Siamese Cascaded RPN with Mean Shift Tracking

Mudit Garg Vijay Vamsi Nadella Arun Teja Muluka
Project 1 Proposal, CSE 586 Computer Vision II, Spring 2020
{mxg5783, vvn5075, avm6604}psu.edu

February 13, 2020

Motivation

The Siam-RPN[3] has been proved to perform with high accuracy and efficacy. However, the performance deteriorated when the distractors similar to the object we are tracking were introduced. To address this issue, Siamese Cascaded RPN (C-RPN)[1] was proposed. This model was successful in discriminating difficult backgrounds (i.e, similar distractors). It also leveraged the advantage of multi-stage regression, which helped in refining the location and shape of the target. Thus, making tracking more accurate.

Similarly, object tracking mean-shift algorithm has also been proved for locating the moving objects very precisely and accurately. It tracks the moving objects using the histograms (independent of the object color) which has turned out to be very efficient. It is also appreciated because it does not require extensive parameter tuning and yields remarkable results.

Proposal Outline

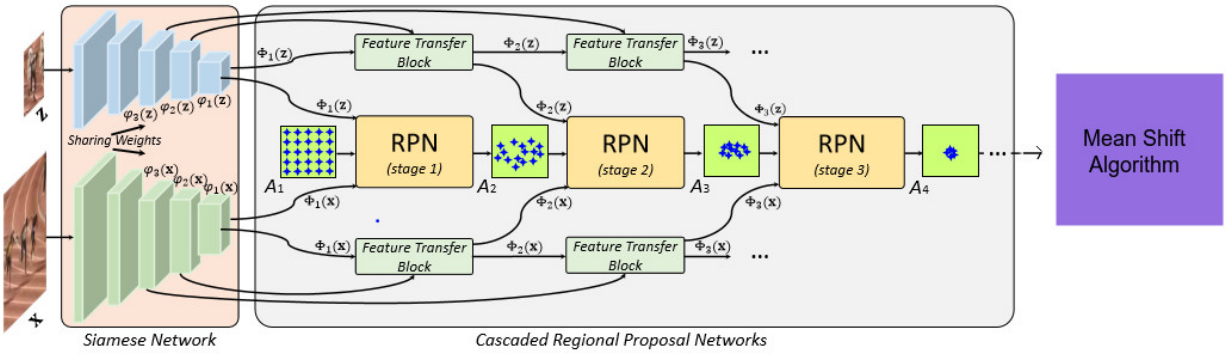
Given the advantages of the C-RPN and mean-shift tracking algorithm, we are proposing to integrate later with the former. We believe even more accurate and efficient results can be obtained with this combination. Siam-RPN model has demonstrated to work with high speeds ($160fps$)[3]. We are keeping the Siamese architecture hoping it would give us considerable speed along with the accuracy of C-RPN and mean shift.

As per [1], C-RPN has two subnetworks: (1) Siamese Network and (2) Cascaded RPN. The first network will extract the features of the target template x and the search region z. The second one will receive these features for each RPN. Feature Transfer Block(FTB) will be applied to merge the features from the convolution layers. Its output will serve as an input to each RPN layer. This architecture rejects the negative anchors, therefore, discriminating the similar distractions more effectively. With the observations from [1], the response map after each stage of RPN has the different discriminative ability. We will apply the mean-shift algorithm on the correlation response map of the last RPN layer. However, this work can be further extended by integrating the mean-shift to the response function of every RPN layer.

In the proposed architecture, we plan to apply the mean-shift algorithm after the images have been processed by the multiple RPN layers. The architecture of the model will be similar to the C-RPN, but

the accuracy of this framework might be further improved using the proposed method. Considering the performance of both C-RPN and mean shift, even though the proposed model might be slower when compared to C-RPN, we believe it can be good enough to act in real-time processing.

Figure 1: Proposed Architecture



Dataset and Model Training

Similar to Siam-RPN, C-RPN is trained offline using a large number of image pairs which are sampled from the same sequence as in [3]. We will use the same methodology. A pre-trained model is already given by [1], which we are planning to use. After building the hybrid model, we will test the same on VOT-2016, VOT-2017, and LaSOT datasets. It will be compared with the available state of art online trackers.

References

[1] Fen, Heng and Ling, Haibin, "Siamese Cascaded Region Proposal Networks for Real-Time Visual Tracking", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] Collins, Robert T., "Mean-shift blob tracking through scale space", *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2003.

[3] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, Xialin Hu, "High Performance Visual Tracking with Siamese Region Proposal Network", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (8971-8980), 2018.

[4] Bertinetto, Luca and Valmadre, Jack and Henriques, Joao and Vedaldi, Andrea and Torr, Philip, "Fully-Convolutional Siamese Networks for Object Tracking", *25th IEEE International Conference on Image Processing (ICIP)*, 2016