Target-Specific Siamese Attention Network for Real-time Object Tracking

Kokul Thanikasalam, Clinton Fookes, Sridha Sridharan, Amirthalingam Ramanan and Amalka Pinidiyaarachchi

***Abstract*—Deep similarity trackers are able to track above real-time speed. However, their accuracy is considerably lower than deep classification based trackers since they avoid valuable online cues. To feed the target-specific information for real-time object tracking, we propose a novel Siamese attention network. Different types of attention mechanisms are used to capture different contexts of target information and then learned knowledge is used to feed target cues at different representation levels of similarity tracking. In addition, an online learning mechanism is employed to utilise the available target-specific data. The proposed tracker reduces the impact of noise in the target template and improves the accuracy of similarity tracking by feeding target cues into the similarity search. Extensive evaluation performed on OTB-2013/50/100 and VOT2018 benchmark datasets demonstrate the proposed tracker outperforms state-of-the-art approaches while maintaining real-time tracking speed.**

***Index Terms*—Visual object tracking, Deep neural networks, Siamese network, Attention network.**

I. Introduction

VISUAL object tracking is a well-known research topic in computer vision and is of great interest in many security-sensitive areas such as video surveillance [1], abnormal event detection [2], traffic monitoring [3], [4], along with other domains such as sport analysis [5] and medical imaging [6]. It is the task of estimating the trajectory of an unknown target in a video sequence, while only the initial location (a bounding box) is given. The objective of much of this research is to reach the tracking accuracy of humans in challenging scenarios such as in situations where the target moves in a cluttered background or becomes occluded by other objects or the appearance is changed due to deformation. Nowadays, in addition to the accuracy, most applications are demanding real-time tracking speed. In this paper, we aim to develop a robust framework for a model-free, single-camera, appearance-based single-object tracker with real-time speed.

Being discriminative and being able to generalise are the important characteristics of a robust tracker. The discriminative power of the tracker is used to differentiate the target from its surrounding context. Target-specific features are used to learn the discrimination and it helps to locate the target in cluttered backgrounds and across illumination variations. On the other hand, if a tracker is too discriminative, it fails to adopt to the significant appearance changes of the target. As a result, a robust tracker should also have a strong generalisation ability, which results in the tracker being able to manage the appearance changes of targets across scale variation or deformation. Maintaining the balance between discrimination and generalisation ability is important to a robust tracker which can be achieved by developing a strong appearance model from offline and online learning. Although capturing cues in an online manner improves the accuracy, the computational cost of online learning significantly reduces the tracking speed. These issues make the process of developing a robust and real-time tracker a challenging task.

Appearance based trackers model the target by learning its features. Most of the recent tracking frameworks [7]–[15] use Convolutional Neural Networks (CNNs) to model the target. CNNs show state-of-the-art performance in various computer vision tasks [16]–[19] due to their rich level hierarchical feature representation capability. Since CNNs employ an extensive set of parameters, they require huge amounts of training data. A massive video tracking dataset [20] became available only recently where previously data deficiency limited the number of CNN based trackers. Previous CNN based trackers [12], [13] manage the limitation by transferring offline learned CNN features to online tracking. Although such approaches allow an adequate generalization capability, they fail to learn the discriminative information of the target well and therefore they are always affected by distractors.

State-of-the-art CNN based appearance trackers are following two major strategies. The first group of approaches [7], [8], [10] follow the classification and update strategy. An online learned classifier network is used in these trackers and the target is located by the tracking-by-detection technique. A pre-trained covolutional network [16], [21], from an image classification task, is used as the baseline network in this strategy, and Stochastic Gradient Descent (SGD) is used to fine-tune the baseline network for object tracking. These approaches learn the target-specific cues online and update the classifier accordingly. Even though such methods achieve state-of-the-art accuracy, their tracking speed is very low due to the computational cost of updating a large number of parameters of the CNN.

The second group of CNN based trackers [9], [11], [22] follow the similarity tracking strategy. Such approaches obtain the target template from the first frame of a video sequence and then search the similarity of that template in upcoming frames to locate the target. Similarity trackers utilise the Siamese network architecture [23] to boost the generalisation power of template matching and are trained offline with a

T. Kokul obtained his PhD from the Queensland University of Technology (QUT), Australia and is working at the Department of Physical Science, Vavuniya Campus of the University of Jaffna, Sri Lanka. C. Fookes and S. Sridharan are with the SAIVT Lab, QUT, Australia. A. Ramanan is with the Department of Computer Science, University of Jaffna, Sri Lanka and U.A.J. Pinidiyaarachchi is with the Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka. T. Kokul is the primary author (e-mail: kokul1984@gmail.com).