A comparative study of speculative retrieval for multi-modal data trails: towards user-friendly Human-Vehicle interactions

Yaohua Wang, Zhentao Huang, Rongze Li, Zheng Zhang, Xu Sun, Xinyu Yin, Min Luo

University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo, 315100, Zhejiang, China.

First published 2020

University of Nottingham
UK | CHINA | MALAYSIA

A Comparative Study of Speculative Retrieval for Multi-modal Data Trails: Towards User-friendly Human-Vehicle Interactions

Yaohua Wang[1*], Zhengtao Huang[2], Rongze Li[2],

Zheng Zhang[2], Xu Sun[2], Xinyu Yin[1], Min Luo[1]

ABSTRACT

In the era of growing developments in Autonomous Vehicles, the importance of Human-Vehicle Interaction has become apparent. However, the requirements of retrieving in-vehicle drivers' multi-modal data trails, by utilizing embedded sensors, have been considered user unfriendly and impractical. Hence, speculative designs, for in-vehicle multi-modal data retrieval, has been demanded for future personalized and intelligent Human-Vehicle Interaction.

In this paper, we explore the feasibility to utilize facial recognition techniques to build in-vehicle multi-modal data retrieval. We first perform a comprehensive user study to collect relevant data and extra trails through sensors, cameras and questionnaire. Then, we build the whole pipeline through Convolution Neural Networks to predict multi-model values of three particular categories of data, which are Heart Rate, Skin Conductance and Vehicle Speed, by solely taking facial expressions as input. We further evaluate and validate its effectiveness within the data set, which suggest the promising future of Speculative Designs for Multi-modal Data Retrieval through this approach.

KEYWORDS

Human-Vehicle Interaction, Facial Recognition, Multi-modal Data Streams

ACM Reference Format:

Yaohua Wang, Zhentao Huang, Rongze Li, Zheng Zhang, Xu Sun Xinyu Yin, and Min Luo. 2020. A Comparative Study of Speculative Retrieval for Multi-modal Data Trails: Towards User-friendly Human-Vehicle Interactions. In *ICCAI '20: ACM International Conference on Computing and Artificial Intelligence, April 23–26, 2020, Tianjin China.*

*1 National University of Defense Technology, 2 User-Centric Computing Group University of Nottingham Ningbo, * yhwang@nudt.edu.cn,*

1 INTRODUCTION

Human-Vehicle Interaction is a rapidly growing sub-discipline of Human-Computer Interactions. More recently, when Autonomous Vehicles comes into daily life, the demand for intelligent and personalized Human-Vehicle Interaction systems becomes intense. Since in-vehicle drivers become free from performing necessary driving behaviors manually. For instance, long-term characterizations of driving styles have already attracted attentions to build intelligent and personalized prototypes in the future [8,11–14].

However, all previous work already assume that Human-Vehicle Interaction systems have already assured there are already constantly reliable data streams to feed in, but we believe this understudied issue would greatly endanger the practicality when attempting to adapt those novel designs. The major issue is that the user unfriendliness could be the key factor, which could forfeit the original design purpose of those systems.

Hence, the domain-specific data support, as an essential part, has bred a large number of advances in Data Science and Computer Vision, e.g. ImageNet [1], CIFAR [6, 7], Kinetics [4] and DEAP [5]. All those databases have greatly imposed and supported the validations of different designs/prototypes in many disciplines, by serving as the source information for system developments.

To this end, we carried out a comparative study to examine the feasibility for a speculative data retrieval for in-vehicle drivers' multi-modal data trails. We first perform a comprehensive data collection procedure, through ten-month user study, to over-engineer and enrich the data source as much as possible. We have also collected multiple other categories of data for future work extensions.

Later, based on the support of the above data set we collected, we further develop a unified predictor through Convolution Neural Networks. Such predictor could supply predicted multi-modal data streams in real time, by taking only Facial Captures as the input. More specifically, we have already supported the prediction for three dimensions of multi-modal states, including Heart rates, Skin Conductance and Vehicle Speed.

Finally, we have evaluated the effectiveness of all involved data models through a proper partition of the collected data set. And the results have shown that our predictor could achieve 57.90% for Heart Rate, 82.96% for Skin Conductance and 58.75% for Vehicle

Figure 1: An Example to Illustrate Data Collection Procedure.

Figure 2: Another Example to Illustrate Data Collection Procedure.

Speed respectively. These results suggest the promising future of speculative designs for multi-modal data retrieval through our approach, and we discuss the future challenges to encourage more variants to be built in the context of Human-Vehicle Interaction.

More specifically, this paper makes the following three key contributions:

- We elaborate key designs and details of our data collection procedure among 27 participants, which aims to integrate as many categories of data streams as possible. We also collect context-based feedbacks through questionnaires, which are left to the future work.
- We explore the design space and build a unified prototype, as a speculative design for in-vehicle multi-modal retrieval, to predict drivers' related data streams in real time. Particularly, we trained three DenseNet models to support the prediction of the corresponding multi-modal values using this data set.
- We quantitatively evaluate the accuracy of these three corresponding Convolution Neural Networks, which supports the prediction of relevant data streams, within the collected data set. Our predictor could achieve 57.90% for Heart Rate, 82.96% for Skin Conductance and 58.75% for Vehicle Speed.

2 DATA COLLECTION PROCEDURE

In this section, we present details about how we collect relevant data to form our own data set. As Figure 1 and 2 shown, the examples of our experimental study have been showcased. This section would be organized as follow. We first perform the recruitment of participants and the necessary training before the experiments. Then We introduced details about our study procedure, including the setup, environmental factors and equipped devices. Finally, we perform a context-based study to collect relevant feedbacks, which we leave them for the future work.

In total, 27 recruited volunteers and they have filled an Internet questionnaire, to collect information about their driving life and their preference of driving styles. Then, we conducted a training course for every participant, by providing details about the process of the experiment. Meanwhile, we explained in detail the differences between different driving styles to help participants to choose their most suitable driving style, from their own perspective.

After all preparations for both participants and experiments, each participant took about one and a half hours to finish the experiment through seven scenarios, which consists of four manual driving scenarios and three automated driving scenarios. During the procedure, each participant was equipped with an Eye Tracker, a Heart Rate Sensor and a Skin Conductance Sensor. We also have placed a camera to record their facial videos, which has contributed to the data set as an alternative data source.

During the bulk of this experiment, each participant would be give a period to understand and get familiar with the usages of the instruments within the simulator, such as the steering wheel, throttle, brakes, the buttons of left and right turn signals and the speakers. Also, each participant would perform a preliminary experiment to familiarize with the signs that appear in the scenarios, such as speed limits, turns, destinations. We finally ended up with a data set with 27 participants, which are scratched in Table 1.

Data Types	Size
Facial Video	217 GB
Heart Rates	10.4 MB
Skin Conductance	38.4 MB
Eye-tracking	21.6 GB
Driving Status	48.7 MB

Table 1: Details about the collected data set.

The composition of manual and automated driving scenarios lead to a major gap, which was bridged by a follow-up investigation. The key bridge between the manual and automated driving scenarios is to be assured by two alternative questionnaires, and this is for us to determine which automated conditions shall be deployed after.

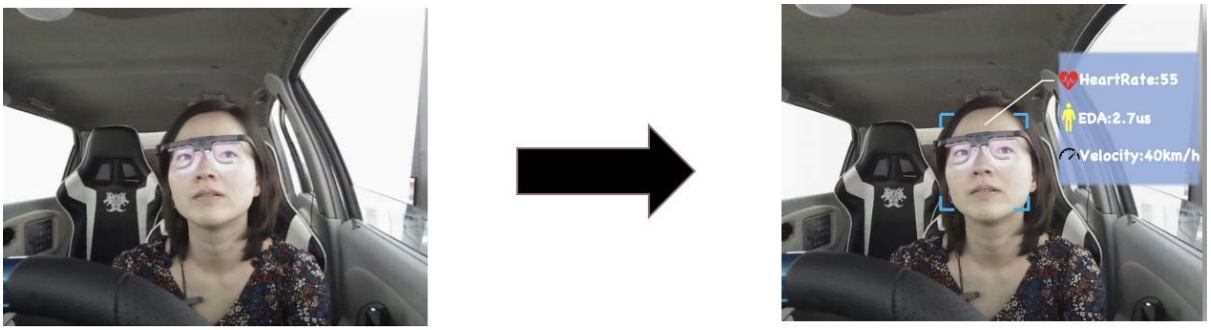

Figure 3: A Conceptual Map about *Face-to-Multimodal* system, which takes the left image as the input and presents the output as shown in the right image.

Also, in order to collect relevant driving status data, We have deployed and modified OpenDS [2, 10], a open-source cognitive driving simulator, for driving scenario generations and maintenance. Based on our built-in framework, we have confirmed that our modifications didn't affect the original functionality of the simulator, but provide better accuracy of data streams. Also, we utilise an alternative software to coordinate sensor data output for storage.

3 FACE-TO-MULTIMODAL SYSTEM DESIGN

In this section, we would explain how we build a unified engine called *Face-to-Multimodal*, which predicts the multi-modal data stream by solely leveraging the facial video of the driver as the input. We first explore its front-end design, and then we move to the backbone of our overall pipeline. Next, we illustrate how we visualize the prediction results. Finally, we reveal several key implementation details during the development.

3.1 Face Detector and Retrieval

To build up an streaming system, the capability, to capture the facial expression and make necessary adjustments for further processing, is essential. Hence, we first need to design the Face Detector and Retrieval to achieve this at high performance.

After multiple frames, captured by the camera, have been assembled and buffered in the vehicle cab for a continuous record of the facial video, the video stream would be first cropped into consecutive frames, and then they would be resized into a fixed size for Neural Network Prediction (i.e. 224x224px). We carried out our implmentation through OpenCV library [9].

3.2 Neural Network-driven Predictor

For data-driven prediction, we utilize Convolution Neural Network models to support such functions. Hereby, we specialize our Convolution Neural Network-based Image Classification model as the backbone of *Face-to-Multi-modal*. Particularly, we choose DenseNet as the model architecture because its hyperparameter adjustment doesn't require too much efforts [3]

Although the original settings of DenseNet is suitable for traditional image classification tasks (e.g. distinguish different objects), this might not be as effective as well, in the context of predicting driver's information. Therefore, we apply some lightweight changes in model settings to adapt our data set and we elaborate a little bit more later.

We have devoted ourselves to improve the algorithm and increase the levels of generalization within the above models, in order to utilize the data set more effectively and obtain higher levels of accuracy. There are two aspects to be highlighted. Firstly, the time flow of our data set could be considered as an important factor in the accuracy of estimator, for better adaptability of all these models. Secondly, we could amplify the range of this data set, to improve the accuracy of predictions, by setting landmarks or using a spatial-temporal representation to encrypt the relevant signals (e.g. Heart Rate) from multiple range of interest(ROI) volumes as the input, and make the models involve more valuable and intuitive combinations of these information.

3.3 Visualization Support

The right part of Figure 3 displays the final visualized version of our unified system. As shown in Figure 3, all three predicted values have showcased a brief understanding about the basic functionalities of our overall system design.

The entire procedure is achieved by three key steps: First, the system takes advantage of the forward propagation algorithm on trained Neural Network model to obtain the prediction results.

Second, they system utilizes the certain function to transform predicted results (i.e. labels) into concrete output as numerical. Third, we display all the results with detailed annotations on the output.

3.4 Implementation Efforts

We reveal key implementation efforts, particularly in the context of Neural Network Training. Specifically, we focus on the adaption and adjustment of DenseNet in this case. For the hyperparameter adjustments, 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. And more details about hyperparameter setting of DenseNet are provided in Table 2.

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

Table 2: Detailed Parameters about DenseNet in this case.

4 EVALUATION RESULTS

In this section, we elaborate our evaluation procedure of all involved system components, in the context of accuracy and latency. We first analyze the accuracy performance, then we perform the latency evaluation. We first describe all three models' accuracy, and then explain its latency performance.

4.1 Accuracy Overview

We first look into the test accuracy and training loss of DenseNet-101 on Heart Rate. Our trained model has achieved 57.90% accuracy in the terms of the prediction in Heart Rate, and presents a relatively sharp training loss, with the increase of training epochs.

We then examine the test accuracy and training loss of DenseNet-101 on Skin Conductance. As for the prediction in Skin Conductance, our trained model has reached 82.96% accuracy, and has a similarly sharp training loss when the number of training epochs grows.

We finally explore the test accuracy and training loss of DenseNet-101 on Vehicle Speed. Towards the Vehicle Speed prediction, 58.75% accuracy has been achieved through our trained model. However, the trend of training loss has not been as sharp as the previous two, when the number of epochs increases.

4.2 Latency Overview

One of the most important characteristics, in the Face-to-Multimodal prototype, should provide instant reflections, in order to feed into more complicated Human-Vehicle Interaction systems, as their input sources. So we evaluate the latency of the overall system by presenting such a system towards all participants in the experiment.

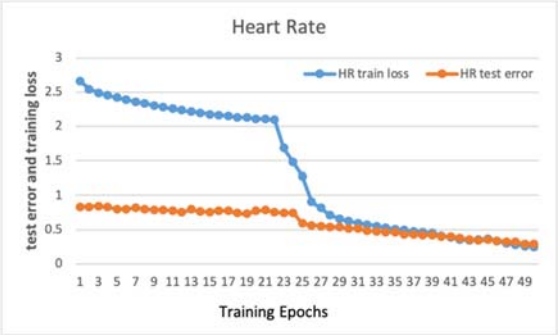

Figure 4: The test accuracy and training loss of DenseNet-101 on Heart Rate.

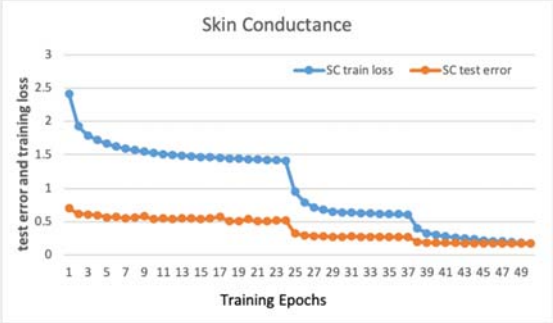

Figure 5: The test accuracy and training loss of DenseNet-101 on Skin Conductance.

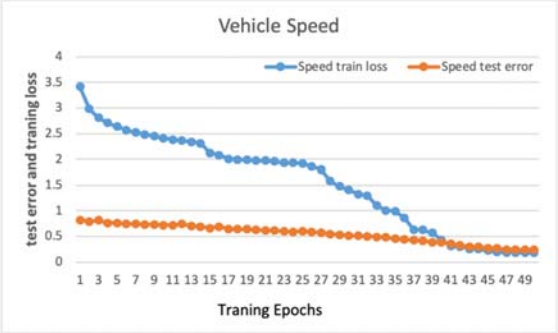

Figure 6: The test accuracy and training loss of DenseNet-101 on Vehicle Speed.

Interestingly, all participants have satisfied with the instant reflections from our Face-to-Multimodal prototype, where they have compared our prototype and the relevant sensor data flows. Also, we validated the latency variation and we found that there are only little differences between different data streams.

5 DISCUSSIONS

Our Face-to-Multimodal shows a promising way to enhance the procedure of Human-Vehicle Interaction (HVI). By estimating driver's status, we present an alternative data source of all Human-Vehicle Interaction Systems in the future. This provides an opportunity for them to gain more dimensions of perspectives during the decision-making procedures.

There are at least two directions for future developments. First, it provides a speculative approach to monitor the health status of drivers and provides necessary reflections for their safety. Second, it could serve as the data source for automated and personalized recommendation to perform relevant services for drivers' instant or long-term preferences, based on their facial expressions only.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we perform a comparative study of speculative retrieval designs for in-vehicle multi-modal data streams. To examine its feasibility and enable comparisons, we first collect relevant data streams through a comprehensive data collection procedure from 27 participants. Then, we built an unified system to predict multi-modal statuses, by taking only facial records as input. Later, we evaluate this design and show impressive accuracy in terms of prediction and acceptable time of reflections.

Our future work would focus on deepening the understanding of facial expressions to improve the accuracy of our prototypes, attempts to combine context-based feedbacks with numerical predictors and emerge them with more novel Human-Vehicle Interaction Systems.

7 ACKNOWLEDGMENTS

We thank for valuable suggestions and feedbacks from all members of User-Centric Computing Group and the reviewers from ICCAI'20. This research is supported by ported by The Science and Technology Planning Project of Hunan Province (2019RS2027).

REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. 248–255. https://doi.org/10.1109/CVPRW.2009.5206848

[2] Paul A. Green, Heejin Jeong, and Te-Ping Kang. 2014. Using an OpenDS Driving Simulator for Car Following: A First Attempt. In *Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Seattle, WA, USA, September 17 - 19, 2014*. 4:1–4:6. https://doi.org/10.1145/2667239.2667295

[3] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[4] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *CoRR* abs/1705.06950 (2017). arXiv:1705.06950 http://arxiv.org/abs/1705.06950

[5] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Trans. Affective Computing* 3, 1 (2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[6] Alex Krizhevsky. 2009. CIFAR: 80 million Images Datasets. http://www.cs.toronto.edu/~kriz/cifar.html.

[7] Alex Krizhevsky. 2010. Convolutional Deep Belief Networks on CIFAR-10. In *Technical Report in University of Toronto*. 1–9. http://www.cs.toronto.edu/~kriz/conv-cifar10-aug2010.pdf

[8] Xiaohan Li, Wenshuo Wang, and Matthias Roetting. 2019. Estimating Driver's Lane-Change Intent Considering Driving Style and Contextual Traffic. *IEEE Trans. Intelligent Transportation Systems* 20, 9 (2019), 3258–3271. https://doi.org/ 10.1109/TITS.2018.2873595

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

[10] OpenDS Development Team. 2017. OpenDS - the Flexible Open Source Driving Simulation. https://opends.dfki.de/.

[11] Hanneke Hooft van Huysduynen, Jacques M. B. Terken, Jean-Bernard Martens, and Berry Eggen. 2015. Measuring driving styles: a validation of the multidimensional driving style inventory. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2015, Nottingham, United Kingdom, September 1-3, 2015*. 257–264. https://doi.org/10.1145/2799250.2799266

[12] Eric Vasey, Sangjin Ko, and Myounghoon Jeon. 2018. In-Vehicle Affect Detection System: Identification of Emotional Arousal by Monitoring the Driver and Driving Style. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2018, Toronto, ON, Canada, September 23-25, 2018*. 243–247. https://doi.org/10.1145/3239092.3267417

[13] Wenshuo Wang, Junqiang Xi, and Ding Zhao. 2019. Driving Style Analysis Using Primitive Driving Patterns With Bayesian Nonparametric Approaches. *IEEE Trans. Intelligent Transportation Systems* 20, 8 (2019), 2986–2998. https://doi.org/10.1109/TITS.2018.2870525

[14] Nidzamuddin Md. Yusof, Juffrizal Karjanto, Jacques M. B. Terken, Frank Delbressine, Muhammad Zahir Hassan, and Matthias Rauterberg. 2016. The Exploration of Autonomous Vehicle Driving Styles: Preferred Longitudinal, Lateral, and Vertical Accelerations. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2016, Ann Arbor, MI, USA, October 24-26, 2016*. 245–252. https://doi.org/10.1145/3003715.3005455