

Learning Spatiotemporal Attention for Egocentric Action Recognition

Minlong Lu^{1,2} Danping Liao³ Ze-Nian Li¹

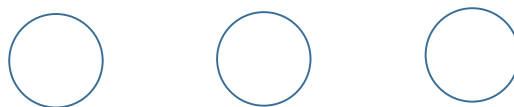
¹School of Computing Science, Simon Fraser University, Canada

²Huawei Technologies, Canada

³College of Computer Science and Technology, Zhejiang University, China

Abstract

Recognizing camera wearers' actions from videos captured by the head-mounted camera is a challenging task. Previous methods often utilize attention models to characterize the relevant spatial regions to facilitate egocentric action recognition. Inspired by the recent advances of spatiotemporal feature learning using 3D convolutions, we propose a simple yet efficient module for learning spatiotemporal attention in egocentric videos with human gaze as supervision. Our model employs a two-stream architecture which consists of an appearance-based stream and motion-based stream. Each stream has the spatiotemporal attention module (STAM) to produce an attention map, which helps our model to focus on the relevant spatiotemporal regions of the video for action recognition. The experimental results demonstrate that our model is able to outperform the state-of-the-art methods by a large margin on the standard EGTEA Gaze+ dataset and produce attention maps that are consistent with human gaze.



1. Introduction

With the increasing popularity of wearable cameras, there is a growing interest in recognizing actions using the first-person/egocentric videos, which has potential applications including remote assistance, health monitoring and human-robot interaction. The wearable camera is usually mounted on the person's head with its optical axis aligned with the wearer's field of view. Action recognition for the camera wearer using the first-person videos is different from that in third-person setting. First, unlike in the third-person video, the camera wearer's pose are mostly unavailable in egocentric videos. The recognition of egocentric actions often requires more fine-grained discrimination of the objects being manipulated and their locations. Second, strong ego-motions are often present in egocentric videos due to the head motion of the person, whereas the third-person videos are usually static or more stable. These aspects make action recognition in egocentric videos very challenging.

