Neural Network-Based Video Tracking Techniques

P.Sriramulu[1], Amaresh[2], Dr.B.Ratnakanth[3]

[1] Assistant Professor, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

[2]Assistant Professor, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

[3]Professor & HOD, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

Abstract: ***Target tracking is an important research direction in computer vision. In order to realize target tracking, this paper proposes a target tracking algorithm based on deep neural network. The basic framework of the algorithm consists of a Siamese network. The main innovation is that the attention mechanism module is added to the Siamese network model. The attention mechanism module can make the extracted features more refined, so that the tracking effect is more accurate. The model proposed in this paper achieves excellent tracking results on both OTB2015 and UAV123 test datasets.***

Keywords: Target tracking, Deep neural network, Siamese network, Attention mechanism

1. Introduction

In the field of computer vision, video tracking has very important research value. And single target tracking is one of the important branches of video tracking. The concept is that, given the target of interest in the first frame of the video, a rectangular frame is used to mark the target from the video in the subsequent video sequence to complete the tracking. At present, single-target tracking plays an irreplaceable role in smart security, modern military and civilian fields.

Currently, there are roughly two algorithms for target tracking. One class is the algorithm based on correlation filters. Correlation filter theory was first applied to the field of signal processing and was later introduced to the field of video tracking. The tracking algorithm based on correlation filter has the advantages of high accuracy [5] and fast tracking speed [6]. Among them, the Minimum Output Sum of Squared Error filter (MOSSE, Minimum Output Sum of Squared Error filter) algorithm should be the first to be proposed, and then the target tracking algorithm based on Circulant Structure with Kernels (CSK, Circulant Structure with Kernels) appeared one after another., Target tracking algorithm based on Kernelized Correlation Filter (KCF, Kernelized Correlation Filter), tracking algorithm based on discriminative scale space (DSST, Discriminative Scale Space Tracker), spatial regularization discriminative correlation filter tracking algorithm (SRDCF, Spatially Regularized Discriminative Correlation Filters) and so on.

Later, with the continuous development of deep learning technology, another tracking algorithm, the tracking algorithm based on the twin network, was born. In 2016, fully-Convolutional Siamese Networks for Object Tracking (SiamFC, Fully-Convolutional Siamese Networks for Object Tracking) was published. This paper put the Siamese network on the video tracking craze. The tracker proposed in the paper converts the video tracking into a similarity measure between the template of the tracked object and the candidate region in the frame, and calculates the measured response value. The position with the largest response value is the target to be tracked. After that, a series of tracking algorithms were derived from SiamFC, such as target tracking based on the region proposal network (High Performance Visual Tracking with Siamese Region Proposal Network, SiamRPN). This paper introduces a region proposal network (RPN) on the basis of SiamFC, where the RPN is composed of a classification network branch and an anchor box regression branch. First, the foreground and background of the proposed region are classified, and then the pre-defined anchor box is used. (anchor) for processing to get an accurate bounding box. Since then, the paper Deeper and Wider Siamese Networks for Real-time Visual Tracking (SiamDW) replaces the backbone network used to extract features in SiamFC and SiamRPN with a center cropped one. Resnet.

This approach further improves the accuracy and robustness of tracking while sacrificing a certain tracking speed. The attention mechanism is a common structure used in convolutional neural networks. This mechanism can extract more refined features by focusing the attention of the convolution kernel on a specific position, so that the

extracted information has more semantics, if we can use this mechanism in the field of video tracking, then the effect of video tracking must be greatly improved.

2. Algorithm Description

The following paper will describe the algorithm in detail. As mentioned above, the algorithm is divided into four modules: backbone network module, attention mechanism module, feature fusion module, and cross-correlation module.

This paper uses a fully convolutional network to build a Siamese network framework for tracking objects. The framework consists of two branches: the target branch takes the tracking template Z as input, and the search branch takes the search area X as input. These two branches share the same CNN network as the backbone network for feature extraction. The feature maps output by CNN are ϕ (Z) and ϕ (X). The next thing to do is to estimate the similarity of the feature maps to get the response map.

2.1 Backbone Network

The first step in building a Siamese network framework is to select a suitable backbone network for feature extraction. SiamFC uses Alexnet as the backbone network. Alexnet is characterized by translation invariance, which can accurately embed images into the feature space without offset. However, its network structure is relatively simple, so it is difficult to extract deeper features of the image using Alexnet.

SiamRPN++ uses Resnet as the backbone network, which can extract deeper features. However, because Resnet uses Padding, it does not fully meet the translation invariance, so it will cause certain errors to the features. This paper adopts some of the views in Deeper and Wider Siamese Networks for Real-Time Visual Tracking, and adds a central cropping module after the last layer of Resnet to crop the features and remove the most boundary part of the feature.

After the derivation of the above paper, the Resnet after center clipping is translation invariant. From this, we determine the composition of the backbone network, that is, the Resnet after center cropping.

2.2 Attention Mechanism

Combining the attention mechanism with the Siamese network is a major innovation of this paper. Using this feature can make the features extracted by Resnet more refined, so as to obtain richer semantics and improve the tracking effect.

Combining the attention mechanism with the Siamese network is a major innovation of this paper. Using this feature can make the features extracted by Resnet more refined, so as to obtain richer semantics and improve the tracking effect. The attention mechanism module is composed of a series of deconvolutions, whose role is to upsample the features. Through upsampling to achieve the purpose of amplifying the features. The enlarged features can make the model focus on one place, so that the features are more refined and contain more semantics. That is to say, the features extracted by the backbone network will be sent to the attention mechanism module for further amplification. In this paper, the features are amplified twice, with a magnification of 1.5 and 2 times respectively. After the attention mechanism module, we will get two sets of features with magnifications of 1.5 and 2 times on both the template side and the search side.

2.3 Feature fusion

The attention mechanism module can output two sets of enlarged features, which are combined with the unscaled features in this paper. The feature fusion module refers to the algorithm in SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. In this paper, the author performs a weighted sum of three sets of features of the same size. This method is also suitable for the model in this paper. Through this method, the three sets of features are fused into a set of new features, and the newly obtained features have the same size as the original features of each set. This method does not increase the number of channels and parameters of the features, so it will not affect the connection. down the speed of the cross-correlation operation. See the formula:

$$S = \sum \beta_i * S_i$$

Among them, S is the new feature after fusion, Si is the feature before fusion, and βi is the weight corresponding to each group of features.

2.4 Correlation

The next step is to cross-correlate the tracking template Z with the aggregated features of the search region X. Cross-correlation is a parameter-free convolution operation. The specific algorithm is as follows:

$$R = \varphi(X) \bigstar \varphi(Z)$$

Where ⋆ represents the cross-correlation operation performed on ϕ (Z) and ϕ (X), and R is the single-channel response graph obtained after the cross-correlation operation of ϕ (Z) and ϕ (X).

The response graph obtained by the cross-correlation operation contains the information of the target to be tracked. That is, the most relevant part of the response graph is the position of the target to be tracked. The size of the response graph is 17*17. In this paper, the response graph is subjected to bicubic linear interpolation, and the size after interpolation is 272*272. The size of the input image at the input end is 255*255, so we can roughly restore the coordinates of the target to be tracked on the search branch through interpolation.

2.5 Loss function

In this paper, the logistic loss function is used to train the model, as shown in the formula:

$$l(y,v) = \log(1+\exp(-yv))$$

$$L(y,v) = \frac{1}{|D|}\sum_{u \in D} l(y[u], v[u])$$

Where v is the response value of a certain coordinate output by the model, y is the label, and the value range is {-1, 1}. When y is equal to 1, it is represented as a positive sample, and when y is equal to-1, it is represented as a negative sample. D is the response graph output by the model, u is a specific coordinate in the response graph, and X is the size of the response graph.

3. Experimental Content

This chapter mainly describes the experimental content and the use of experimental datasets. The brief experimental steps can be divided into model training and testing.

3.1 Model training

This paper adopts the strategy of online training and uses the ILSVRC2015 VID dataset for training, which is a dedicated A dataset for computer vision, which is used by many object tracking models for experiments. First, we download the data set compression package that has been preprocessed for the target tracking task on the Internet. The compressed package contains the picture files cut from the video frame by frame and the information files of the targets to be tracked in the pictures. After decompression, we divided the image files into five folders, four of which are used as training set and one is used as validation set. The training epoch is set to 15, the batch is set to 32, the learning rate is gradually decayed from 0.005 to 0.0005, and the warm-up is set in the first three epochs, the learning rate of the first epoch is set to 0.001, and the base is reached in the third epoch. learning rate. The training adopts the stochastic gradient descent algorithm (SGD), and the training of the overall model reaches convergence in the ninth epoch. The pre-training model of the backbone network has been given in the paper SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks.

3.2 Model testing

In this paper, the offline tracking strategy is adopted, and the first frame of the image sequence is always used as the tracking template of the Siamese network. We use the ninth epoch model parameter in the training process as the test model for the test session. The model tracks with 13.5 Fps on the test dataset. This paper uses OTB2015 and UAV123 two sets of test sequences to test on the OPE (one-pass evaluation) benchmark. The specific test content and test results will be described in the next chapter.

4. Experimental Results and Analysis

This chapter mainly describes and analyzes the experimental results. This paper conducts the OPE

Benchmark test on the OTB2015 and UAV123 datasets respectively.

4.1 Results at OTB2015

OTB2015, also known as OTB100, includes 100 video sequences with substantial variation challenges, which were published in Object tracking benchmark [J]. TPAMI, 2015. Paper. We can download it from the Visual Tracker Benchmark website. Compared to OTB50 and OTB2013 on the same website, OTB2015 has twice the number of video sequences, so it is more challenging. The challenges are mainly divided into eleven aspects: illumination change, scale change, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out of view, background clutter, and low resolution. Under this dataset, our algorithm achieves a success rate of 48.5% and an accuracy of 66.4%. And it shows good performance to deal with various challenges of OTB2015. In comparative experiments, it is shown that the tracker in this paper also beats SiamFC in several challenges such as fast movement, scale change, and rotation. But the fly in the ointment is that since this paper does not have the network structure used to improve the two challenges of illumination change and low resolution, the tracker performs relatively general on this challenge.

4.2 Results on UAV123

The UAV123 dataset is mainly composed of 91 drone videos, and individual long video sequences are split, and finally 123 video sequences are formed. These video sequences also contain various classification challenges such as fast motion, huge scale, and occlusion. Although not as famous as OTB2015, UAV123 is also widely used in object tracking due to its challenging authority. Our proposed tracker performs equally well on this dataset, achieving a success rate of 49.0% and an accuracy of 68.8%.

5. Conclusion

In this paper, a video tracking model based on deep neural network is proposed. The model is mainly divided into four modules: backbone network module, attention mechanism module, feature fusion module, and cross-correlation module. The backbone network part is composed of Resnet which is cut by the center. The attention mechanism module adopts the method of upsampling to focus on a certain part of the feature to make the feature more refined. The response graph is obtained after feature fusion and cross-correlation operation, and the tracking is completed. The tracker proposed in this paper achieves excellent success rate and accuracy on both OTB2015 and UAV123 test datasets. At the same time, compared with multiple trackers, the tracker proposed in this paper also shows strong robustness in the challenges of OTB2015 scale change and low resolution, which shows the innovation and advancement of the tracker.