Face2Multi-modal: In-vehicle Multi-modal Predictors via Facial Expressions

Zhentao Huang
scyzh3@nottingham.edu.cn
User-Centric Computing Group,
University of Nottingham Ningbo China

Rongze Li
scyr11@nottingham.edu.cn
User-Centric Computing Group,
University of Nottingham Ningbo China

Wangkai Jin
scywj1@nottingham.edu.cn
User-Centric Computing Group,
University of Nottingham Ningbo China

Zilin Song
User-Centric Computing Group,
University of Nottingham Ningbo China

Yu Zhang∗
School of Computer Science,
University of Nottingham

Xiangjun Peng†
Memory and Storage System Research Group, The Chinese University of Hong Kong

Xu Sun
Faculty of Engineering, University of Nottingham Ningbo China

ABSTRACT
Towards intelligent Human-Vehicle Interaction systems and innovative Human-Vehicle Interaction designs, in-vehicle drivers’ physiological data has been explored as an essential data source. However, equipping multiple biosensors is considered the limited extent of user-friendliness and impractical during the driving procedure. The lack of a proper approach to access physiological data has hindered wider applications of advanced bio-signal-driven designs in practice (e.g. monitoring systems and etc.). Hence, the demand for a user-friendly approach to measuring drivers’ body statuses has become more intense.

In this Work-In-Progress, we present **Face2Multi-modal**, an In-vehicle multi-modal Data Streams Predictors through facial expressions only. More specifically, we have explored the estimations of Heart Rate, Skin Conductance, and Vehicle Speed of the drivers. We believe **Face2Multi-modal** provides a user-friendly alternative to acquiring drivers’ physiological status and vehicle status, which could serve as the building block for many current or future personalized Human-Vehicle Interaction designs. More details and updates about the project Face2Multi-modal is online at https://github.com/unnc-ucc/Face2Multimodal/.

∗ Work was done during the summer research internship at User-Centric Computing Group
† Work was done as a Research Affiliate to User-Centric Computing Group, University of Nottingham Ningbo China

CCS CONCEPTS
• Human-centered computing → Human computer interaction (HCI).

KEYWORDS
Human-Vehicle Interactions; Computer Vision; Ergonomics.

ACM Reference Format:

1 INTRODUCTION
With the emerging practices of autonomous vehicles, the demands for Human-Vehicle Interaction has become more intense. With various techniques emerging into productions, such as Voice Interaction and GUI-based Navigation, the focus of driving experiences has centered on Human-Vehicle Interaction, rather than Vehicle-assisted approaches. More interests, in terms of the next-generation Human-Vehicle Interaction techniques, have come to explore various novel functionalities and services such as health monitoring assistant[20] and etc.

Reliable and user-friendly data sources of drivers’ body status are crucial to adapt various designs into practice [1, 20]. The status of drivers’ body refers to the measurement through a variety of biosensors, which could be further used to examine drowsiness, tiredness, and the emotion of drivers. Current Human-Vehicle Interaction Systems are facilitated with communicative and physical actions[16]. However, with the access to drivers’ body statuses, the designs of Human-Vehicle Interaction Systems could explore higher dimensional perspectives during the decision-making procedure (e.g. warning driver when the heart rate is unstable, and etc.). However, the equipment of multiple biosensors would be extremely not user-friendly in the driving procedure, which has been considered
as the key obstacle to access drivers’ physiological information in practice.

To this end, we present our Work-In-Progress **Face2Multi-modal**, an in-vehicle design to estimate drivers’ multi-modal states (skin conductance and heart rates) and driving status (vehicle speed) through their facial expressions only. Without the burdens of multiple biosensors, our approach is more efficient and user-friendly to acquire drivers’ physiological states. We first brief the system design of **Face2Multi-modal** in Section 2. Then we present quantitative results from our evaluations of **Face2Multi-modal** in Section 3. Finally, we discuss relevant optimization and design spaces, within or enabled by **Face2Multi-modal** in Section 4.

## 2 METHODOLOGY

**Front-end: Face Detector.** To provide the data input for the **Face2Multi-modal**, a camera capture need to be assembled in the vehicle cab to record the facial video of drivers continuously. The recorded videos will be cropped into consecutive frames, and the input is produced by resizing the frames into 224x224px facial images because Neural Network for image classification takes the same size of images as input[9]. OpenCV, an external library of Python, is used to perform these tasks[15].

**Backbone: Neural Network Model.** Neural Network is the backbone of the **Face2Multi-modal**. We select DenseNet as the architecture of neural network because of its strength in traditional image classification tasks (e.g. distinguish different objects), training efficiency and hyperparameter adjustment[9]. Other neural network models (e.g. ResNet[4], SENet[8]), that have excellent performance in image classification could also be adapted in our network design, further studies would find out which one fits the driving context better. Although DenseNet has the aforementioned advantages, in the context of predicting drivers’ status, it might not effective as expected since the inputs are drivers’ highly similar facial captures. Therefore, we apply some lightweight changes in model settings to adapt to the context. We choose 100 to be the depth of our model instead of the suggested depth from the paper, which would reach a higher level[2]. Correspondingly, the Neural Network (e.g. temporal information), the accuracy of the model[3]. Both spatial and temporal illumination variance has occurred in the BROOK database which would harm the accuracy of the model[3]. Both spatial and temporal illumination variance has occurred in the BROOK database which would result in the faulty allocation of pixel values of the skin. Although an approach that takes the background region of each picture as a reference is widely used to reduce the effects, illumination variance is still an obstacle for estimating status by facial captures[10].

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth/Layers</td>
<td>100</td>
</tr>
<tr>
<td>Growth Rate</td>
<td>12</td>
</tr>
<tr>
<td>Dense Blocks</td>
<td>4</td>
</tr>
<tr>
<td>Compression Factor</td>
<td>0.5</td>
</tr>
<tr>
<td>Batch Size</td>
<td>128</td>
</tr>
<tr>
<td>Initial Learning Rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Training Epochs</td>
<td>50</td>
</tr>
</tbody>
</table>

Figure 1: The pivotal Parameters of DenseNet for Face2Multi-modal in details.

**Visualization.** This component first acquires the results which are predicted labels from the Neural Network model. All the predicted labels are single column vectors that contain zeros and ones. To visualize the results, predicted labels are transformed into numerical results. In the end, all the results are displayed on the screen. OpenCV is used to perform these steps.

## 3 RESULTS

Figure 2 which displays three predicted results is an auxiliary understanding for the basic functionalities of three models. After the initial hyperparameter adjustment and optimization, the accuracy for estimating skin conductance, vehicle speed and heart rate are 83.78%, 59.89%, and 58.60% respectively. Estimating drivers’ skin conductance is the most accurate one in this case, the reason for it might be that the skin conductance changes slightly during the whole driving procedure. The details of the test accuracy of three models for each training epoch are illustrated in Figure 3.

The accuracy of the **Face2Multi-modal** might not reach the level of commercial use, but it does show a promising way to acquire drivers’ multi-modal status. There are several reasons for these results: first, the input of the Neural Network is unprocessed 224x224px facial captures. If more features were extracted and input to the Neural Network (e.g. temporal information), the accuracy would reach a higher level[2]. Correspondingly, the Neural Network model should be modified to take the temporal information, for example, applying a Recurrent Neural Network (RNN) layer[12]. Second, illumination variance in the BROOK database could harm the accuracy of the model[3]. Both spatial and temporal illumination variance has occurred in the BROOK database which would result in the faulty allocation of pixel values of the skin. Although an approach that takes the background region of each picture as a reference is widely used to reduce the effects, illumination variance is still an obstacle for estimating status by facial captures[10].
Vehicle Interaction system in which equipping multiple biosensors while monitoring systems such as Holter monitors or mobile poly-ments. Besides, task-specific hardware should be designed to meet the require-
ments. In the context of the application, the train set is never used. In addition, the curve of test accuracy does not rise steadily all the time (e.g. during phase 1 highlighted in Figure 3, the test accuracy would rise sharply in several epochs). Further training would find out how many epochs are suitable for each status.

4 DISCUSSIONS
While monitoring systems such as Holter monitors or mobile polygraphs can be used in research projects, the application of biosignal monitoring in production cars is limited by the lack of practical and user-friendly solutions for integration of biosensors in vehicle[6]. Face2Multi-modal is aiming to solve this problem by releasing the burden of wearing multiple sensors. It could be the alternative data source for the personalized innovative user-friendly Human-Vehicle Interaction system in which equipping multiple biosensors is not user-friendly for driving tasks.

There are many applications for biosignals, not only could these applications determine drivers’ stress level, but it could also ensure that the driver is at a stable state to perform the driving tasks[5,7,17]. Besides, Face2Multi-modal has gone beyond predicting biosignals, it could also predict the vehicle speed. We believe that this is related to the minor changes in drivers’ facial expressions in different velocity conditions.

In the application level, the Face2Multi-modal is trained on an existing database by PC. In real-life driving conditions, the cockpit would not have enough space to integrate the hardware, therefore task-specific hardware should be designed to meet the requirements.

In the security aspects, the use of the Face2Multi-modal might raise several privacy concerns. A webcam is assembled to capture drivers’ facial expressions, the estimated multi-modal status would be sent to the HVI system for further uses. Both facial images and status could have potential leaks. Typical approaches to protecting driver’s privacy include blacking out or blurring driver’s faces[11]. These approaches would make a trade-off between the level of protection and accuracy of the system.

5 CONCLUSION & FUTURE WORK
In this paper, we present a Work-In-Progress driver’s multi-modal status estimator Face2Multi-modal. This prototype shows a promising way to estimate drivers’ heart rate, skin conductance, and vehicle speed through facial expressions only. The system details and the evaluation of the estimation are provided simultaneously. Eventually, We also discussed the limitations of the prototype and approaches to improve it.

Our future work would aim to increase the accuracy of the Face2Multi-modal from several aspects. Currently, the training data is collected on simulated driving tasks rather than real-world driving tasks. Hence, creating a more realistic driving scenario would make drivers’ facial reactions more authentic, combining with the algorithm to reduce the effect of illumination variance would create a database with higher quality[18]. Additionally, even though DenseNet shows its effectiveness, exploring more a sophisticated Neural Network model or applying different layers to the current model and trying to input temporal data are promising ways to improve accuracy.

We hope Face2Multi-modal could inspire new ideas and stimulate more outstanding contributions to the field of Computer Vision and Human-Vehicle Interaction. Not only that, we believe that Face2Multi-modal has application prospects in our daily life.

ACKNOWLEDGMENTS
We thank for the anonymous reviewers from AutomotiveUI’20 Work-In-Progress Track and all members of User-Centric Computing Group for their valuable and insightful feedbacks. An earlier version, to present a comparative study between different models, has appeared in ICCAI’20 [21]. This project is a part of the BROOK project from the User-Centric Computing Group in the University of Nottingham Ningbo China [14].

REFERENCES


