Action Recognition Model with First-Person Videos: Final Report

G011 (s1884147, s1802373, s1260269)

Abstract

Human action recognition research is mostly focusing on the third-person vision data. There are some approaches for first-person action recognition, but they are heavily problem-dependant. In this project, we evaluated third-person action recognition methods with first-person datasets, and compare the differences between the third and first-person methods. We found that third-person methods generate relatively similar performance on some first-person dataset compared to those first-person focused methods.We also showthat the performance of third method methods could be better than that of first person methods when optimised. Furthermore, we propose and study on a new model combining MobileNet and Two-stream Pyramid. Our new model presents remarkable performance with GTEA and DogCentric compared to third-person methods studied in this project.

1. Introduction

Human action recognition from video data is one of the most popular research areas in computer vision. They have various applications such as smart home, robot vision, self-driving cars, and police body-worn cameras. Most of the video data used to be taken from third-person viewpoints, but recently, the number of first-person video (FPV) data and demands to recognise actions from the data are increasing rapidly because of rising popularity of wearable cameras (e.g. Google Glass) and action cams (e.g. GoPro). In fact, depending on how people mount their action cams, or even smart phones, the devices can capture videos from third-person viewpoint as well as first-person viewpoints. Namely, various types of video data can be produced and it is better for action recognition methods to work with any kinds of general data without being biased to a certain viewpoint, or application.

However, most of the popular action recognition datasets consist of third-person video (TPV) because they are usually collected from movies, so the evaluation of action recognition methods are biased to TPV data (Wang et al., 2017; Carreira & Zisserman, 2017; Tran et al., 2018). Most of these popular action recognition methods does not contain evaluations with FPV datasets, because the datasets are usually not popular. In this report, we will call the popular action recognition methods as TPV action recognition methods, since it has not been proved to work generally.

Researchers have proposed some FPV action recognition methods by extracting useful features for FPV recognition (e.g. hand pose of the wearer, gaze, active objects, and ego-motion), but they are again biased to work well with FPV data. Furthermore, much of the FPV action recognition research is problem-dependant. In other words, they assumed a very specific application scenario and tested with the biased dataset. For example, Fathi et al. (2011); Pirsiavash & Ramanan (2012); Matsuo et al. (2014); Singh et al. (2016) have proposed daily living FPV action recognition methods. The datasets they have evaluated consist of daily actions (e.g. cooking), so most of the sample data do not have any moving objects (e.g. humans), enabling the methods to focus mostly on the hand features.

Some of the first-person action recognition research tried to extract more features using more sensors other than a camera. For example, Ma et al. (2016); Fathi et al. (2012); Li et al. (2015) use gaze features obtained with eye tracking devices, and Garcia-Hernando et al. (2017) uses depth features as well as RGB colour video (i.e. RGB-D). However, we neglect comparison among the result with those kinds of research as there are more auxiliary information from sensors compared with pure video-based recognition.

We train and test TPV recognition models with FPV action datasets, and compare the result with FPV recognition models using only videos as inputs (i.e. no additional sensor data). In addition, we report a performance of an existing model with some modifications to make it work well in general data.

Experiments with FPV methods using TPV data are not deployed since all of the FPV methods assume that the video will have ego-features such as hand segments and ego-motion, which is often not existing in TPV data. General deep-learned action recognition models (i.e. TPV recognition methods) may work fairly well on even FPV datasets because it is supposed to learn relevant and crucial deep feature extractors from the input data no matter what kind of view they have.

The structure of the sections below is as follows. Section 2 describes the core tasks in our project and datasets with which we implemented training and test. Details on methodology we adopted in our research is described in section 3. In section 4, we present the experiment settings and results as well as the analysis on results, and lastly we provide a conclusion and possible future work in section 5 of the report.