Transformer Tracking Using Spatiotemporal Features

J. Kugarajeevan [1,2,*], T. Kokul [3], A. Ramanan [3], S. Fernando [2]

[1] Department of Computational Mathematics
University of Moratuwa
Moratuwa, Sri Lanka

[2] Department of Interdisciplinary Studies
University of Jaffna
Jaffna, Sri Lanka

[3] Department of Computer Science
University of Jaffna
Jaffna, Sri Lanka
* Email: jananitha@univ.jfn.ac.lk

***Abstract*— Recent approaches in single-object tracking leverage Transformer-based architectures to achieve state-of-the-art performance by utilizing the attention mechanism between target templates and search region patches. However, most existing trackers fail to capture the temporal information of the target object present in the sequences of frames, making them unsuitable for capturing changes in the target's appearance over time. This leads to suboptimal performance in real-world scenarios where the target's appearance can change significantly due to variations such as occlusions, deformations, and abrupt movements. To overcome this limitation, a Transformer-based tracking framework is introduced that simultaneously captures the spatial and temporal features of the target object and utilizes that knowledge to accurately locate the target. This is achieved by adapting the standard Transformer encoder architecture, enabling the direct extraction of spatiotemporal features of the target object from a sequence of frames. As the first step, the sequences of template frames and search region are partitioned into non-overlapping spatiotemporal patches, which are then tokenized and processed by the Transformer encoder to capture spatiotemporal features. Additionally, the Transformer encoder is initialized with spatiotemporal masked autoencoders, which are trained to capture the relationship between patches through self-supervised learning. Extensive experiments on the GOT-10k benchmark dataset demonstrate that our tracker surpasses current state-of-the-art methods, achieving an average overlap of 70.9%, a success rate of 82.4% at a 0.5 threshold, and a success rate of 66.3% at a 0.75 threshold on test data, while maintaining real-time processing speed of 47 frames per second on an NVIDIA P100 GPU.**

Keywords— Transformer Tracking, Spatiotemporal Tracking, Visual Object Tracking

979-8-3503-8969-2/24/$31.00 ©2024 IEEE