GLOBALLY SPATIAL-TEMPORAL PERCEPTION: A LONG-TERM TRACKING SYSTEM

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

Although siamese trackers have achieved superior performance, these kinds of approaches tend to favour the local search mechanism and are thus prone to accumulating inaccuracies of predicted positions, leading to tracking drift over time, especially in long-term tracking scenario. To solve these problems, we propose a siamese tracker in the spirit of the faster RCNN's two-stage detection paradigm. This new tracker is dedicated to reducing cumulative inaccuracies and improving robustness based on a global perception mechanism, which allows the target to be retrieved in time spatially over the whole image plane. Since the very deep network can be enabled for feature learning in this two-stage tracking framework, the power of discrimination is guaranteed. What's more, we also add a CNN-based trajectory prediction module exploiting the target's temporal motion information to mitigate the interference of distractors. These two spatial and temporal modules exploit both the high-level appearance information and complementary trajectory information to improve the tracking robustness. Comprehensive experiments demonstrate that the proposed Globally Spatial-Temporal Perception-based tracking system performs favorably against state-of-the-art trackers.

Index Terms**—** Visual object tracking, siamese network, motion model

1. INTRODUCTION

Object tracking [1, 2, 3] is a challenging problem in the field of computer vision, which aims to establish the positional relationship of the object to be tracked in a continuous video sequence. The popular siamese trackers [4, 5, 6] are typically based on the local search mechanism: searching the target within a small neighborhood centered on the target position of the previous frame to determine its current position. This mechanism works well if the target only has a small displacement between two adjacent frames. It also brings benefits in another aspect, which is to avoid interference from the distractors in the background.

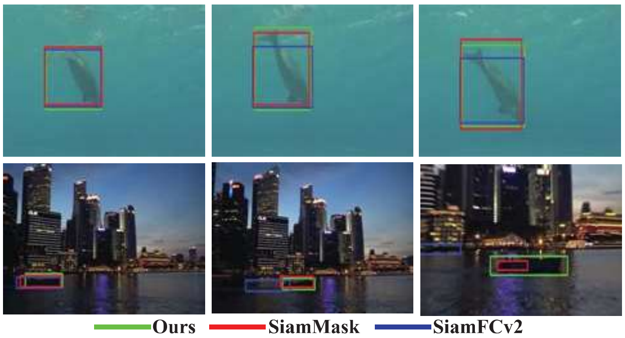

Fig. 1. A comparison of our method with the stage-of-the-art trackers SiamMask [6] and SiamFCv2 [4] in challenging situations. The example frames are from the GOT-10k [7] testing set.

However, the local search mechanism bears some shortcomings. First, it could cause irreversible cumulative errors if the predictions of the target positions in the previous frames drift away due to challenging illumination change, motion blur, *etc.*, because the search area generated in the current frame may not cover the target leading to a complete failure in subsequent frames. Second, it is difficult for the local mechanism-based trackers to meet the needs of long-term tracking [8, 9]. Under the long-term scenario, the target frequently re-enters and re-exits the screen. Since the tracker cannot set the correct search area when the target leaves and re-enters the screen, it often fails to retrieve the target due to the wrong search area without the target covered.

Inspired by faster RCNN's two stage detection paradigm [10], we propose a siamese tracker based on the global perception mechanism. During the tracking process, our tracker is always able to perceive the target over the entire image. Therefore, even if the tracker makes a mistake due to the challenging target appearance variations, the target can still be retrieved in time once its appearance returns to normal. Especially under the long-term out-of-view disappearance scenario, where the tracker cannot find the target in the full image when the target leaves the screen, our tracker can continue to work when the target re-enters the screen from any position.

Besides the above globally spatial perception mechanism, we also propose a temporal motion model to mitigate the

*Corresponding author.

978-1-7281-6395-6/20/$31.00 ©2020 IEEE