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
scyrl1@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 biosignal-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/.

CCS CONCEPTS

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

KEYWORDS

Human-Vehicle Interactions; Computer Vision; Ergonomics.

ACM Reference Format:
Zhentao Huang, Rongze Li, Wangkai Jin, Zilin Song, Yu Zhang, Xiangjun Peng, and Xu Sun. 2020. Face2Multi-modal: In-vehicle Multi-modal Predictors via Facial Expressions. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20 Adjunct), September 21–22, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3409251.3411716

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

[*] 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

AutomotiveUI '20 Adjunct, September 21–22, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8066-9/20/09.
https://doi.org/10.1145/3409251.3411716

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 statues. 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 could reduce execution time and memory storage for each image. For the hyperparameter adjustments, after a fair number of attempts, the initial learning rate is set to 0.1 and is divided by 10 at 50% and 75% of the total number of training epochs. More details about the hyperparameter setting of Densenet are provided in Table1. We use BROOK which is a public multi-modal database with facial video records as the training and validation dataset[14]. The dataset contains 22 driver's facial videos labeled with heart rate, skin conductance, and vehicle speed. We split the training set, test set, and validation set in a ratio of 8:1:1 followed the recommended settings. The PyTorch is used for the implementation of the Neural Network model[13].
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.

Parameters	Value
Depth/Layers	100
Growth Rate	12
Dense Blocks	4
Compression Factor	0.5
Batch Size	128
Initial Learning Rate	0.1
Training Epochs	50

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

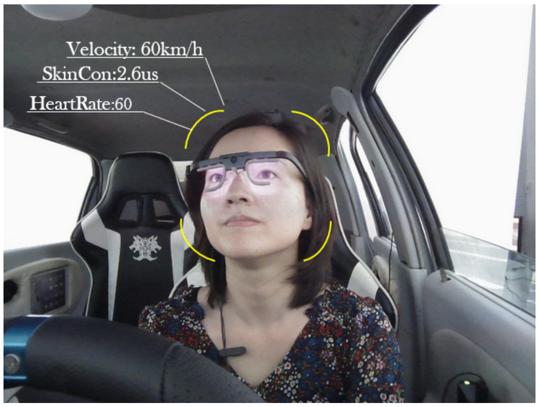

Figure 2: The estimating process of the current Face2Multi-modal. The SkinCon is referred to skin conductance.

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].

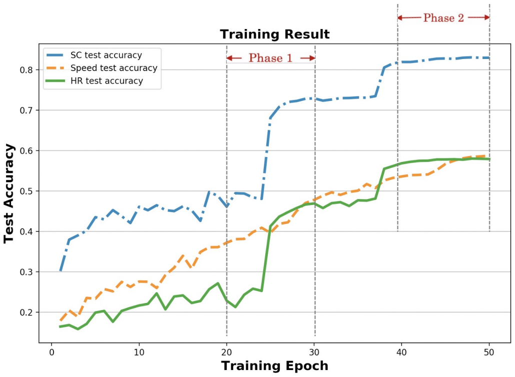

Figure 3: An Overview of the Test Accuracy between skin conductance(blue), heart rates(green) and Vehicle Speed(orange).

The training process contains 50 epochs. This is because from epoch 40 to 50 which has highlighted as phase 2 in Figure 3, the test accuracy has stopped increasing. Another reason for this is that too much training would result in overfitting[19]. Overfitting means that the Neural Network model is over-trained, it reaches very high accuracy in the train set and relatively low accuracy in the test set. We mainly consider the accuracy of the test set, because 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

[1] J. Choi, K. Kim, D. Kim, H. Choi, and B. Jang. 2018. Driver-adaptive vehicle interaction system for the advanced digital cockpit. In *2018 20th International Conference on Advanced Communication Technology (ICACT)*. 307–310.

[2] G. Du, T. Li, C. Li, P. X. Liu, and D. Li. 2020. Vision-Based Fatigue Driving Recognition Method Integrating Heart Rate and Facial Features. *IEEE Transactions on Intelligent Transportation Systems* (2020), 1–12.

[3] Mohamed Abul Hassan, Aamir Saeed Malik, David Fofi, Naufal Saad, Babak Karasfi, Yasir Salih Ali, and Fabrice Meriaudeau. 2017. Heart rate estimation using facial video: A review. *Biomedical Signal Processing and Control* 38 (2017), 346–360.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[5] Jennifer A Healey and Rosalind W Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems* 6, 2 (2005), 156–166.

[6] Stephan Heuer, Bhavin Chamadiya, Adnene Gharbi, Christophe Kunze, and Manfred Wagner. 2010. Unobtrusive in-vehicle biosignal instrumentation for advanced driver assistance and active safety. In *2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, 252–256.

[7] Stefan Hoch, Manfred Schweigert, Frank Althoff, and Gerhard Rigoll. 2007. The BMW SURF project: A contribution to the research on cognitive vehicles. In *2007 IEEE Intelligent Vehicles Symposium*. IEEE, 692–697.

[8] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[10] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. 2014. Remote Heart Rate Measurement From Face Videos Under Realistic Situations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[11] S. Martin, A. Tawari, and M. M. Trivedi. 2014. Toward Privacy-Protecting Safety Systems for Naturalistic Driving Videos. *IEEE Transactions on Intelligent Transportation Systems* 15, 4 (2014), 1811–1822.

[12] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

[13] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[14] Xiangjun Peng, Zhentao Huang, and Xu Sun. 2020. Building BROOK: A Multimodal and Facial Video Database for Human-Vehicle Interaction Research. In *the 1st Workshop of Speculative Designs for Emergent Personal Data Trails: Signs, Signals and Signifiers, co-located with the 2020 CHI Conference on Human Factors in Computing Systems, (CHI), Honolulu, HI, USA, April 25-30, 2020.* arXiv, 1–9. arXiv:2005.08637 https://arxiv.org/abs/2005.08637

[15] Kari Pulli, Anatoly Baksheev, Kirill Kornyakov, and Victor Eruhimov. 2012. Realtime computer vision with OpenCV. *Commun. ACM* 55, 6 (2012), 61–69. https://doi.org/10.1145/2184319.2184337

[16] Antoine Raux, Ian Lane, and Rakesh Gupta. 2016. System and method for multimodal human-vehicle interaction and belief tracking. US Patent 9,286,029.

[17] Eike A Schmidt, Willhelm E Kincses, Michael Scharuf, Stefan Haufe, Ruth Schubert, and Gabriel Curio. 2007. Assessing drivers' vigilance state during monotonous driving. (2007).

[18] Zilin Song, Shuolei Wang, Weikai Kong, Xiangjun Peng, and Xu Sun. 2019. First attempt to build realistic driving scenes using video-to-video synthesis in OpenDS framework. In *Adjunct Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2019, Utrecht, The Netherlands, September 21-25, 2019*, Christian P. Janssen, Stella F. Donker, Lewis L. Chuang, and Wendy Ju (Eds.). ACM, 387–391. https://doi.org/10.1145/3349263.3351497

[19] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[20] Xiaohua Sun, Honggao Chen, Jintian Shi, Weiwei Guo, and Jingcheng Li. 2018. From HMI to HRI: Human-Vehicle Interaction Design for Smart Cockpit. In *Human-Computer Interaction. Interaction in Context*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 440–454.

[21] Yaohua Wang, Zhentao Huang, Rongze Li, Zheng Zhang, and Xu Sun. 2020. A Comparative Study of Speculative Retrieval for Multi-modal Data Trails: Towards User-friendly Human-Vehicle Interactions. In *Proceedings of the 2019 6th International Conference on Computing and Artificial Intelligence, ICCAI 2020, Tianjin, China, April 23-26, 2020.* ACM.