Visual Tracking based on Regression Convolutional Neural Network

Yibo Min[a], Jianwei Ma [b, *], Shaofei Zang [c], Hongyun Zheng [d]

School of Information Engineering, Henan University of Science and Technology Luoyang, China.

[a]yibo_min@163.com,[b] lymjw@163.com, [c]zangshaofei@163.com, [d]1669628873@qq.com

*Corresponding author

Keywords: Object tracking; Deep learning; Convolutional neural network.

Abstract: Deep learning methods have been widely used in target tracking with superior performance, but most target trackers still require a large amount of video training data to ensure the robustness of the network, resulting in low real-time performance. In order to address this problem, we propose a deep feedforward network that does not require online training to learn the general relationship between target motion and appearance. Then, target position can be output through multi-layer regression network. Finally, the tracker shows excellent results on the benchmark data sets.

1. Introduction

As a subject in the field of computer vision, target tracking plays an important role in many fields such as missile guidance, video surveillance, and drone tracking. Although there has been a lot of research, there are still many problems, such as illumination changes, deformation, occlusion and camera movement [1,2,3]. The stat of art trackers generally utilized target features such as texture features, edge features, and HOG features, which need to be manually labeled [4]. When the appearance of the target changes greatly, these features are not well adaptive. In the field of object detection and recognition, CNNs (convolutional neural networks) have been used to extract the target depth features, which were very effective. Therefore, the deep learning strategy was into the tracking field. In general, different level of deep neural networks have different characteristics. Shallow features contain more location information, but the semantic information is not obvious. Deep features contain more semantic information whose anti-interference ability is stronger, but the position information is weakened. Therefore, the combination of different convolutional layers improve performance. However, deep neural networks require a large amount of video and image data to get good performance, and it is difficult to track untrained targets online. Held et al. [5] proposed to use the general target tracking framework of the regression network to search the target of the previous frame and the search area of the current frame through 5 convolutional layers at the same time. The cascaded feature output through the fully connected layer and returns to the position of the current frame target, while realized higher tracking speed. In this work, the target area of the previous frame and exhaustive area in the current frame are convolved through the 13-layer small scale of VGG16 to increase the receptive field, strengthening the nonlinearity and reducing the number of parameters. And the generalization performance is higher, which is suitable for real-time online object tracking.

2. Related Work

The combination of deep learning and correlation filtering is not limited to the application of deep network extraction features. Some algorithms combine feature extraction and classifiers, using neural networks to simulate the entire neural networks to simulate the entire process of correlation filtering [6,7,8]. In the correlation filtering, the template information needs to be saved and the search area features are extracted. Therefore, the network based on the relevant filtering idea generally adopts the Siamese Network structure, in which one branch saves the target template information and the other branch uses the search area. Feature extraction, and finally the two parts of the feature are related operations to obtain a response image. Wang et al. [9] proposed fully convolutional network-based tracking (FCNT), using VGG-16 network. It is proposed that the features of different layers of deep neural network have different characteristics. Shallow features contain more position information, while deep layers contain more semantic information, and there are a lot of redundancy in depth features. Therefore, the algorithm targets the feature maps of the Conv4-3 and Conv5-3 two-layer output, and the training feature selection network extracts the effective features and reduces the feature dimensions. The selected features are then transported to their respective positioning networks for target location. The algorithm complements each other with different layer features to achieve effective suppression of tracker drift and robustness to the shape change of the target itself. The image resolution in image classification is relatively high. The main purpose is to judge the category of the object in the image. The resolution of the image in the target tracking is generally low. The image resolution in target tracking is generally low. As long as the purpose is to locate the target in the image, you may encounter background clutter and the target is too small. Although the network trained in the large-scale image classification competition had good performance in the target tracking task, it is still not fully applicable to the target tracking task. In the target tracking, the types of objects not included in the image classification task may appear. As the network deepens, the weakening of the target location information is not conducive to the target positioning. In order to extend the capabilities of CNN in the field of target tracking, a large amount of training data is required. Nam et al. [10] proposed multi-domain Network (MD-Net), which uses the model VGG-M pre-trained in the image classification task as the network initialization model, followed by multiple fully connected layers for the classifier. During tracking, the fully connected layer during training is removed, and a fully connected layer is initialized using the first frame sample. The new fully connected layer continues to fine tune during the tracking process to accommodate new target changes. But this tracking strategy is very time consuming and it is difficult to meet real-time requirements.

3. Method

3.1 Convolutional Neural Network

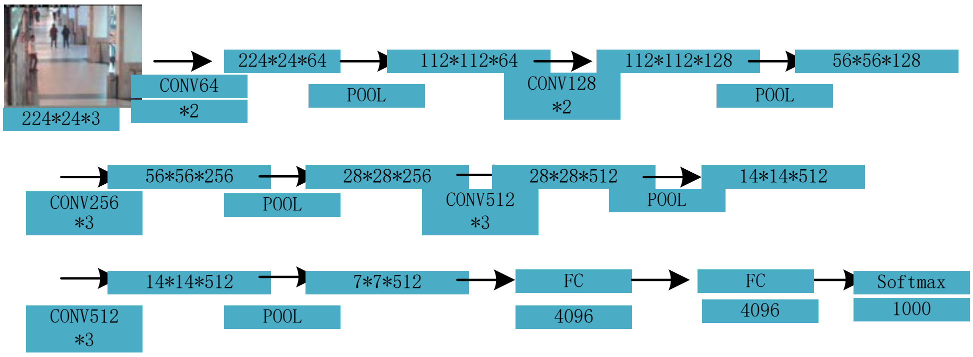

Figure 1. VGG16 network architecture

The VGG16 neural network is a version of VGG-net [11,12], which is a 16-layer deep convolutional neural network constructed by continuously stacking 2*2 maximum pooling layers and 3*3 small convolution kernels. The VGG16 neural network has a total of 5 convolutional segments, 13 convolutional layers, and each convolution segment contains 2-3 convolutional layers. In order to achieve the purpose of reducing the image size, one end of each convolution segment is connected the largest pooling layer. And the number of convolution kernels per segment is the same, the more the number of segment convolution kernels near the end of the network: 64-128-256-512 There will be 3*3 stacks in this structure. Together, this has the advantage of increasing the receptive field, increasing nonlinearity and reducing the number of parameters. The network is mainly generalized and has good performance and easy to migrate to other image recognition projects. One of the great advantages of the VGG16 network is that it simplifies the structure of the neural network.

The network structure of the VGG16 is very regular and not complicated. Several convolutional layers are followed by a pooling layer that can compress the image size, and the pooling layer can reduce the image height and width.

3.2 Algorithm Framework

The framework of the whole algorithm is as shown in the figure 2. We combine the target of the previous frame and the search area of the current frame through the Conv Layers of CNN, and then obtain the output of the convolution layer through Fully-Connected Layers, used to regress the position of the current frame target. The entire framework can be divided into three parts:

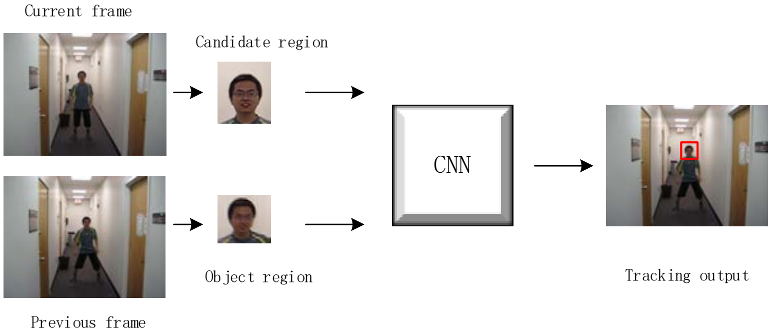

Figure 2. Tracking framework

1) Convolution layer for extracting features of the target area and the search area;

2) Fully connected layer, used as a regression algorithm for comparing target features and search area features;

3) Output new target location

For single-target follow-up, we define a generic image alignment follow-up framework, as shown in the figure above. In this model, we enter the detection target and search range data into the convolutional layer sequence. The output of these convolutional layers represents the representation of the image in high dimensional features. The output of these convolutional layers is then input into some of the fully connected layers. The role of the fully connected layer is to compare the features from the target image block with the features of the current frame to find the specific moving distance to follow the target. Between two frames, the target may have undergone transformation, rotation, ray transformation, occlusion, or deformation. The function learned through the fully connected layer is undoubtedly a complex feature comparator. This comparator can learn a large number of examples to cope with various factors and simultaneously output the motion relationship of the target object. We concatenate the output of the 13 convolutional layers into a vector. This vector is then entered the fully connected layer. Finally, we connect the output of the last fully linked layer to an output layer that contains 1000 nodes representing the coordinate relationship of the output rectangle. Divide the output data by 10 and select it by the validation set (the same processing for all parameters). The network parameters are obtained from the default settings of VGG-net, and we use dropout and ReLU as nonlinear between each fully connected layer [13,14].

4. Implement Details

Real-world objects have a smoothing effect in spatial motion, giving ambiguous blurred image of a target object. A tracker can predict the position of the target tracking object near the location saved at the previous moment. This is especially important in videos with similar objects, such as multiple fruits of the same kind. Therefore, we hope to teach the neural network that, in other cases, the small movement takes precedence over the big movement.

We define the relationship between the center coordinates as $\left(c'_x, c'_y\right)$ of the current frame target object rectangle and the center coordinates as $\left(c_x, c_y\right)$ of the previous frame target object:

$$c'_x = c_x + w \cdot \Delta x \tag{1}$$

$$c_y' = c_y + h \cdot \Delta y \tag{2}$$

Where w and h are the width and height of the calibration frame of the previous frame, respectively. The parameters Δx and Δy are a random number used to obtain the amount of change in the center position of the target object of the previous frame. In the training phase, we find that the variables of the target object position, such as Δx and Δy, can be modeled as a Laplace distribution and a mean of zero. Such a distribution model has a much greater probability of appearing on small movements than large movements.

The scale changes are as follows:

$$w' = w \cdot \mu_w \tag{3}$$

$$h' = h \cdot \mu_h \tag{4}$$

Where w' and h' are the framed width and height of the target object of the current frame, w and h are the width and height framed by the target object of the previous frame, and the two parameters represent the frame size change. We found in training that these two parameters can be modeled as a Laplace distribution and a mean of 1. Such a distribution is more likely to keep the calibration frame size the same as the previous frame.

5. Experiment

We use the five target tracking algorithms that are outstanding on the OTB100 dataset: DLT [15], KCF [16], CF [17], CT [18], comparative experiments were performed on 100 test video sequences. We use one-time evaluation (OPE) accuracy curve and success rate graph to judge the performance of the algorithm.

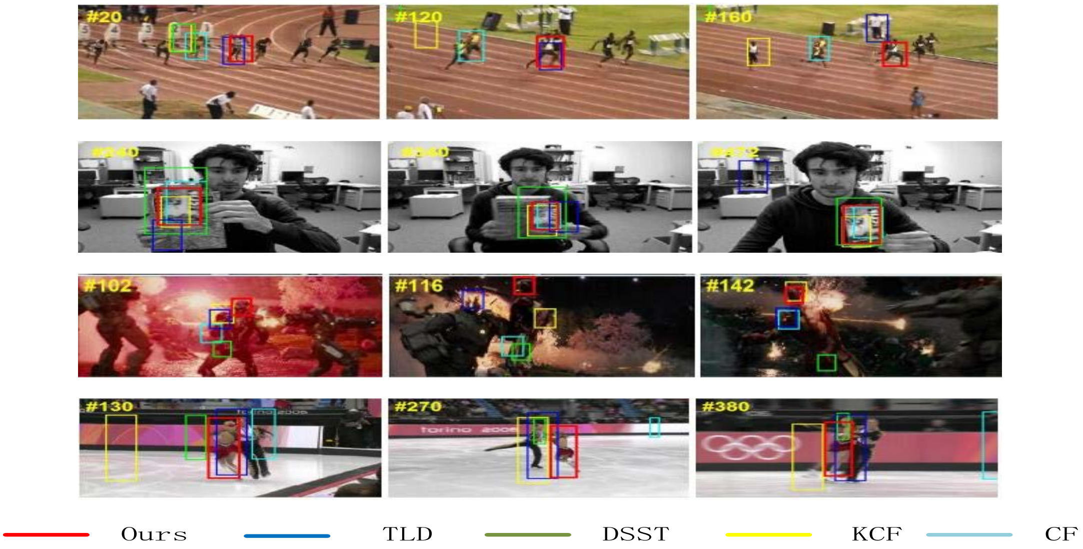

Figure 3. Comparison of various algorithm tracking results

As shown in Figure 3, our algorithm achieves very good tracking results in most frames, and the overall position error is lower than other algorithms. The overlap rate of our algorithm with the real border in the tracking process can achieve considerable results, and the overlap ratio is higher than other algorithms. In the video of Figure 4, Ironman is affected by the illumination changes and the tracking target has a scale change. The algorithm achieves good tracking effect throughout the process. In Bolt2 video, the tracking target is occluded, motion blurred, etc. At the same time, the tracking target has rotated. Only the algorithm in this paper achieves a good tracking effect. In the ClifBar video, some parts of the tracking target are beyond the field of view and scale changes occur, although various algorithms can track the target. However, the algorithm of this paper has the best tracking effect; in the video, the tracking target scale changes and illumination changes occur in the entire video sequence. The algorithm in this paper achieves good tracking of the target; in Skating video, the tracking target Influenced by disturbance factors such as motion blur, illumination change, and fast motion, only the algorithm of this paper achieves accurate tracking; the tracking target rotates in the image plane, and the image plane rotates out, and is disturbed by illumination changes. It also achieves good tracking. In summary, the algorithm in this paper has good robustness and can accurately track the target.

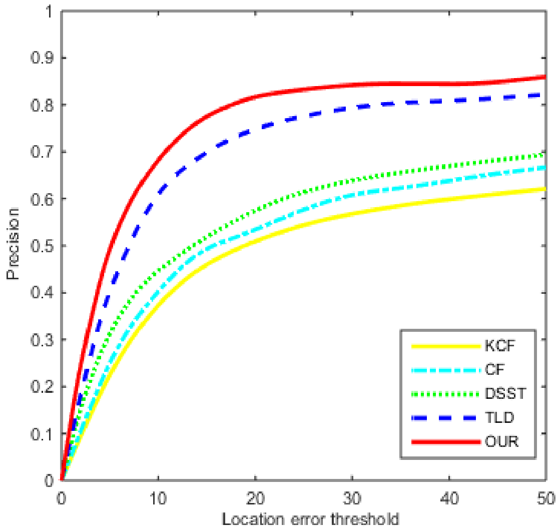

Figure 4. Precision plot

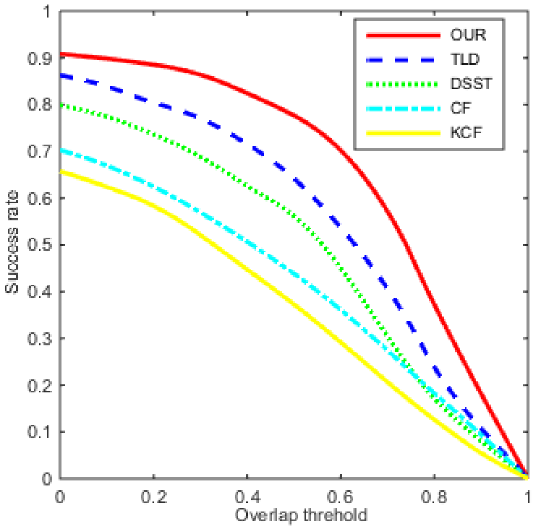

Figure 5. Success rate plot

It can be seen from Fig. 4 and Fig. 5 that the accuracy and success rate of the algorithm in this paper are ranked first, and a good tracking effect is achieved. The tracking performance is better than the other four tracking algorithms.

6. Conclusion

The excellent feature representation performance of convolutional neural networks improves the accuracy of tracking. However, training robust networks requires a large amount of offline training of image data. In order to improve the real-time tracking, we propose a general target tracker to learn the appearance and motion characteristics of the target, and then determine the target position through the VGG regression network, which further improves the tracking accuracy. Experiments show that compared with other related tracking algorithms, the performance of each algorithm is compared. It can be seen that the method can achieve more effective tracking of targets and has higher tracking efficiency.

Acknowledgments

Many thanks to the National Key R&D Program [No. 2016YFE0104600].

Your Name	Title*	Research Field	Personal website
Yibo Min	master student	Machine vision and object tracking	
Jianwei Ma	professor	Research on Control and Integrated Navigation Technology	
Shaofei Zang	lecture	Reinforcement learning	
Hongyun Zheng	master student	Object detection	

References

[1] Chinea-Rios M, Sanchis-Trilles G, Casacuberta F. Discriminative ridge regression algorithm for adaptation in statistical machine translation[J]. Pattern Analysis & Applications, 2018(1):1-13.

[2] Tang W, Li B, Tan S, et al. CNN-based Adversarial Embedding for Image Steganography[J]. IEEE Transactions on Information Forensics and Security, 2019, PP (99):1-1.

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

[4] Li Z, Gao S, Ke N, Robust object tracking based on adaptive template matching via the fusion of multiple features[J]. Journal of Visual Communication & Image Representation, 2017,44:1-20.

[5] HELD D, THRUN S, SAVARESES. Learning to Track at 100FPS with Deep Regression Networks / / Proc of the IEEE Conference on Computer Vision. Washington, USA: IEEE, 2016: 749-765.

[6] Yang H, Zhong D, Liu C, et al. Robust visual tracking based on deep convolutional neural networks and kernelized correlation filters[J]. Journal of Electronic Imaging, 2018, 27(2):1.

[7] Ma C, Huang J B, Yang X, et al. Hierarchical convolutional features for visual tracking [C]. IEE International Conference on Computer Vision, IEEE, 2016:3074-3082.

[8] CUI Z, XIAO S T, FENG J S et al. Recurrently Target-Attending Tracking / / Proc of the IEEE Conference on Computer Vision and Pattern Recognition Washington, USA: IEEE,2016: 1449 -1458.

[9] Wang L, Ouyang W, Wang X, et al. Visual tracking with fully convolutional networks [C]. IEEE International Conference on Computer Vision, IEEE, 2015:3119-3127.

[10] Nam, H, Han, B.: Learning multi-domain convolutional neural networks for visual tracking. arXiv. preprint arXiv:1510.07945 (2015).

[11] Zhang K, Liu Q, Wu Y, et al. Robust visual Tracking via convolutional networks without training [J]. IEEE Transactions on Image Processing, 2015,24(4):1779-1792.

[12] 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.

[13] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1717-1724.

[14] Danelljan M, Hager G, Shahbaz K F, et al. Convolutional features for correlation filter based visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015: 58-66.

[15] VALMADRE J, BERTINETTO L, HENRIQUES J F, et al. End to End Representation Learning for Correlation Filter Based Tracking [C/OL]. [2017-10-2]. https://arxiv.org / pdf /1704.06036.pdf.

[16] 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, 2014, 37(3):583-596.

[17] WANG N Y, YEUNG D. Learning a Deep Compact Image Representation for Visual Tracking // BURGES C J C, BOTTOU L, WELLING M, et al. eds. Advances in Neural Information Processing Systems 26. Cambridge, USA: The MIT Press, 2013: 809-817.

[18] Yunxia Wu, Ni Jia, Jiping Sun. Real-time multi-scale tracking based on compressive sensing[J]. The Visual Computer,2015,31(4).