

Recurrent Tubelet Proposal and Recognition Networks for Action Detection

Dong Li¹, Zhaofan Qiu¹, Qi Dai², Ting Yao³, and Tao Mei³

¹ University of Science and Technology of China, Hefei, China

² Microsoft Research, Beijing, China

³ JD AI Research, Beijing, China

{dongli1995.ustc, zhaofanqiu, tingyao.ustc}@gmail.com

qid@microsoft.com, tmei@live.com

Abstract. Detecting actions in videos is a challenging task as video is an information intensive media with complex variations. Existing approaches predominantly generate action proposals for each individual frame or fixed-length clip independently, while overlooking temporal context across them. Such temporal contextual relations are vital for action detection as an action is by nature a sequence of movements. This motivates us to leverage the localized action proposals in previous frames when determining action regions in the current one. Specifically, we present a novel deep architecture called Recurrent Tubelet Proposal and Recognition (RTPR) networks to incorporate temporal context for action detection. The proposed RTPR consists of two correlated networks, i.e., Recurrent Tubelet Proposal (RTP) networks and Recurrent Tubelet Recognition (RTR) networks. The RTP initializes action proposals of the start frame through a Region Proposal Network and then estimates the movements of proposals in next frame in a recurrent manner. The action proposals of different frames are linked to form the tubelet proposals. The RTR capitalizes on a multi-channel architecture, where in each channel, a tubelet proposal is fed into a CNN plus LSTM to recurrently recognize action in the tubelet. We conduct extensive experiments on four benchmark datasets and demonstrate superior results over state-of-the-art methods. More remarkably, we obtain mAP of 98.6%, 81.3%, 77.9% and 22.3% with gains of 2.9%, 4.3%, 0.7% and 3.9% over the best competitors on UCF-Sports, J-HMDB, UCF-101 and AVA, respectively.

Keywords: Action Detection · Action Recognition.

1 Introduction

Action detection with accurate spatio-temporal location in videos is one of the most challenging tasks in video understanding. Compared to action recognition, this task is more difficult due to complex variations and large spatio-temporal search space. The solutions to this problem have evolved from handcrafted feature-based methods [18, 34, 40] to deep learning-based approaches [7]. Promising progresses have been made recently [22, 28, 36, 39] with the prevalence of deep Convolutional Neural Networks (CNN) [10, 16, 30].

