Multi-Branch Siamese Networks with Online Selection for Object Tracking

Zhenxi Li[1], Guillaume-Alexandre Bilodeau[1], and Wassim Bouachir[2]

[1]LITIV lab, Polytechnique Montreal
{zhenxi.li, guillaume-alexandre.bilodeau}@polymtl.ca
[2]TELUQ University
wassim.bouachir@teluq.ca

Abstract. Model-free visual object tracking is one of the most fundamental problems in computer vision. Given the object of interest marked in the first video frame, the objective is to localize the target in subsequent frames, despite object motion, changes in viewpoint, lighting variation, among other disturbing factors. One of the most challenging difficulties with model-free tracking is the lack of prior knowledge on the target object appearance. Since any arbitrary object may be tracked, it is impossible to train a fully specialized tracker.

Recently, convolutional neural networks (CNNs) have demonstrated strong power in learning feature representations. To fully exploit the representation power of CNNs in visual tracking, it is desirable to train them on large datasets specialized for visual tracking, and covering a wide range of variations in the combination of target and background. However, it is truly challenging to learn a unified representation based on videos that have completely different characteristics. Some trackers [1] train regression networks for tracking in an entirely offline manner. Other works [2, 3, 6] propose to train deep CNNs to address the general similarity learning problem in an offline phase, and evaluate the similarity online during tracking. However, since these works have no online adaptation, the representations they learned offline are general but not always discriminative.

Rather than applying a single fixed network for feature extraction, our contribution is to use multiple network branches with an online branch selection mechanism. It is well known that different networks designed and trained for different tasks have diverse feature representations. Therefore, there are two strategies to improve the discriminative ability of the tracking networks. The first one is training the network in different contexts, while the second one is to use multiple networks designed and trained for different tasks. In our approach, we utilize context-dependent branches pretrained in different contexts in addition to another branch pretrained for the image classification task to improve our tracking performance. We note that more branches could be added with other pretrained networks at the cost of slower performances. With the online branch selection mechanism, our tracker dynamically selects the most efficient and robust branch for target representation, even if the target appearance changes. Our goal is to improve the generalization capability with multiple networks.

To verify the contribution of each branch and the online branch selection mechanism of our algorithm, we implemented several variations of our approach and evaluated them on the OTB benchmarks [4, 5]. Firstly, we compared our full branches algorithm with various combination of branches. Results demonstrate that the proposed multiple branches architecture allows a better use of diverse feature representations. Then, we conducted experiments on the branch selection interval T. Our experiments showed that a frequent execution of the selection mechanism increases the possibility of selecting an inappropriate branch, while the tracking performance is also decreased if we keep for a too long period a branch that is not discriminative anymore. These experiments demonstrate that the optimal branch selection interval T is 7 frames.
We also compare our Multi-Branch Siamese Tracker (MBST) with several state-of-the-art trackers on OTB benchmarks. The overall comparison demonstrated that the proposed MBST achieves the best performance among state-of-the-art trackers on OTB benchmarks [4, 5]. Notably, it outperfroms SiamFC [2] as well as its variation CFNet [3] on all datasets. This demonstrates that diverse feature representations are important to improve tracking, as feature maps from various CNNs can be quite different. On the other hand, the experiments on challenging situations demonstrates that our tracker effectively handles all kinds of challenging situations (illumination variation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutters, low resolution) that often require high-level semantic understanding.

References

1. Held, D., Thrun, S and Savarese, S.: Learning to Track at 100 FPS with Deep Regression Networks. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., ECCV 2016, pp. 749–765. Springer
2. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A. and Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: ECCV 2016, pp. 850–865. Springer
3. Valmadre, J., Bertinetto, L., Henriques, J.F., Vedaldi, A. and Torr, P.H: End-to-end representation learning for correlation filter based tracking. In: CVPR 2017, pp. 5000–5008. IEEE
4. Wu, Y., Lim, J. and Yang, M.H.: Online object tracking: A benchmark. In: CVPR 2013, pp. 2411–2418
5. Wu, Y., Lim, J. and Yang, M.H.: Object tracking benchmark. TPAMI**37**(9), 1834–1848(2015)
6. He, A., Luo, C., Tian, X. and Zeng, W.: A twofold siamese network for real-time object tracking. In: CVPR 2018, pp. 4834–4843