An fNIRS dataset for driving risk cognition of passengers in highly automated driving scenarios

Xiaofei Zhang1,+, Qiaoya Wang1,+, Jun Li1 , Xiaorong Gao2, Bowen Li2, Bingbing Nie1 , Jianqiang Wang1 , Ziyuan Zhou1, Yingkai Yang3, Hong Wang1,\*

1. School of Vehicle and Mobility, Tsinghua University, Beijing, 100084, China
2. School of Medicine, Tsinghua University, Beijing, 100084, China
3. Department of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ

\* corresponding author(s): Hong Wang (hong wang@mail.tsinghu)

+ these authors contributed equally to this work

Table of Contents

1 Supplementary Notes 1

2 Supplementary Example 1

2.1 Data preprocessing 1

2.2 Data content 2

2.3 Data analysis 4

# 1 Supplementary Notes

This dataset contains the raw intensity data which can be changed to and based on the modified Lambert-Beer law, and driving scenario data. We have adjusted our data to be BIDs-complaint, and placed the data at OpenNeuro. its DOI <doi:10.18112/openneuro.ds004973.v1.0.1> Homer3 is an open-source software toolkit used for fNIRS, we provide one example about data preprocessing using Homer3 and data analysis using machine learning. Besides, we also provide four functions in Python to process the raw data and , one function which can transform MNI coordinate to Talairach coordinate, and one project file which can rebuild those fourteen types of highly automated driving scenarios using VTD software, those related content have been update at github <https://github.com/benchidefeng/fNIRS-experiment-for-automated-driving-scenarios.git>

# 2 Supplementary Example

In this example, the data of each scenario was divided into low-risk and high-risk episodes in accordance with one split point (stimulus point). The obtaining process of this data is shown in Fig 1.

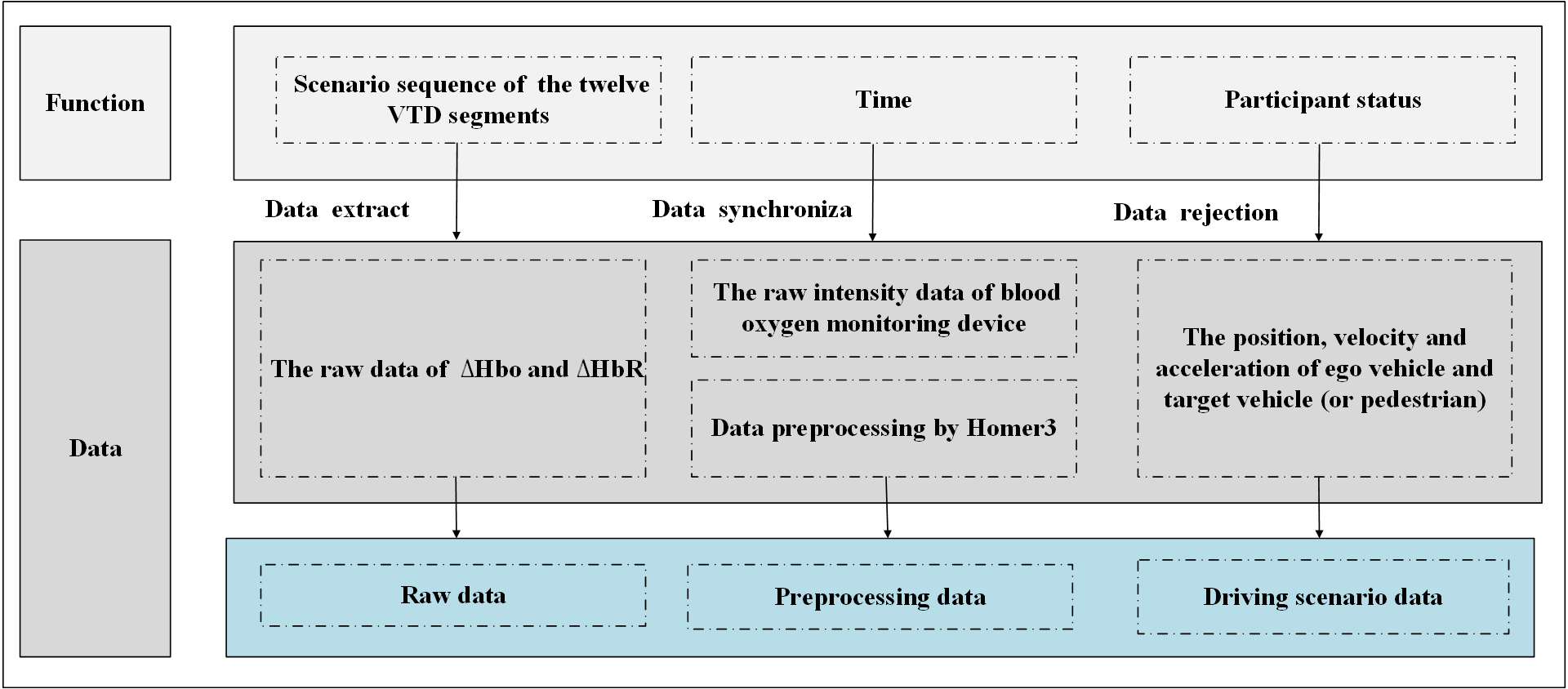


Fig 1 The obtaining data process of this example

Firstly, we judge whether or not the participants were focusing on the given task by comparing the time delays at which participants pressed the keyboard when they heard a stimulating sound, and we accordingly reject those data that were collected when participants were not focusing on those task. Then the remaining experimental data is used for driving risk cognition of passengers based on the type of highly automated driving scenario. In these experiments, few data is removed because the data are not recorded correctly, and the detailed information about the total number of data points and the total number of valid points in each of the fourteen types of highly automated driving scenarios are shown in Table 1.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Number | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 | Scenario 6 | Scenario 7 |
| Total | 180 | 280 | 260 | 560 | 500 | 440 | 280 |
| Valid | 170 | 263 | 242 | 507 | 478 | 417 | 261 |
| Number | Scenario 8 | Scenario 9 | Scenario 10 | Scenario 11 | Scenario12 | Scenario13 | Scenario 14 |
| Valid | 380 | 380 | 660 | 620 | 380 | 500 | 340 |
| Total | 358 | 350 | 614 | 581 | 339 | 470 | 319 |

Table1 Total number and valid number about fourteen types of highly automated driving scenarios

Secondly, The kinetic energy field indicates a danger degree of driving scenario, and it involves relative longitudinal distance and the speed of target vehicle (or pedestrian), and it is shown in (1). In those experiments, the kinetic energy field is adopted as an objective indicator for indicating the danger degree of driving scenario, and the value 0.05 of the kinetic energy field is chosen as the split point based on experience. The raw intensity data from this blood oxygen monitoring and driving scenario data is divided into low-risk and high-risk episodes according to this split point.

（1）

where, , , and are five constants, , , , , and . The velocity of target vehicle (or pedestrian) is denoted by , represents the distance between ego vehicle and target vehicle (or pedestrian), and denotes the angle between and .

Finally, the raw intensity data from this blood oxygen monitoring is preprocessed using Homer3, then the raw and , the preprocessing and and driving scenario data at low-risk and high-risk episodes are available.

## 2.1 Data preprocessing

In order to remove respiration, heart rate, blood pressure fluctuations, Mayer waves noises and others noises. We did preprocessed for the raw intensity data of blood oxygen monitoring device by Homer3, and obtained the preprocessed data of and . The preprocess contains removing motion artifact operation and filtering operation, and the functions and parameters which were employed are: hmrR\_PruneChannels(dRang: 0.01-1, SNRthresh: 2, Sdrange:10.0,45.0), hmrR\_Intensity2OD, hmrR\_MotionArtifactByChannel(tMtion:0.5, tMask: 1.0, STDEVthresh: 5.0, AMPthresh: 0.05), hmrR\_MotionCorrectWavelet(iqr: 1.50, turnon: 1), hmrR\_BandpassFilt:Bandpass\_Filter\_OpticalDensity(hpf: 0.015, lpf:0.085), hmrR \_OD2Conc(ppf: 1.0, 1.0) and hmrR\_MotionCorrectCbsi (turnon: 1). The preprocess is shown in Fig 2.

Fig 2 The preprocessing process of the raw intensity data of blood oxygen monitoring device by Homer3

## 2.2 Data content

The data of one participant in one highly automated driving scenario is stored in a csv file. The data of one participant in same highly automated driving scenario is stored in a subfolder. The csv file which contains the raw data and and the csv file which contains the preprocessing data of and by Homer3 were stored in the folder of “RawfNIRSDataset”, “PreprocessingfNIRSDataset” respectively. The details of their csv files are shown Table 2.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Raw data CSV | | | | Preprocessing data CSV | | | |
| Raw | Content | Unit | Sampling rate | Row | Content | Unit | Sampling rate |
| 1 | The of channel 1 | mol/L | 100Hz | 1 | The of channel 1 | M mol/L | 100Hz |
| 2 | The of channel 1 | mol/L | 100Hz | 2 | The of channel 1 | M mol/L | 100Hz |
| 3 | The of channel 2 | mol/L | 100Hz | 3 | The of channel 2 | M mol/L | 100Hz |
| 4 | The of channel 2 | mol/L | 100Hz | 4 | The of channel 2 | M mol/L | 100Hz |
| 5 | The of channel 3 | mol/L | 100Hz | 5 | The of channel 3 | M mol/L | 100Hz |
| 6 | The of channel 3 | mol/L | 100Hz | 6 | The of channel 3 | M mol/L | 100Hz |
| 7 | The of channel 4 | mol/L | 100Hz | 7 | The of channel 4 | M mol/L | 100Hz |
| 8 | The of channel 4 | mol/L | 100Hz | 8 | The of channel 4 | M mol/L | 100Hz |
| 9 | The of channel 5 | mol/L | 100Hz | 9 | The of channel 5 | M mol/L | 100Hz |
| 10 | The of channel 5 | mol/L | 100Hz | 10 | The of channel 5 | M mol/L | 100Hz |
| 11 | The of channel 6 | mol/L | 100Hz | 11 | The of channel 6 | M mol/L | 100Hz |
| 12 | The of channel 6 | mol/L | 100Hz | 12 | The of channel 6 | M mol/L | 100Hz |
| 13 | The of channel 7 | mol/L | 100Hz | 13 | The of channel 7 | M mol/L | 100Hz |
| 14 | The of channel 7 | mol/L | 100Hz | 14 | The of channel 7 | M mol/L | 100Hz |
| 15 | The of channel 8 | mol/L | 100Hz | 15 | The of channel 8 | M mol/L | 100Hz |
| 16 | The of channel 8 | mol/L | 100Hz | 16 | The of channel 8 | M mol/L | 100Hz |

Table 2 Organization of the content in blood oxygen monitoring device CSV

The information corresponding to the position, velocity and acceleration of ego vehicle and target vehicle (or pedestrian) was contained in a CSV file, and those CSV files were stored in the folder of "VehicleStatusDatase". The CSV file also contains a kinetic energy field and a flag bit, The flag bit indicates the presence of absence of a stimulating sound during the sample time period. If its value is true, then the period contains a stimulating sound. The details of this csv files are shown in Table 3. In this experiment, data is divided into low-risk and high-risk episodes according to one spite point, the data in csv files at front half and later half parts are considered as low-risk data and high-risk data.

|  |  |  |  |
| --- | --- | --- | --- |
| Row | Content | Unit | Sampling rate |
| 1 | The lateral position of ego vehicle | m | 100Hz |
| 2 | The longitudinal position of ego vehicle | m | 100Hz |
| 3  4 | The lateral velocity of ego vehicle  The longitudinal velocity of ego vehicle | m/s  m/s | 100Hz  100Hz |
| 5 | The lateral acceleration of ego vehicle | m/s2 | 100Hz |
| 6 | The longitudinal acceleration of ego vehicle | m/s2 | 100Hz |
| 7 | The lateral position of target vehicle (or pedestrian) | m | 100Hz |
| 8 | The longitudinal position of target vehicle (or pedestrian) | m | 100Hz |
| 9  10 | The lateral velocity of target vehicle(or pedestrain)  The longitudinal velocity of target vehicle (or pedestrain) | m/s  m/s | 100Hz  100Hz |
| 11 | The lateral acceleration of target vehicle(or pedestrain) | m/s2 | 100Hz |
| 12 | The longitudinal acceleration of target vehicle (or pedestrain) | m/s2 | 100Hz |
| 13 | The kinetic energy field28 | kg/s | 100Hz |
| 14 | a flag bit | \* | 100Hz |

Table 3 Organization of the content in highly automated driving scenario CSV. The \* indicates that there is no a unit for the corresponding content.

## 2.3 Data analysis

There are four indexes which can be used to explore mental activities based on fNIRS: , , cerebral blood volume exchange and cerebral oxygen exchange . has be used in many studies. Scenarios 04, 05, 07 and 10 belong to four relatively dangerous scenarios for these three typical single-car scenarios. The difference of passengers’ mental activities between low-risk and high-risk segments were compared based on . Training and testing data were extracted from those low-risk and high-risk according to a four-second window and a one-second step, and those training and testing data was classified using common there machine learning methods which contains two integration algorithm, Adaboost and composite classifier. The classification results are shown in Table 4. Compared with other algorithms, the testing accuracy of AdaBoost algorithm for scenarios 4, 5, and 7 are the highest. Besides, the data of in low-risk and high-risk segments can be distinguished partly by those three machine learning methods.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Scenario | Accuracy | Ada Boost31 | Composite Classifier | RandomForest |
| Scenario4 | Training accuracy | 100% | 96.79% | 100% |
| Testing accuracy | **90.54%** | 76.51% | 87.32% |
| Scenario5 | Training accuracy | 100% | 93.90% | 100% |
| Testing accuracy | **87.35%** | 69.52% | 83.48% |
| Scenario7 | Training accuracy | 100% | 93.66% | 100% |
| Testing accuracy | **85.54%** | 69.60% | 81.27% |
| Scenario10 | Training accuracy | 100% | 97.05% | 100% |
| Testing accuracy | 69.31% | 66.06% | **76.22%** |

Table 4 The classification results of low-risk and high-risk for relatively dangerous scenarios. Composite Classifier is also a integration algorithm, in this paper, it is consists of support vector machine, decision tree, logistic regression and RandomForest, and its result is decided by voting method.