871               |
| CXT                | 0.649                | 0.818               |
| KCF                | 0.647                | 0.76                |
| TLD                | 0.564                | 0.67                |
| CT                 | 0.487                | 0.5                 |
| MS-V               | 0.184                | 0.072               |
| DSST               | 0.661                | 0.77                |

6. conclusion
DSST algorithm is presented in this paper with the SSD algorithm by SSD detection algorithm is proposed on the basis of supervision DSST update and tracking algorithm based on monitoring results of face tracking algorithm, namely the inspection supervision and tracking algorithm, experiments show that the presented algorithm in serious background interference and object shelter and deformation under the condition of real-time tracking can still be good, The accuracy of the algorithm is improved compared with other algorithms.
As the detection and monitoring tracking algorithm adopt deep learning SSD algorithm as the detection part, the calculation amount of the algorithm is improved. The frame rate is lower than DSST, but still higher than CXT algorithm, and meets the real-time requirements. The follow-up research work can focus on solving the problem of speed decrease caused by large amount of computation, such as introducing faster target detection algorithms such as YOLO, etc., or focusing on using depth features to replace manual features, etc., so that the improved algorithm can take into account the requirements of accuracy, success rate and speed.

References

[1] Kalman R E. A New Approach to Linear Filtering and Prediction Problems[J]. Journal of Fluids Engineering, 1959, 82D:35-45.
[2] Numbers K,Koller-Meier E,Gool L V.An adaptive color-based particle filter[J].Image and Vision Computing.2003,21(1):99-110.
[3] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” IEEE T. PAMI, vol. 24, no. 5, pp. 603-619, May 2002
[4] Ali Z, Mikolajczyk K, Matas J. Tracking-Learning-De- section[J]. IEEE Transactions on Software Engineering, 2011, 34(7):1409-1422.
[5] S. Hare, A. Safari,and P. H. S. Torr, “Struck: Structured output tracking with kernels,” in Proc. IEEE Int. Conf. Comput. Vis., 2011, pp. 263–270
[6] Henriques João F,Caseiro Rui,Martins Pedro,Batista Jorge. High-Speed Tracking with Kernelized Correlation Filters.[J]. IEEE transactions on pattern analysis and machine intelligence,2015,37(3).
[7] Bertinetto L, Valmadre J, Henriques J F, et al. Fully- convolutional siamese networks for object tracking[C]// European Conference on Computer Vision (ECCV), October, 8-16, 2016, Amsterdam, the Netherlands, 2016. 9914:850-865.
[8] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters. Computer Vision & Pattern Recognition, vol. 119, no. 5, pp. 2544-2550, 2010.
[9] Wu Y, Lim J, Yang M H. Online objects tracking: A benchmark [C]// CVPR, 2013.
[10] Henriques J F, Caseiro R, Martins P, et al. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels[C]// Proceedings of the 12th European conference on Computer Vision - Volume Part IV. Springer, Berlin, Heidelberg, 2012.
[11] Danelljan M,Khan F S,Felsberg M,et al.Adaptive color attributes for real-time visual tracking[C] // Proc of 2014IEEE Conference on Computer Vision and Pattern Recognition,2014:1090-1097.
[12] Li Y, Zhu J. A Scale Adaptive Kernel Correlation Filter Tracker of Feature Integration[C]// European Conference on Computer Vision. Springer, Cham, 2014.
[13] Wang N, Yeung D Y. Learning a deep compact image representation for visual tracking[J]. Advances in Neural Information Processing Systems, 2013.
[14] MA C,HUANG J B,YANG X K,et al. Hierarchical convolutional features for visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015:3074-3082.
[15] Bertinetto L, Valmadre J, Henriques J F, et al. Fully- convolutional siamese networks for object tracking[C]// European Conference on Computer Vision (ECCV), October, 8-16, 2016, Amsterdam, the Netherlands, 2016. 9914:850-865.
[16] Li Xi,Zha Yufei,Zhang Tianzhu,et al. Survey of visual objects tracking algorithms based on deep learning[J]. Journal of Image and Graphics2019,24(12):2057-2080.
[17] Danelljan M,Hager G,Khan F S,et al.Discriminative scale space tracking[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(8):1561-1575.
[18] LIU W,ANGUELOV D,ERHAN D,et al.SSD:Single Shot Multibox Detector[C]// European Conference on Computer Vision.Springer,Cham:Springer,2016:21-37.