An Improved Tracking Algorithm Based on MDNet

Biyang Wang, Baolong Guo, Qianying Li and Runzhi Liu

School of Aerospace Science and Technology, Xidian University, Xi’an 710071, Shaanxi – P.R. China
Email: 17709382583@163.com; blguo@xidian.edu.cn

Abstract. The MDNet algorithm works well on tracking problems of the video sequence, but the speed is very slow. We have made some improvements to accelerate the process of feature extraction. It enhances the expressive ability of the feature map by removing the max pooling layer and using the method of expanding convolution to increase the receptive field of each point on the feature map. In addition, MDNet is building on CNN, and there are problems that similar targets have a large interference to the results. To address this problem, We use RNN to capture the long-term dependence of the target before and after the target data in the sequence data, and introduce the RNN to model the structure information of the target object, and then fuse the RNN feature and CNN feature of the tracked target object. In addition, another new loss term is introduced to make the targets in different domains away from each other in the shared feature space, thereby improving the algorithm’s ability to discriminate similar interferers. Compared with MDNet, our improved algorithm is much faster and the accuracy is improved.

1. Introduction

The moving target detection and tracking technology directly operates on the input image, and through the reasonable image data expression and analysis, mines the position and shape characteristics of the target of interest, and transmits the information to the subsequent computer vision analysis system to complete the further analysis and processing. With a variety of applications, such as video surveillance, robotics, human-computer interaction and so on[1]. Because for sports targets, the scenes of their movements are very complex and often change, or the goals themselves are constantly changing. So how to identify and track changing targets in complex scenarios becomes a challenging task.

Deep neural networks[2] has made many breakthroughs in the field of image analysis and processing. Because of its powerful ability to extract high-level feature representations[3], it has been highly evaluated and widely used in the computer vision; such as image classification[4], recognition[5], saliency detection[6] and semantic segmentation[7]. Many CNN-based tracker[8,9,10] have been proposed so far. Among them, MDNet is an online tracking method based on multi-domain CNN architecture. It samples candidate regions by pre-training CNN on large-scale data sets and fine-tuning the first frame of the test video. However, its algorithm runs slower. In addition, MDNet is based on CNN. The focus is mainly on inter-class classification. In the presence of interferences, MDNet is likely to misclassify objects and backgrounds.

In recent years, the recurrent neural network (RNN)[11], which made great breakthroughs in the neural language process (NLP)[12], has been introduced into the field of computer vision. RNN can model the structure of the object itself and use this structural information to distinguish between tracking objects and similar interferers. Furthermore, the proposed RoIPooling[13] improves the speed of feature extraction, but due to the coarse quantization of feature map, a simple implementation leads
to poor positioning. To alleviate this harsh quantification of the RoI collection, Reference [14] proposed RoIAlign by bilinear interpolation.

In this paper, we have made some improvements to the network structure of MDNet. First, this paper use RoIAlign for feature extraction. In addition, this paper removes the max pooling layer and use the method of expanding convolution to increase the receptive field of each point on the feature map to enhance the expression ability of the feature map. Secondly, the RNN is introduced to model the structure information of the target object. Then, the RNN feature and CNN feature of the tracked target object are merged to enhance the discriminating ability of the tracking network between the tracked target and similar interferers. And at the same time, this paper introduces a new loss term, which is to make the targets in different domains far away from each other in the shared feature space. It will improve MDNet's ability to identify similar interferers.

The rest of this paper is organized as follows. Section 2 describes the related work in detail. Section 3 presents the improved network structure and experimental results. This article is summarized in Section 4.

2. Related Work
This section describes the MDNet architecture with the RoIAlign layer, which accelerates the feature extraction process. We also discuss the process of RNN modeling the structure of the tracking object, and introduce a new multi-domain learning method by introducing RNN and new loss item.

2.1. Efficient Feature Extraction
The original MDNet produce a lot of redundant parts like R-CNN when extracting features. so naturally I think of an improved version of ROI Pooling: RoIAlign . In this paper, a RoIAlign layer is added after the conv3 of the MDNet network to extract the fixed-size feature maps required for the fully connected layer. The feature extracted by RoI Align is a bit rough in itself. In order to improve the quality of RoI in the feature map before RoI Align, a high resolution and rich semantic information can be constructed by computing a denser full convolutional feature map and expanding each of the activated receptive fields. To do this, this paper remove the max pooling layer behind the conv2 layer in the original MDNet network, and perform dilated convolutions in conv3 layer with rate r=3 to increase the receptive field of each point on the feature map. The feature map output by conv3 after this processing is twice larger than the feature map output by conv3 in the original MDNet, making it possible to extract high-resolution features and improve the characterizing ability of the feature map. The improved network structure is compared with the original MDNet network structure as shown in figure 1.

![Figure 1. The improved network structure is compared with the original MDNet network structure](image-url)
2.2. RNNs for Object Self-Structure Modeling and Discriminative Feature Learning

2.2.1. RNNs for Object Self-Structure Modeling. The RNN can capture the long-term dependencies of the frames before and after the target in the sequence data. Using RNN to model the self-structure information of the target object. For a given input sequence \( \{x^t\} \), the output of the hidden layer \( h(t) \) and the output layer \( y^{(t)} \) at each time step is shown formula (1)(2).

\[
h(t) = \phi(M x^{(t)} + N h^{(t-1)} + z) \quad (1)
\]

\[
y^{(t)} = \sigma(K h_t + r) \quad (2)
\]

In formula (1)(2), \( M, N, K \) represent the input and hidden layers, the former hidden layer and the current hidden layer, the weight matrix between the hidden layer and the output layer, respectively; \( z \) and \( r \) represent deviation terms; \( \phi(\bullet) \) and \( \sigma(\bullet) \) are nonlinear activate the function.

Unlike one-dimensional data, the target tracking task needs to process two-dimensional image data. Therefore, we need to encode the self-structure of the two-dimensional image into an undirected cyclic graph, but such a graph cannot be directly applied to the RNN directly, so Approximate them using the four different directed acyclic graphs shown in figure 2.

For the four directed acyclic graphs in figure 2, we can conclude that:

\[
h^{(v)}_m = \phi(M x^{(v)} + \sum_{v_j \in P g_m(v)} N h^{(v)}_m + z_m) \quad (3)
\]

\[
y^{(v)} = \sigma(\sum_{g_m \in g''} K_m h^{(v)}_m + r) \quad (4)
\]

In the above formula, \( g_m (m = 1, 2, 3, 4) \) represents the four directed acyclic graphs, \( g'' = \{g_1, g_2, g_3, g_4\} \). \( P g_m(v_j) \) represents the predecessor node set of \( v_j \).

Finally, we can get the error back-propagated to previous convolutional layer at \( v_i \), where is expressed as

\[
\nabla x^{(v)} = \sum_{g'' \in g''} M^T_m dh^{(v)}_m \circ \phi'(h^{(v)}_m) \quad (5)
\]

In our improved network structure, we only join the RNN layer after CONV1 and adopt a skip concatenation strategy. Its network structure is shown in figure3.
2.2.2 Pre-training for Discriminative Instance Embedding. MDNet only distinguishes between targets and backgrounds in each domain, and it may not be able to distinguish foreground targets in different domains, especially if the current scene target belongs to the same semantics or has a similar appearance. Therefore, on the basis of the MDNet system, we propose a new loss term, which embedding foreground objects from multiple videos into each other, and letting targets in different domains move away from each other in the shared feature space, and can learn discriminative representations of the invisible target objects in new test sequences.

Our network minimizes a multi-task loss $S$, which is given by formula 6.

$$S = S_{cls} + \alpha S_{inst}$$

where $S_{cls}$ and $S_{inst}$ are loss terms for binary classification and discriminative in-stance embedding, respectively, and $\alpha$ is a hyper-parameter that controls balance between the two loss terms. Figure 4 illustrates the effects of discriminative feature embedding.

3. Improved Network Structure and Experiments
The overall structure of the improved network is shown in figure 5.
Figure 5. The overall structure of the improved network.

We analyzed our algorithm on the OTB100 dataset. Figure 6 shows the qualitative results of the improved algorithm in a partial dataset of the OTB100, and figure 7 shows the accuracy and success rate of the OTB100 dataset. In figure 6 and figure 7, IMP-MD is our improved algorithm.

Figure 6. The qualitative results of the proposed method

Figure 7. The accuracy and success rate of the OTB100 dataset.

In the experiment we compared our method with other trackers. Figure 7 shows the comparison of our method with other trackers in terms of accuracy and success. As can be seen from figure 7, the methods we propose is superior to other trackers. Our algorithm takes into account the structure of the object in the learning process and improves the ability of the network to distinguish between objects.
and backgrounds. Figure 6 illustrates that our tracker can effectively handle a variety of challenging situations.

4. Conclusion
We improved MDNet by learning discriminative representation of target in a multi-domain learning framework and considering the self-structural information of a target object. We accelerate the feature extraction process through the improved RoIAlign technology. At the same time, we use RNN to model the object's own structure and fuse the resulting RNN features with CNN features. In addition, we also introduce a new loss term, and finally get a new algorithm named IMP-MD. This algorithm can not only distinguish between targets, but also distinguish similar interferers. Our improved algorithms were evaluated on the public visual tracking benchmark dataset and demonstrated outstanding performance.

5. References
[1] Yilmaz A. Object tracking: A survey[J]. Acm Computing Surveys, 2006, 38(4):13.
[2] Lecun Y, Boser B, Denker J S, et al. Backpropagation Applied to Handwritten Zip Code Recognition[J]. Neural Computation, 1989, 1(4):541-551.
[3] Girshick R, Donahue J, Darrelland T, et al. Rich feature hierarchies for object detection and semantic segmentation[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014.
[4] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2).
[5] Yan Z, Hao Z, Piramuthu R, et al. HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition[C]. IEEE International Conference on Computer Vision. 2016.
[6] Wang L, Wang L, Lu H, et al. Saliency Detection with Recurrent Fully Convolutional Networks[C]. European Conference on Computer Vision. 2016.
[7] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.
[8] Ma C, Huang J B, Yang X, et al. Hierarchical Convolutional Features for Visual Tracking[C]. 2015 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society, 2015.
[9] Nam H, Han B. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking[J]. 2015.
[10] Tao R, Gavves E, Smeulders A W M. Siamese Instance Search for Tracking[J]. 2016.
[11] Elman J L. Finding Structure in Time[J]. Cognitive Science, 1990, 14(2):179-211.
[12] Graves A. Sequence Transduction with Recurrent Neural Networks[J]. Computer Science, 2012, 58(3):235-242.
[13] R. Fast R-CNN[J]. Computer Science, 2015.
[14] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C]. 2017 IEEE International Conference on Computer Vision (ICCV). 2017.