END-TO-END TEMPORAL FEATURE AGGREGATION FOR SIAMESE TRACKERS

Zhenbang Li[a,c], Qiang Wang[a,c], Jin Gao[a], Bing Li[a,], Weiming Hu[a,b,c]*

[a]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[b]CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China
[c]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

While siamese networks have demonstrated the significant improvement on object tracking performances, how to utilize the temporal information in siamese trackers has not been widely studied yet. In this paper, we introduce a novel siamese tracking architecture equipped with a temporal aggregation module, which improves the per-frame features by aggregating temporal information from adjacent frames. This temporal fusion strategy enables the siamese trackers to handle poor object appearance like motion blur, occlusion, *etc*. Furthermore, we incorporate the adversarial dropout module in the siamese network for computing discriminative target features in an end-to-end-fashion. Comprehensive experiments demonstrate that the proposed tracker performs favorably against state-of-the-art trackers.

Index Terms— Visual object tracking, siamese network, feature aggregation, adversarial training

1. INTRODUCTION

Visual object tracking is the task of estimating the state of an arbitrary target in each frame of a video sequence. Recently, siamese networks have demonstrated the significant improvement on object tracking performances. However, the learned generic representation may be less discriminative because of the deteriorated object appearances in videos (Fig. 1), such as motion blur, occlusion, *etc*. Researchers try different ways to improve the feature representation. For example, SA-Siam [1] separately trains two branches to keep the heterogeneity of semantic/appearance features. In DaSiamRPN [2], a novel distractor-aware incremental learning module is designed, which can effectively transfer the general embedding to the current video domain and incrementally catch the target appearance variations during inference. SiamRPN++ [3] introduces a simple yet effective sampling strategy to drive the siamese tracker with more powerful deep architectures. These efforts have produced some impact and improved state-of-the-art accuracy. However, all above siamese algorithms perform tracking based on features cropped from only the current frame, which limits the power of siamese trackers.

*Corresponding author.

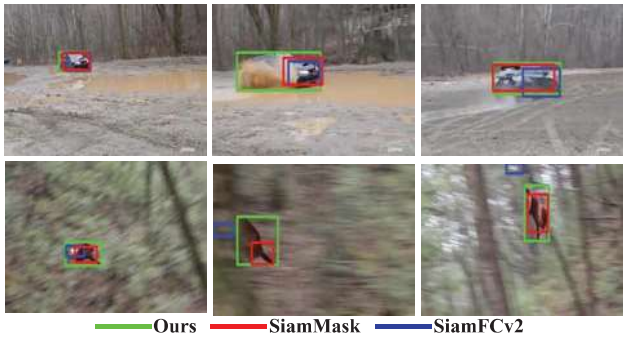

Fig. 1. A comparison of our method with SiamMask and SiamFCv2. The example frames are from the GOT-10k testing set. Our approach effectively handles poor object appearance compared to existing approaches.

Actually, the video has rich information about the target and such temporal information is an important basis for video understanding and tracking. For example, in video object detection, FGFA [4] leverages temporal coherence on feature level. It improves the per-frame features by aggregation of nearby features along the motion paths, and thus improves the video recognition accuracy. In video object segmentation, STCNN [5] introduces a temporal coherence module, which focuses on capturing the dynamic appearance and motion cues to provide the guidance of object segmentation. In discriminative correlation filter-based object tracking, FlowTrack [6] focuses on making use of the rich flow information in consecutive frames to improve the feature representation and the tracking accuracy. However, how to utilize the temporal information in siamese trackers has not been widely studied yet.

In this paper, we aim to take full advantage of temporal information in siamese trackers. We introduce a novel siamese tracking architecture equipped with a temporal aggregation module, which improves the per-frame features by aggregating features from adjacent frames. This temporal fusion strategy enables the siamese tracker to handle poor object appearance like motion blur, occlusion, *etc*. To achieve this, we shift the channels along the temporal dimension [7] in the backbone of the siamese network. Note that features of the same object are usually not spatially aligned across frames due to