

Deep Reinforcement Learning with Iterative Shift for Visual Tracking

Liangliang Ren^{1,*}, Xin Yuan^{1,*}, Jiwen Lu^{1,†}, Ming Yang², Jie Zhou¹

¹Tsinghua University; ²Horizon Robotics, Inc.

{renll16, yuanx16}@mails.tsinghua.edu.cn; {lujiwen, jzhou}@tsinghua.edu.cn; ming.yang@horizon-robotics.com

Abstract. Visual tracking is confronted by the dilemma to locate a target both accurately and efficiently, and make decisions online whether and how to adapt the appearance model or even restart tracking. In this paper, we propose a deep reinforcement learning with iterative shift (DRL-IS) method for single object tracking, where an actor-critic network is introduced to predict the iterative shifts of object bounding boxes, and evaluate the shifts to take actions on whether to update object models or re-initialize tracking. Since locating an object is achieved by an iterative shift process, rather than online classification on many sampled locations, the proposed method is robust to cope with large deformations and abrupt motion, and computationally efficient since finding a target takes up to 10 shifts. In offline training, the critic network guides to learn how to make decisions jointly on motion estimation and tracking status in an end-to-end manner. Experimental results on the OTB benchmarks with large deformation improve the tracking precision by 1.7% and runs about 5 times faster than the competing state-of-the-art methods.

Keywords: Visual object tracking, reinforcement learning, actor-critic algorithm

1 Introduction

Visual object tracking (VOT) aims at locating a target efficiently in a video sequence, which remains a challenging problem in unconstrained applications due to deformation, abrupt motion, occlusions and illumination, after several decades of intensive research [5, 10, 20, 36, 41, 42, 51]. Essentially VOT needs to address 3 key issues: 1) How to represent a target, i.e., the observation model; 2) How to efficiently leverage the motion smoothness assumption to locate a target in the next frame; 3) How to update tracking models online, if necessary, to handle dynamic scenarios.

The appearance models have evolved from intensity templates [19], color histograms [14], and sparse features [4], to the dominating deep features [47] extracted by CNN models. Thus, naturally tracking may be formulated as a

* Indicates equal contribution.

† Corresponding author.

