Single Object Tracking for Dynamic Programming based on Fine-grained Network

Jianguo Jiang$^{1,2}$.b, Kaiyuan Jin$^{2,a}$, Jingjing Wu$^{2,d}$, Meibin Qi$^{1,2}$.c and Qian Wang$^{2,e}$

$^1$Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education
$^2$School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230601, China

Corresponding Author: a) nolj_kim@foxmail.com; b) jgjiang@hfut.edu.cn; c) qimeibin@163.com; d) hfutwu@jingjing@mail.hfut.edu.cn; e) tinaw25@foxmail.com

Abstract. Single target tracking has always been a key and challenging research field in computer vision. Currently, an increasing number of researchers are focusing on extracting better tracking features and designing the best tracker. This paper proposes a new single target tracking network that uses fine-grained features and dynamic programming (DPFNet). In order to extract superior features, we added an attention module to the regression network enabling us to extract finer-grained and discriminative features to achieve regression. Besides, we did observe that different objects have varying moving rates; for different moving targets, the magnitude of the changes in target position within two adjacent frames is not the same either. Although an area search of 4 times the target's size can be applied to most objects, targets with large position changes may appear in other image areas outside the search area and the target would not be located as a result. Aiming at solving this problem, when designing the tracker, this paper analyzes some of the indicators for predicting the location and uses the analysis results to determine whether the search area is appropriate, so as to dynamically adjust the extent of the search area thereby significantly improving the tracking function. In other words, the size of the search area can be dynamically recalibrated for different images. Subsequent experiments prove that the method put forward in this paper achieves State-of-the-Art results.

1. Introduction

Visual object tracking has been one of the hottest research fields of computer vision in recent years. It has extremely extensive research prospects for intelligent monitoring systems and vehicle navigation systems. This task's objective is to locate the desired target in subsequent frames of a video based on a known target in the first frame of a given video.

Single object tracking usually entails two solutions: target classification and target estimation. Regarding target classification tasks, the primary goal is to classify objects into two categories, foreground, and background, in order to robustly provide a rough location of the target in the image. Whereas target estimation approximately determines the target's current state to optimize the precision of the coarse position rendering it as close as possible to the accurate state.

The majority of existing object tracking algorithms apply multi-scale search methods to boost the
accuracy of the target's bounding box [1,2]. Recently, two algorithms have become particularly prominent in solving the object tracking problem, namely the correlation filtering algorithm [3] and the RPN-based tracking algorithm [4]. According to the experimental results, it was discovered that a recent article (ATOM) [5] using the correlation filtering algorithm method has achieved a significant tracking accuracy, hence this paper shall be focalized on improving and studying this particular method.

The ATOM article proposed a new training idea, which consists of maximizing overlap. In detail, the idea is to maximize the overlap between the predicted bounding-box and the actual bounding-box to improve the tracking effect. The ATOM algorithm achieves object tracking in three steps: 1) Selecting a fixed size search area of the current frame. 2) Determining the approximate position of the target in the search area through the target classification task. 3) Regressing the approximate position of the target using target regression to OTBain the target's precise position by IoU regression technology.

Although ATOM has achieved superior results, it still faces problems and deficiencies: 1) Different objects have different motion rates, implying that the amplitude of the target position changes in two adjacent frames is not the same for different moving targets. 2) Only the basic Resnet structure is used in feature extraction, and the results are easily affected by the background information.

Consequently, this paper introduces Dynamic Programming based on Fine-grained Network (DPF) to solve the above two shortcomings. In a nutshell, this paper has made the following contributions: 1) The IoU value is estimated through the IoU prediction network, and different targets in various search areas are dynamically selected through the estimated IOU value. 2) An attention mechanism is added to the Resnet structure to ensure that the network constantly pays attention to the areas of interest in the picture and improves the discriminability of feature extraction.

2. Proposed Method

This paper proposes a new tracking method which is learned offline, which is introduced in Section 2.1. Meanwhile in Section 2.2, this paper introduces a novel dynamic programming online tracking method.

![Figure 1. Structure of Target Estimation Network](image-url)
bounding box. Inspired by the paper [5], in response to some of the deficiencies described above, we further improved the network structure based on that paper and proposed a more advanced target estimation network. The target estimation network structure is illustrated in Figure 1.

The network mainly consists of two branches, which are the template and Search branch in the upper and lower part, respectively. The input of the template branch is a frame in the video and a given bounding box, this branch’s main function is to extract the target features in the given bounding box. In contrast, the input of the search branch is another picture in the video that is not adjacent to the input frame of the reference branch, and 16 bounding boxes randomly generated around the target. The main function of this branch is to extract features in the 16 bounding boxes respectively. After that, the feature extracted from the template branch will be multiplied with the features extracted from the 16 bounding boxes, and the IOUs of the 16 randomly generated bounding boxes will eventually be predicted by the multiplied features.

As mentioned above, although the structure proposed in the paper [5] has achieved favourable results, the final IOU prediction will still contain some error margin since the effect of feature extraction is far from perfect. In view of this dilemma, we note that fine-grained features that are easily overlooked can serve as an important basis for the classification tasks. Inspired by the paper [6], we combine the Residual attention mask with Resnet-18 block3 in the network, as displayed by the yellow box in Figure 1, we call it the attention module. This design’s purpose is to focus the network attention on some notable fine-grained features to increase the classification effect.

The attention module’s schematic diagram is shown in Figure 2. Compared with the previous setting, the attention constraint is performed before feature G1 is input to Resnet18 Block3. Feature G1 is sent to both the Mask Unit and the Trunk Unit. The Trunk Unit’s output is a feature map, while the Mask Unit’s is equivalent to a feature weight map. The two outputs also perform a feature fusion operation to OTBain the final output G_f, which is equivalent to adding weight to each corresponding pixel on the feature G1.

![Figure 2. The schematic diagram of the attention module (better viewing of color pictures)](image)

It is worth noting that feature fusion is not a simple matrix point multiplication of two outputs, but operates according to following formula:

\[ G_j(x) = (1 + M(x)) \ast T(x) \]  

(1)
Where $M(x)$ means the output of the Mask Unit, and $T(x)$ means the output of the Trunk Unit. Since the value range of $M(x)$ is $[0,1)$, in order to avoid the problem of disappearing features caused by the value of $M(x)$ being zero, we add 1 to $M(x)$, it guarantees that even in the worst case, Resnet Block3 can get the feature input.

2.2. Dynamic Programming Online Tracking (DPF tracker)
It runs at over 30 FPS on a Nvidia GTX2080Ti GPU. In order to solve the situation described in the foregoing, concerning the fact that the target may appear in other image areas outside the search area and the target would not be located, this paper presents a new type of tracker based on dynamic programming during the online tracking, we named the tracker as DPF tracker. Our DPF tracker is implemented in Python, using PyTorch. The main steps to accomplish tracking comprise of the following six steps.

**Step 1:** The targets in the first frame of the video used for the test are collected and used to train the online classifier. This online classifier's training and updating strategy is identical to ATOM's [5].

**Step 2:** Use the online classifier for the search area of the video's second frame (the initial value is 4 times the target size) to get the online classification score map. According to the original image position corresponding to the maximum score in the classification score map, it is determined as the second frame target's rough position.

**Step 3:** The rough position is input to the DPF Network trained offline, and the rough position's IOU value is predicted.

**Step 4:** If the prediction IOU is higher than a preset threshold, the operation proceeds to Step 5. Conversely, if the prediction IOU is lower than the threshold, the classification result may be biased. In that case, the search area is expanded based on the target's heuristic knowledge a few frames before the video (the expansion principle will be described later), finally returning to Step 2 to re-execute according to the new search area. The new rough position is then fed into the Step 5 operation. The threshold is set to 0.2 based on the validation set results.

**Step 5:** The coarse position OTBained by dynamically changing the size of the search area is sent to the DPF Network, and then optimized to OTBain an accurate position that can maximize the IOU.

**Step 6:** The other frames in the video are processed in accordance with Step2-Step5.

For the principle of expanding the search area mentioned in Step 4, we have the following definition:

1) If $k = 2$, expand the search area to 5 times the size of the target area;
2) If $k > 2$, in the $k-1$ and $k-2$ frames, the horizontal or vertical displacement of the target to be tracked is considered to be greater than 1.5 times the width of the target template, then the search area is enlarged to 6 times the target area's size; otherwise, the search area is broadened to 5 times the size of the target area.

3. Experiment
We evaluated the proposed tracker DPF on two benchmarks: VOT-2018[7] and OTB-100[8]. In the following lines, this paper compares the DPF tracker proposed in this paper with some of the best trackers in the world at these two benchmarks and demonstrates its superiority. Besides, this paper equally scrutinizes the changes made in the previous experiment to prove the effectiveness of each component.

3.1. Comparisons on VOT-2018 benchmarks
This paper tests DPF on VOT2018. VOT2018 contains 60 test video sequences. The evaluation standard is the Expect Average Overlap Rate (EAO). This indicator can simultaneously display the tracking accuracy and robustness. The Accuracy indicates the average overlap rate between the tracking frame and the actual frame in the tracking success state. The results are outlined in Table 1.

Table 1 delineates our comparison of the six currently most advanced trackers. It can clearly be
depicted that the DPF tracker set forth in this paper has achieved the best results in the main EAO. Furthermore, it is worth noting that compared with the ATOM tracker, the method proposed in this paper exhibits a 2.2\% improvement on EAO, indicating that our method can effectively solve the aforementioned challenges faced in this paper.

**Table 1.** Comparison on the VOT-2018 dataset in terms of expected average overlap (EAO(\%)), accuracy(\%) and robustness(\%).

|               | DRT [9] | RCO [7] | UPDT [10] | DaSiam-RPN N[11] | ATOM [5] | Siam-RPN++ [12] | DPF (Ours) |
|---------------|---------|---------|-----------|------------------|----------|------------------|------------|
| **EAO**       | 35.6    | 37.6    | 37.8      | 38.3             | 40.1     | 41.4             | **42.3**   |
| **Accuracy**  | 51.9    | 50.7    | 53.6      | 58.6             | 59.0     | 60.0             | **60.0**   |
| **Robustness**| 20.1    | 15.5    | 18.4      | 27.6             | 20.4     | 23.4             | **18.3**   |

3.2. **Comparisons on OTB-100 benchmarks**

We tested the performance of the DPF tracker on OTB-100 benchmarks here and compared it with some papers, including ATOM which is the baseline of this paper. The evaluation index is the AUC metric. The results are revealed in Table 2.

**Table 2.** Comparison on the OTB-100 dataset in terms of AUC(\%).

|               | ATOM [5] | DaSiam-RPN N [11] | MDNet [13] | DPF (Ours) |
|---------------|----------|-------------------|------------|------------|
| **OTB-100**   | 66.9     | 65.8              | 67.8       | **68.2**   |

As can be seen in Table 2, compared with all methods, the DPF tracker achieved the best results with an excellent performance of 68.2\%.

3.3. **Analysis of our Approach**

We performed ablation experiments on the VOT2018 benchmarks for the two-point design we described in Chapter 2 in order to prove their effectiveness. The experimental results are shown in Table 3. The baseline points out that according to ATOM's design, the IOU OTBained from the network without the attention module is used exclusively as a pre-training label, using only an online classifier and no dynamic programming.

**Table 3.** Different structural designs comparison on the VOT-2018 dataset in terms of expected average overlap (EAO(\%)), accuracy(\%) and robustness(\%).

|                     | EAO    | Accuracy | Robustness |
|---------------------|--------|----------|------------|
| **Baseline**        | 40.1   | 59.0     | 20.4       |
| **Baseline +Attention module** | 41.1   | 60.4     | 19.2       |
| **Baseline +dynamic programming** | 40.8   | 60.7     | 19.2       |
| **Baseline +Attention module +dynamic programming** | 42.3   | 60.0     | 18.3       |

The results in Table 3 are in perfect consonance with the experimental results in the three cases. The results suggest that, in these three cases, the method proposed in this paper cannot be applied to achieve optimal results. Judging from the EAO value, separate addition of the attention module and dynamic programming can increase the baseline EAO by 1\% and 0.7\%, respectively. However, a
combination of both would be optimal, since it can augment the baseline EAO by 2.2%, providing irrefutable proof that the design proposed in this paper is effective.

4. Conclusion
This paper proposes an offline training object tracking network DPFNet and a dynamic programming based online tracker DPF Tracker. This design effectively solves the problems of target disappearance and inaccurate feature extraction in the current object tracking research field. It laid a certain foundation for future research.

Acknowledgments
This work is supported by the National Natural Science Foundation of China Grant (61771180, 61876056). Kaiyuan Jin is the corresponding author of this paper.

References:
[1] Lukezic, Alan, et al. "Discriminative correlation filter with channel and spatial reliability." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[2] Sun, Chong, et al. "Correlation tracking via joint discrimination and reliability learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[3] Danelljan, Martin, et al. "Learning spatially regularized correlation filters for visual tracking." Proceedings of the IEEE international conference on computer vision. 2015.
[4] Zhu, Zheng, et al. "Distractor-aware siamese networks for visual object tracking." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[5] Danelljan, Martin, et al. "Atom: Accurate tracking by overlap maximization." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[6] Wang, Fei, et al. "Residual attention network for image classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[7] Kristan, Matej, et al. "The sixth visual object tracking vot2018 challenge results." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[8] Wu Y, Lim J, Yang M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.
[9] Sun, Chong, et al. "Correlation tracking via joint discrimination and reliability learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[10] Bhat, Goutam, et al. "Unveiling the power of deep tracking." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[11] Zhu, Zheng, et al. "Distractor-aware siamese networks for visual object tracking." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[12] Li, B., et al. "Evolution of Siamese Visual Tracking with Very Deep Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA. 2019.
[13] Nam, Hyeonseob, and Bohyung Han. "Learning multi-domain convolutional neural networks for visual tracking." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.