Driver Monitoring Dataset Guide

Last modification: 19/01/2024

1. Introduction

The DMD stands for Driver Monitoring Dataset. This dataset was built due to the lack of data for training and testing driver monitoring systems based on computer vision. It contains data for tasks like distraction and drowsiness detection and gaze estimation, making it possible to do a temporal or discrete multi-sensor analysis since the dataset is in video format with the information from 3 cameras, each recording RGB, IR and depth data. All of this is available for download.

Processing a big amount of information requires time, especially annotating. As we progressed on the data, we published more material. All the videos have been exported, but some work is still to be done around annotations. Right now, the focus is on temporal annotations. We are still working to offer more.

There is a GitHub repository with tools and documentation you might find useful, please check the Wiki. There is a new Known Issues section in the readme.

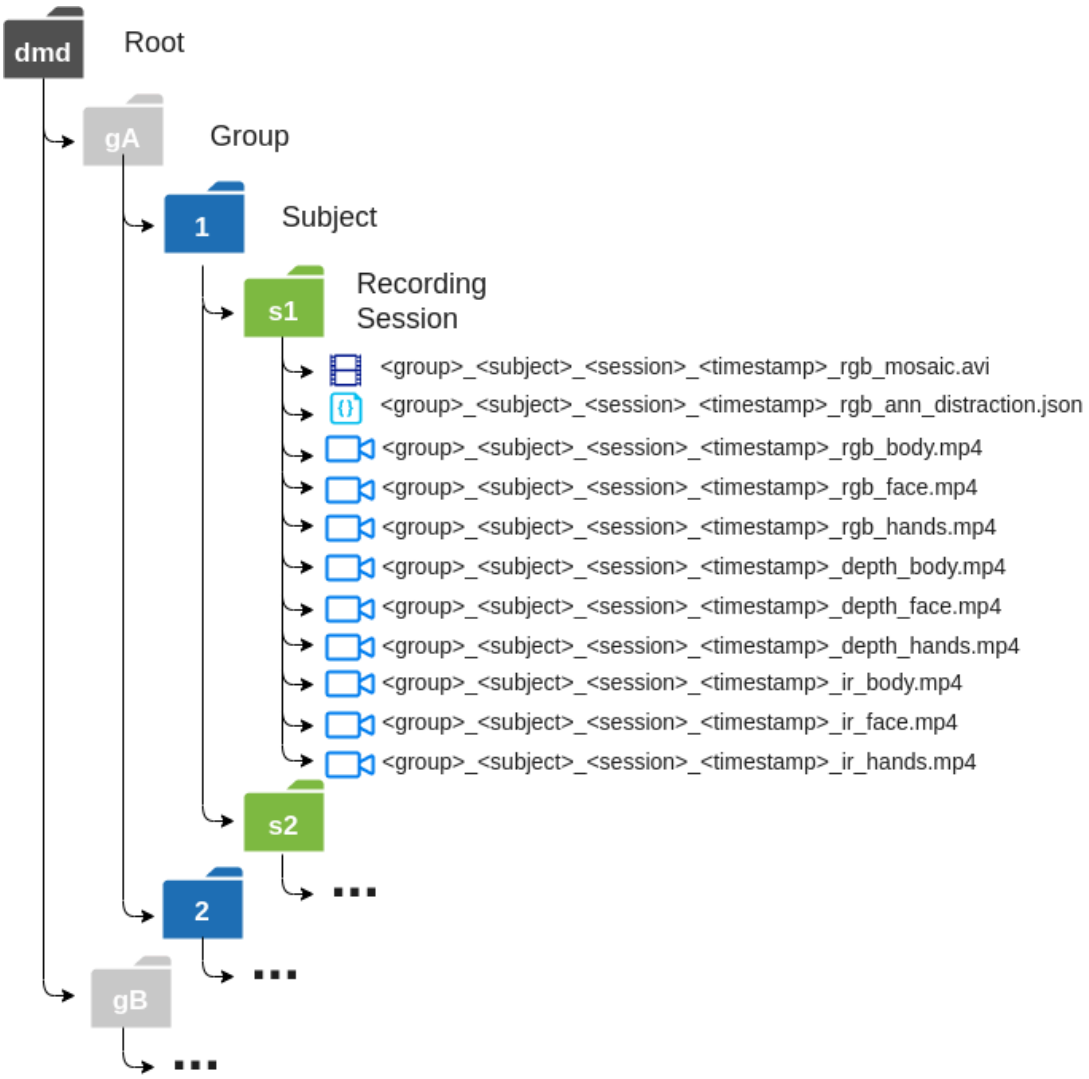
Check our website to read our scientific papers and don't forget to cite us!

This document hopes to be a guide for the understanding and use of the DMD.

2. Structure

You now have access to the DMD. The material is organized to be easily used, but it can be confusing. As shown in the Figure below, the material is first divided into groups of subjects (gA, gB, gC, etc.). Inside each group, there is the material of about 5 participants. Inside a subject folder, the material is organized by recording sessions (s1, s2, s3, etc.). These are explained in Section 3.1. Finally, inside each recording session folder are the videos (three cameras, three streams and mosaic video) and the annotation file. All of the files follow the same nomenclature as shown in the Figure.

** There was a mistake on the IR videos. They were uploaded in .avi format; the correct format is .mp4. By 17/01/2024 this problem is fixed.



3. Dataset Specifications and Set-up

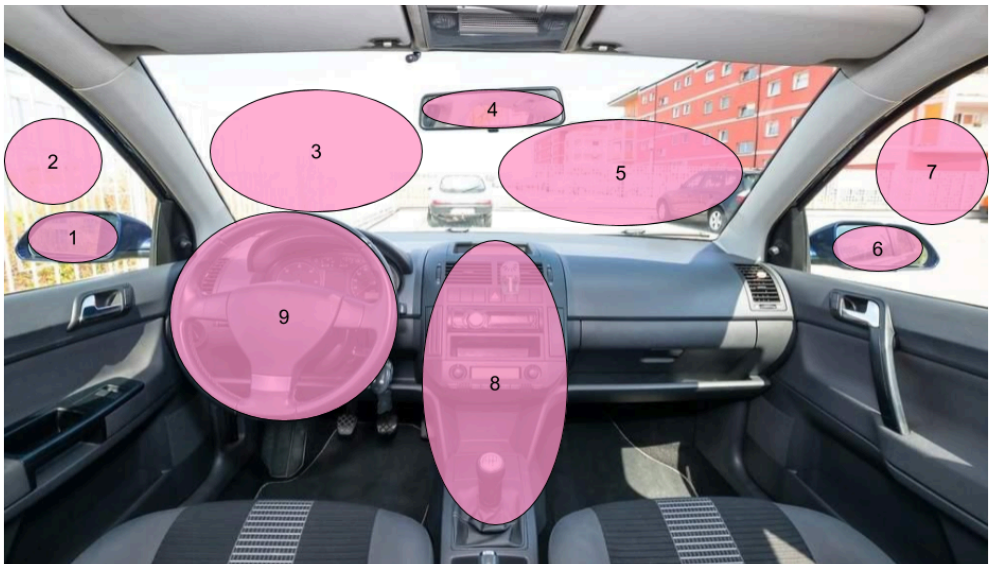
3.1. Recording Sessions

A protocol was created for the recordings, trying to capture different activities with different conditions in every recording. These were made continuously; this means that after performing one activity, the driver starts a new one, as realistically as possible, until the completion of all the activities within a protocol. There were many recording sessions, each covering different activities and/or conditions. The material is presented in this same way, organized by recording sessions.

Below is the description of each recording session, including a list of the recorded activities. Note that the activities are not equal to the list of annotations.

- **S1:** This protocol was recorded in the real car while the driver was driving. There are activities related to **distraction** that could be legally recorded while driving and were not dangerous. The activities are:
    - Safe driving
    - Reaching side
    - Operating the radio
    - Drinking
    - Talking to passenger
- **S2**: In these sessions, there are more activities related to **distraction** recorded with an unmoving real car; the driver acted the driving task. The activities are:
    - Safe driving
    - Reaching side
    - Hair and makeup
    - Talking on the phone - right
    - Texting - right
    - Talking on the phone - left
    - Texting - left
- **S3**: This is also recorded in an unmoving car, and it is the shortest recording session. It was created for one **distraction** activity, which is grabbing something from the back seat.
- **S4**: This recording session is similar to S2, also **distraction** activities, but it is recorded in the driving simulator. The activities are:
    - Safe driving
    - Reaching side
    - Hair and makeup
    - Talking on the phone - right
    - Texting - right
    - Drinking
    - Talking on the phone - left
    - Texting - left
- **S5**: In this session, all the drowsiness activities are recorded. By the time of the writing of this document, this was only performed in an unmoving real car, not the simulator. The activities are:
    - Safe driving
    - Sleepy driving
    - Yawn no hand
    - Yawn with hand
    - Microsleep
- **S6**: Lastly, this recording session contains the material for **gaze & hands** position estimation. The participants were asked to look at specific regions of the car and the simulator and then to place their hands as asked, alternating from quiet and moving hands. Inside this folder, **there is data from the real car and in the simulator**. Take this into account in case you want to analyze them separately. The activities were:
    - Stare at nine regions (shown in the figure below)
    - Both hands on (quiet)
    - Right hand on (quiet)

- Left hand on (quiet)
- Both hands off
- Both hands on (moving)
- Right hand on (moving)
- Left hand on (moving)



3.2. Recording environments

For this project, two environments were prepared to accomplish the recordings for the dataset: a real car and a simulator. The performance of some of the activities that wanted to be included in the dataset, in a real scenario, could be somehow illegal or dangerous to perform. For example, is not legal to text while driving. Hence, to still have the real experience and performance of the activities, it was proposed to have both a properly equipped car and a simulator.

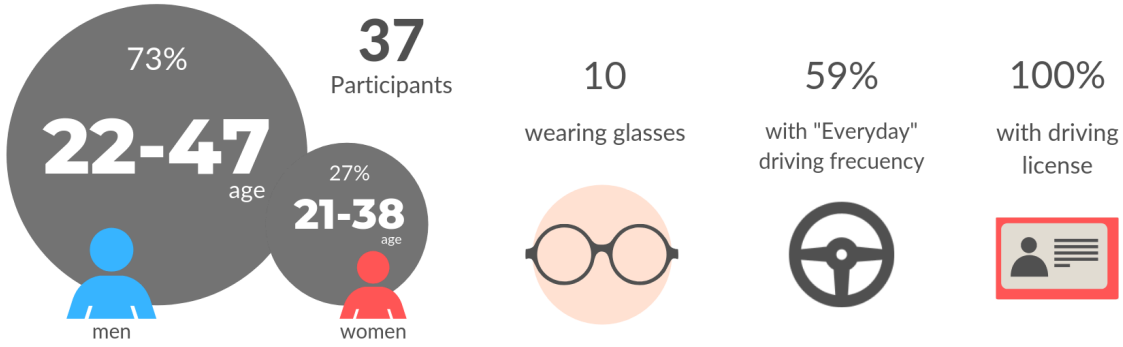
3.3. Cameras

For the DMD, we used three devices installed in both environments, the car and simulator, and placed them correctly to capture images of the driver's face, body and hands. All cameras recorded RGB, IR and depth information.

All the devices are configured to work with a considerable resolution and exposure level without risking the frame rate. This way, the images obtained fit the dataset requirements. The resolution was set to 1280x720 px in all channels. This resolution is offered in the videos, even if it implies heavier files.

3.4. Participants

37 experienced drivers volunteered for the project, as represented in the figure below.



All participants were grouped into six groups of 5 and one of 7 people. Knowing the availability of the participants, first, they were organized by groups depending on the recording sessions they could attend (morning, afternoon or both). Then, a double participation schedule was assigned to 3 groups of volunteers whose availability was "both". All the protocols were recorded twice for these participants, one in the morning and another in the afternoon, not exactly on the same date. This is because the variability in lighting was important. Therefore, you will find groups that have a double size of data. 20% of the data is reserved for possible benchmarking purposes and will not be public.

4. Dataset Post-Processing and Annotation

4.1. Raw to videos

First, each frame was exported into an image in png format. Images were converted into video files using H.264 codec (libx264 library). The videos are created at the rate they were originally recorded with. For the body camera, the fps was 29.98, and the face and hands camera was 29.76, so the videos are created with those fps.
For the next step, the frame rate of all videos must be the same. Therefore, the actual body video is processed with FFMPEG to have 29.76fps, creating another video with this frame rate.

4.2. Align the videos

The next thing in the process is to align the three perspective videos. The cameras did not start recording simultaneously, which is why there is a shift between each video. With a script, it was possible to synchronize the cameras. As a result, we obtained the shifts between the three cameras. This way, knowing the starting order of the videos and their offsets between them, a mosaic (see Figure below) is created in which the three cameras are aligned. These mosaics are available in the dataset, and it was used to annotate.

4.3. Annotation

A tool (TaTo) was developed to annotate temporal actions in the DMD dataset.

The DMD annotations are in ASAM OpenLabel description format based on a JSON Schema. To create annotations, the **VCD (Video content description)** library was used. It can include descriptions of actions, objects, relationships, and events, all in one file. This library is compatible with the standard ASAM OpenLabel labelling format. More information about VCD can be found here.

The DMD will have one file per dimension (distraction, drowsiness, gaze and hands). In the case of annotating in another format, they must be temporal annotations, meaning they are time intervals or can be derived into time intervals. This way, a converter to OpenLabel can be built after the annotation process finishes.

On the other hand, a tool to access the annotations easily was built. It is called DEx Tool. It offers to prepare DMD data for training. There are configuration variables that can be changed to extract specific material.

If you wish, you can also get the annotations from the .json file. It will be a little bit confusing, but inside the file, there is information about each frame (inside frames object). Each frame in the frame object list shows the ID of the action that is present in that frame. To know the name of the action, you can go to the actions object list in the JSON file, and you will find the name in the type field. Also, each action object has the frame intervals in which that action is present, in the frame_intervals field. This way, annotations can be accessed by frame or by action. For a better understanding, check the documentation of the OpenLabel format.

To perform the temporal action annotation of the DMD we have defined an **annotation criterion** that has to be followed by all annotators to guarantee consistent annotation production. Here, the list of annotations is presented

Check the documents for more details:

- Distraction: DMD Distraction-related actions criteria
- Drowsiness: DMD Drowsiness-related actions criteria
- Gaze: DMD Gaze-related actions criteria
- Hands: DMD Hands-related actions criteria