# **Action Recognition**

**Problem Statement:**

Given a set of images of a video the work aims to analyze the different actions in the video. The performance of the work is evaluated using performance parameters like precision

**Dataset Details:**

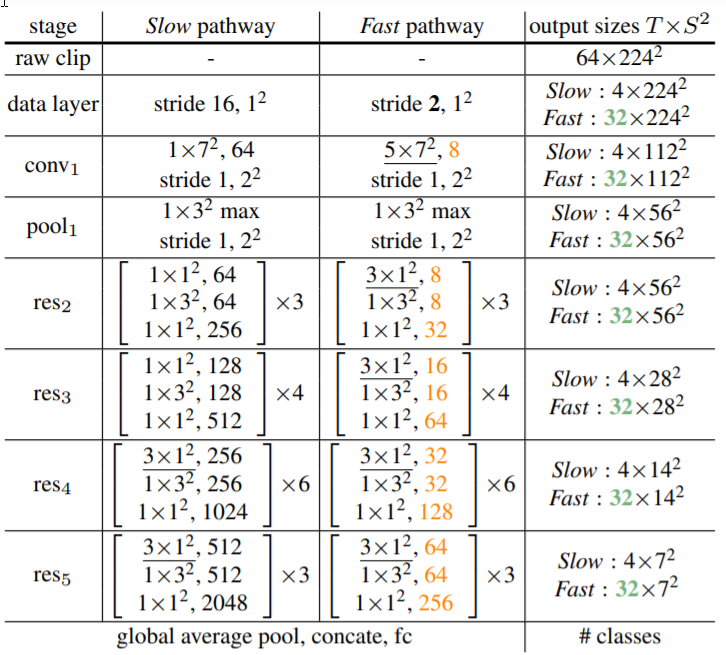
The dataset contains 400 human action classes, with at least 400 video clips for each action. Each clip lasts around 10s and is taken from a different YouTube video. The actions are human focussed and cover a broad range of classes including human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands. We describe the statistics of the dataset, how it was collected, and give some baseline performance figures for neural network architectures trained and tested for human action classification on this dataset. We also carry out a preliminary analysis of whether an imbalance in the dataset leads to bias in the classifiers.

**Description of the Scene:**

Details of the scene with images and the type of actions:

|  |  |  |  |
| --- | --- | --- | --- |
| Action Name | Description | Sample Image | Output with probability |
| Riding motorcycle | The video obtained from the camera fixed o the helmet of a person riding a motorcycle. |  | [motorcycling] 0.953.  [pushing wheelchair] 0.023.  [driving car] 0.015.  [pumping gas] 0.003.  [texting] 0.002. |
| Climbing a tree | The video of a man is climbing a tree by stepping on its branches. |  | [climbing tree] 0.994.  [trimming trees] 0.004. |
| Riding a bicycle | The video of a kid riding a bicycle on a road |  | [motorcycling] 0.809.  [riding a bike] 0.145.  [riding unicycle] 0.012. |
| Playing the piano | The video of a person playing the piano. |  | [playing recorder] 0.658.  [shuffling cards] 0.046.  [playing trumpet] 0.030. |
| Javelin Throw | The video of a person throwing a javelin |  | [javelin throw] 0.998.  [pole-vault] 0.001. |
| Riding a scooter | The video of a small kid riding a toy scooter |  | [riding scooter] 0.895.  [hoverboarding] 0.105. |

**Deep Learning Architecture:**



SlowFast networks for video recognition. This model involves

1. A Slow pathway, operating at a low frame rate, to capture spatial semantics, and a Fast pathway, operating at a high frame rate, to capture motion at fine temporal resolution. The Fast pathway can be made very lightweight by reducing its channel capacity, yet can learn useful temporal information for video recognition.

**Inference:**

From the results, we can infer that the actions are being recognized on the abstract level.

SlowFast model gives an mAP of 42.5 for the kinetic400 dataset. The accuracy for each sample image has been shown in the table above.

**References:**

1. [Literature Survey: Human Action Recognition](https://towardsdatascience.com/literature-survey-human-action-recognition-cc7c3818a99a) (Towards Data Science)
2. [SlowFast model](https://arxiv.org/pdf/1812.03982.pdf)
3. [Kinetic400](https://arxiv.org/pdf/1705.06950.pdf)