Action Recognition Thesis - WIP

Nicolas Fleece
University of Ottawa
75 Laurier Ave. E, Ottawa, ON K1N 6N5
nflee092@uottawa.ca

Abstract

The ABSTRACT is to be in fully justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

2. Related Work

2.1. CNN-Based Action Recognition

CNN's have been widely used for many years on the problem of human action recognition. They have shown to be capable of detecting actions performed by people in simple and complex settings. This quite often involves the use of 3D-Convolutions, which operate on multiple frames of video at a time. Resnet-3D is an example, and is used later in this paper.

2.1.1 Person-Based Action Recognition

2.2. Pose-Based Action Recognition

Pose involves extracting the skeleton of the person and using this data over multiple frames of a video to classify an action. Pose is a common addition used in action recognition as it relates most to how humans view actions and the movement of different bones.

2.2.1 Intermediate Representations

The approach of the majority of this thesis involves creating intermediate representations for pose data over multiple frames. This typically has the aim of creating some kind of image that represents either the motion of the persons bones and/or joints through the image at different points in the video. These images can then be used either by the model independently or added to traditional two-stream architectures.

The advantage of these types of representations is that the model can quite often be a small CNN that can be trained end-to-end very quickly and with little memory. This quite often allows for real-time evaluation and in some cases mobile-capable models.

2.2.2 PoTion

Pose motion representation for action recognition [1] was largely the inspiration for the work that was done within this thesis. This approach aims to take the joints extracted from the pose representation and use the movement over f frames, creating j images where j is the number of joints.

The approach begins by extracting j joint heatmaps from each frame of the video, these individual frames are then combines using their colour coding where depending on what time t the frame is at in the video, the joint heatmap is made to be that colour. They then perform their temporal aggregation where for each joint j, they combine all frames together into one image, perfoming a simple addition through all frames. This leaves an image that demonstrates the movement of one joint through all frames of a video.

2.2.3 PA3D

Pose action 3D [2] is a similar approach to PoTion, where it involves the use of the generated joint heatmaps from pose estimation models. The difference is that instead of using the color coding similar to potion, PA3D stacks the joint heatmaps such that they create j cubes of every heatmap frame.

2.2.4 Simple yet efficient real-time pose-based action recognition

References

[1] Vasileios Choutas, Philippe Weinzaepfel, Jerome Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. pages 7024–7033, 06 2018. 1

[2] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Pose-action 3d machine for video recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7914–7923, 2019. 1