5.1 Dataset Update

The dataset that will be used in this study for model training is the MRL Eye Dataset(cite). It in itself a large-scale dataset filled with human eye images. The dataset was created as the team behind it has been involved in solving tasks in the behaviour of driver behaviour. These tasks do not exclude eye detection, gaze estimation and eye-blinking frequency.

5.2 Data Collection

The dataset itself contains 84898 images of eyes. According to the website, the images in the dataset were captured through infrared means, in both low and high resolution. The eye detector used to obtain the images of the eye were based on a histogram of oriented gradients (HOG) combined with an SVM classifier. The images were also captured in various lighting conditions as well as taken by different devices. The data was collected based on 37 different people, 34 of which were men and the remaining 4 were women. The participants included people who wore and did not wear glasses. The dataset includes images of both open and closed eyes of approximately equal amounts, which prevents oversampling issues. The images also varied in terms of quality of images, which was based on the amount of reflection occurred in the image.

5.3 Data Description

The images produced from this collection were annotated according to the following properties, in the specific order;

subject ID;

image ID;

gender [0 - man, 1 - woman];

glasses [0 - no, 1 - yes];

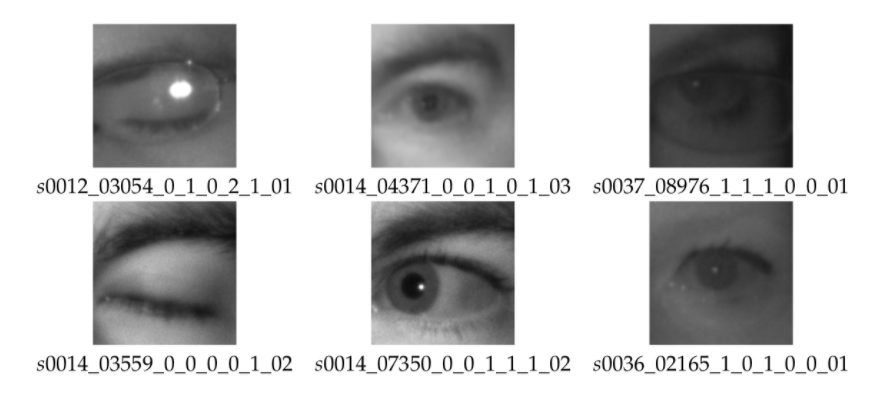
eye state [0 - closed, 1 - open];

reflections [0 - none, 1 - small, 2 - big];

lighting conditions [0 - bad, 1 - good];

sensor ID [01 - RealSense, 02 - IDS, 03 - Aptina];

An example of the annotations is as shown below;



5.4 Data Preprocessing

The data has to be arranged and cleaned if possible before it could be fed into a classifier. The data was primarily cleaned using Python. First the data had to be split into two categories; open and closed eyes. This was done by manually sifting through the folders in the dataset and dividing them accordingly. The reason this was done manually was so that imperfections such as wrongly detected facial parts could be filtered and removed. The two folders, once segregated, each contained an approximate of 42000 samples, which negates the issue of oversampling.

Next the data had to be further pre-processed using Python. The data was first loaded into a 2-dimensional array. The first column stored the images (attributes), which had been saved in grayscale and resized to 50x50 for consistency. The second column stored a number which corresponded to their state (0=Open,1=Close). This was to be the base structure of the dataset for this study. At this stage, the data contained 84606 samples.

Finally, the data was shuffled as to avoid overtraining for a specific class and split into its “X” and “y” components. The two arrays were then “pickled” using the python library *pickle* which saves the data to be loaded once again for model training purposes.