**Video Action Recognition**

**Introduction 600**

Human action recognition is a process of examining and assigning a label to an action given a video sequence. In the past few years, HAR has been a dynamic research topic in many areas, such as human-robot interaction, video surveillance, and patient monitoring. Human action is a physical movement of one or more body parts. Activity can be described as the purpose of a series of human actions. For example, we can consider “take the cup”, “pour water into mouth”, and “swallow the water” as a series of actions, the activities of such actions is “drink water”. Given a series of human actions, the viewer can easily recognize the nature of each action and could often predict the next action with certain confidence. What if the viewer knows the activity of the action series? It is highly likely that the viewer would predict the next action with higher confidence. In real life, it is too expensive to tackle most of action recognition and prediction tasks by human. However, machines can deal these tasks efficiently. On the other hand, the variety of the dataset often challenges the performance of using machines. One obvious example would be the variety of people presents in each video, different people would act in different action orders and each action would take different execution time. Such challenges are considered in the experiment**. (segment)**

Past prior works has researched in to varies topic to break through the current achievement in human action prediction.

Most human actions are goal-directed, where the goal can be seen as activity or the ground truth label of the video sequence. In this research, we use the breakfast dataset that closely reflects the daily life human actions. The effect of use of the ground truth label is to date still a fresh topic in human action prediction. This paper will look into the role of activity in relation to the performance of action recognition.

In the following sections of this paper, we firstly review some recent works in improving the performance of human action prediction tasks as well as the use of activity in the prediction process, in Chapter 2. We describe the dataset used in the research in Chapter 3. We present the methodologies of the experiment in Chapter 4. We then describe the models used in Chapter 5. After that, we show the experimental result in Chapter 6. Last, we conclude the paper in Chapter 7.

**Related Works 500**

**Dataset 500**

The breakfast dataset in used in our experiment, which contains 1127 samples and each sample is a video sequence. There are 52 participants, and each accomplishes 10 unity cooking activities, which enhances the randomness of the dataset and is to better emulate the real-world environment.  Activities are achieved by 48 unique actions including the “silence” action, seen as the “SIL”. Due to the number of actions in the dataset, we use one example to further explain the relationship between activity and action. One of the cooking activities is “the preparation of making a bowl of cereal”, the related actions could be done for this activity are “take bowl”, “pour cereals”, “pour mild”, “stir cereals” without a certain order.

Each video sequence starts and ends with the one or more frames with “SIL” action. Each action could be represented in multiple frames depending on the execution time of the participant. The dataset also includes a mapping of actions is to its integer representation, each action is encoded by a unique integer from 0 to 47 and a video sequence could be look like [0, 0, 0, 0, 1, 1, 2, 2, 2, 3, 3, …… 0, 0, 0]. 52 participants are labelled as P03 to P54, and there is only one participant present in each video sequence.

|  |  |  |
| --- | --- | --- |
| 1 | coffee | n=200 |
| 2 | juice | n=187 |
| 3 | chocolate milk | n=224 |
| 4 | tea | n=223 |
| 5 | bowl of cereals | n=214 |
| 6 | fried eggs | n=198 |
| 7 | pancakes | n=173 |
| 8 | salad | n=185 |
| 9 | sandwich | n=197 |
| 10 | scrambled eggs | n=188 |

Split

We used four-fold cross validation during the training and testing process, the dataset is split by participants. As described in the dataset section, participants are labelled from P03 to P54, the four split is shown in table.

|  |  |
| --- | --- |
| Split 1 | P03 – P15 |
| Split 2 | P16 – P28 |
| Split 3 | P29 – P41 |
| Split 4 | P42 – P54 |

**Pre-processing**

In this section, we describe the pre-processing approaches of the dataset. Firstly, we removed all “SIL” actions in all sample because it is a silence action and only appears at the beginning and the end of the video sequence. Moreover, the “SIL” action does not contribute to any process of an activity, thus the existence of the “SIL” action does not offer more information of the video sequence.

One of the challenges is the spatiote spatiotemporal variations

les. Although there are 51 participants been recorded, test video sequence with unseen participants may not work. Because different person has different habits and styles even doing the same action or activity. Annotating more samples could help, but it will take time and human labours to accomplish. Our approach is to segment the video actions in order to decrease the effect of the data variation to its performance on unseen test videos. Therefore, action segmentation is adopted during this stage. Each action in the dataset is mapped to a pre-defined integer label and the process of action segmentation is to assign a pre-defined label to each segmented action sequence. For example, we could convert a video sequence [1,1,1,2,2,2,2] to [1,2]. The resulted action segments are inputted to the model.

**Model 2000**

**Result 1600**

**Conclusion 600**

(segment)

B. M. Hard, G. Recchia, and B. Tversky. The shape of action. Journal of experimental psychology: General, 140(4):586– 604, Nov. 2011. 2

youtube

https://openaccess.thecvf.com/content\_CVPR\_2020/papers/Chen\_Action\_Segmentation\_With\_Joint\_Self-Supervised\_Temporal\_Domain\_Adaptation\_CVPR\_2020\_paper.pdf