DEEP MATCH TRACKER: CLASSIFYING WHEN DISSIMILAR, SIMILARITY MATCHING WHEN NOT

T.Kokul[⋆†] *C.Fookes*[⋆] *S.Sridharan*[⋆] *A.Ramanan*[‡] *U.A.J.Pinidiyaarachchi*[◇]

[⋆]Image and Video Lab, SAIVT Program, Queensland University of Technology, Australia
[‡]Dept. of Computer Science, University of Jaffna, Sri Lanka
{[†]PGIS, [◇]Dept. of Statistics and Computer Science}, University of Peradeniya, Sri Lanka

ABSTRACT

Visual tracking frameworks employing Convolutional Neural Networks (CNNs) have shown state-of-the-art performance due to their hierarchical feature representation. While classification and update based deep neural net tracking have shown good performance in terms of accuracy, they have poor tracking speed. On the other hand, recent matching based techniques using CNNs show higher than real-time speed in tracking but this speed is achieved at a considerably lower accuracy. To successfully manage the trade-off between accuracy and speed, we propose a novel CNN architecture for visual tracking. We achieve this trade-off balance by using an approach in which consecutive similar frames are processed with a similarity matching technique, and dissimilar frames are processed with a classification approach within the CNN architecture. The tracking speed is improved by avoiding unnecessary model updates through the measurement of similarity between adjacent frames, while the accuracy is maintained by adopting a classification approach when needed, with deeper level features. Extensive evaluation performed on a publicly available benchmark dataset demonstrates our proposed tracker shows competitive performance while maintaining near real-time speed.

Index Terms— Object tracking, CNN, deep tracking

1. INTRODUCTION

Visual object tracking is one of the fundamental tasks in computer vision and has been receiving a rapidly increasing attention due to numerous applications [1]. In this paper, we consider single object tracking, where an unknown target identified in the first frame is to be tracked in subsequent frames automatically. It still remains a challenging problem since there exist enormous variations, which strongly influence both the target and background.

For several years, appearance-based tracking frameworks relied on hand-crafted features [2] to address the tracking challenges. However, these features do not generalize well and fail to capture the semantic information of targets, therefore they are not robust to significant appearance changes of targets. Recent tracking frameworks [3, 4, 5] which use convolutional neural networks (CNNs) have demonstrated a significant increase in tracking performance because of the rich feature representation of CNNs. However, significant challenges which still remain open for CNN based trackers include the scarcity of supervised training data and the ability to tune a large number of parameters in the architectures.

Several trackers have managed this limitation by transferring offline learned CNN features from a related task to online tracking. While some of these approaches [6, 7] directly use pre-trained CNN features to model the tracker, others [3, 8] fine-tune a couple of layers with the available limited data. Even though these approaches showed state-of-the-art performance, they are too slow for practical use because of the expensive online training of CNNs.

A few recent trackers [5, 9] boost the tracking speed by the use of fixed CNNs. They are trained offline and avoid online adaptation to increase the tracking speed to real-time. They use a similarity learning approach or regression network to predict the target's location. However their accuracy is considerably lower than the online learning approaches as they avoid video-specific cues in tracking.

A robust tracker should have high accuracy while maintaining considerable tracking speed which is important for many practical applications. Also online learning is necessary to a tracker as to adapt the appearance changes and to avoid drift. Based on that, we propose a deep tracking approach (that we refer to as MATCHNET) which improves the tracking speed while maintaining state-of-the-art accuracy. It is developed on the concept that if a target's appearance is similar to its previous frame, the target can be located by searching similarity in the current frame at early layers. Otherwise much deeper level features are needed to locate the target and the model should be updated online to adapt to the new appearance change of the target. Our main contributions are:

- Proposal of a novel dual CNN architecture, which learns the generic tracking features and is also able to measure the similarity between image patches.
- We maintain state-of-the-art performance while considerably increasing tracking speed in benchmark dataset.