Review for Know Your Surroundings: Exploiting Scene Information for Object Tracking

Irem ARPAG

May 3, 2021

Although traditional frame-by-frame detection-based object tracking that is one of the main computer vision problems is based on estimating the state of the target object in each frame of a video sequence with its initial frame, tracking an object needs a much more global view of the scene that is dealing with not only the target object in the frame but also other objects in the scene. The challenges like fast appearance change or the presence of distractor objects make it harder for the existing appearance-based approaches to locate the target object that update the model with only previously tracked frames that cannot be sufficient for capturing the locations and characteristics of the other objects esp. in a competitional scene. To take the advantage of aweing the presence and locations of other objects for robust tracking, inspired from the human's experience of tracking objects, the authors of the article purpose a significant architecture which can prophage valuable scene knowledge through the sequence by using a dense set of localized state vector for encoding the position of the local region to decide if it belongs to the target, background or a distractor object. This precious scene information combined with the target appearance model is utilized to estimate the target state in each frame by using a learned predictor module which is updated through a recurrent neural network model.

Performing expansive experiments on six challenging benchmarks and achieving state-of-art results on five of them are the significant strengths of the paper that also demonstrate the generalization capacity of the module. Obtaining an average overlap (AD) score of 63.6 percentage, better than 2,5 percentage former best result is the highlight of the model, with a large scale TrackingNet dataset they also confirm the success on real world videos via an AUC (Area-under the-curve) of 74. Meanwhile, detailed ablation study analyzing the impact of key components in their tracker supports the comprehensive aspect of the work.

Since the model is designed for short-term tracking its evaluation with long-term tracking dataset LaSOT decrease the performance that can be explained with the fast update of the state vectors in the approach.

The researchers installed an elaborated ablation study to examine the function of each component in the architecture by using overlap precision (OP) metric and are-under-the-curve (AUC) score. The first components in impact of scene information for tracking in which they compare to "only appearance model" with their original approach, showing apparently the importance of scene knowledge with a 1.3 percentage improvement in AUC score. The results of the analysis on state propagation indicate the critical importance of the component, and the outcomes on propagation reliability demonstrate that exploiting reliability score is valuable with its 0.3 percentage AUC progress. Although they know that for long term robustness esp. on occlusions, their tracker depends on appearance model, it dramatically decreased the performance over 17 percentage in AUC score. The method is also evaluated on 6 different datasets; VOT 2018, GOT 10k, Tracking Net, OTB100-NFS and LaSOT, and the proposed model's tracker is compared with the SOTA approaches of each dataset. The comparison with VOT 2018 dataset's previous best method. DIMP-50, it achieves an increase of 5 percentage in EAO (Expected Average Overlap). The results on challenging GOT10k dataset which forbids to use external training data, confirm clearly the benefit of using scene knowledge for tracking with an AO score of 63.6 while affirming its high speed once more on NFS dataset with its top results.

The reasons of low performance with the long-term dataset LaSOT seems to be a good future work for the approach that indicates its state-of-art performance on the other tracking benchmark.