The Importance Distribution of Drivers’ Facial Expressions Varies over Time!

Conference Paper · September 2021
DOI: 10.1145/3473682.3480283

6 authors, including:

Jiahao Wang
University of Nottingham Ningbo China
2 PUBLICATIONS 2 CITATIONS
SEE PROFILE

Zeyu Xiong
University of Nottingham Ningbo China
6 PUBLICATIONS 6 CITATIONS
SEE PROFILE

Yicun Duan
University of Nottingham Ningbo China
2 PUBLICATIONS 2 CITATIONS
SEE PROFILE

Junyu Liu
University of Nottingham Ningbo China
3 PUBLICATIONS 2 CITATIONS
SEE PROFILE

All content following this page was uploaded by Zeyu Xiong on 27 September 2021.
The user has requested enhancement of the downloaded file.
The Importance Distribution of Drivers’ Facial Expressions Varies over Time!

Jiahao Wang
User-Centric Computing Group, University of Nottingham Ningbo
China

Zeyu Xiong
User-Centric Computing Group, University of Nottingham Ningbo
China

Yicun Duan
User-Centric Computing Group, University of Nottingham Ningbo
China

Junyu Liu
User-Centric Computing Group, University of Nottingham Ningbo
China

Zilin Song
Department of Computer Science, The University of Sheffield
England

Xiangjun Peng
Department of Computer Science and Engineering, The Chinese University of Hong Kong
China

ABSTRACT

Facial Expressions are valuable data sources for advanced Human-Vehicle Interaction designs. However, existing works always consider the whole facial expressions as input, which restricts the design space for detailed optimizations. In this work, we make the hypothesis that facial expressions can exhibit significant variations during the driving procedure. Our goal in this work-in-progress is to justify this hypothesis, by performing detailed characterizations on the drivers’ facial expressions. To this end, we leverage Local Binary Fitting, a novel mechanism for selecting representative feature points from facial images on the fly, for our characterizations. Our characterizations reveal that, among six major components of facial feature points, there are significant variations of correlations with a certain vehicle status (i.e. Vehicle Speed), in terms of (1) the time spots during the driving procedure; and (2) the gender of the drivers. We believe our works can serve as a starting point to incorporate the characteristics of our findings with a great amount of adaptive and personalized Human-Vehicle Interaction designs.

ACM Reference Format:

1 BACKGROUND AND MOTIVATION

There is a growing amount of research interests to improve the applicability of Human-Vehicle Interaction designs by incorporating data-driven techniques. Recent works show that video data is a valuable resource to estimate in-vehicle drivers’ statistics (i.e. Face2Multi-modal) [5]. Such an approach is more accessible and intuitive. Though video data contains a huge volume of information, the state-of-the-art approaches still take the whole facial expression as the input: the design takes whole facial expressions usually focus on the overall facial changes, while ignoring the changes in local details. This incurs two major limitations.

First, the whole face’s information of drivers is highly combined and encapsulated, and in different situations, the changes in a particular part of the face will be much more drastic than the overall. For instance, when drivers are fatigued, they will blink more, increase eye movement and yawn frequently [7], which can cause significant and regular changes in the eyes and mouth parts. While for other parts where the changes are relatively unobvious, they may waste part of the computing power and even prevent the program from reaching correct judgments. So, in this case, performing local facial detection will be more effective compared with the whole facial expressions. Second, due to some irresistible factors (i.e. gender [3], age, driving styles [13]) and etc., the overall facial expressions of a single group may hardly be used to characterize other groups. For example, they will react differently to the same situation and affect facial expressions. However, when taking the whole facial expression, such difference may cause predication errors under the mutual influence of different facial parts.

To this end, we envision that the needs to understand these information are both essential and beneficial. All these restrictions outlined above can cause a lot of valuable information to be overlooked in practice, and ultimately lead to relatively poor results, which substantially restricts the overall design space. To address this issue, we perform detailed characterization in the driver’s facial expressions in this work to show that local facial features are strongly correlated with the state of the vehicle when driving. Based on the hypothesis that facial expressions can exhibit significant variations during the driving procedure, we take Local Binary Fitting (LBF) in our characterization procedure and group facial feature points into six major components (forehead, eyes, eyebrows, nose, mouth and jaw). Our results reveal that, by comparing the changes in the number of each component during the driving process, different components have different correlation degrees with the state of driving the vehicle and this correlation of the same part will also vary according to the gender of the driver.
2 OUR METHODOLOGY

Our methodology behind this work will be shown in this section. Figure 1 comprehensively provides an execution pipeline to figure out how the selected types of facial expression are distributed with the change of vehicle speed and there are three main steps. The first step is the preprocessing of facial expressions, which aims to replace the whole face expressions with feature points. The second step is feature extraction which clusters feature points into preset 6 categories through Local Binary Fitting. And in the final step, the distribution results of each type after clustering will be displayed.

2.1 Step 1: Preprocessing of Facial Expressions

The preprocessing of facial expressions step consists of two components: acquisition and normalization. For the acquisition, the core part is face detection. As the input images usually contain the whole facial expressions and other irrelevant graphics which have some negative effects when clustering, to better perform feature extraction, Face Mesh [2] is used to locate and intercept the whole facial expressions from input images. This technique takes advantage of a comprehensive lightweight face detector which produces face bounding rectangles and several landmarks of basic feature area (e.g. eyes, ears, and noses). The landmarks are used to rotate a facial rectangle to align the eye centers with the horizontal axis of the rectangle [8]. Such a face detector will output the modified images which contain 468 points arranged in fixed quads (figure 2). The areas that are evaluated to possess higher variability and higher status in human perception will be allocated with higher point density [9]. As a result, we can build a reasonable smooth surface representation of the face and depict the entire facial expression in the form of points.

Normalization is also not negligible for preprocessing and during this procedure, we re-edit all the input images in the dataset with the same configuration. The entire experiment is carried out based on the Brook [10], which contains 34 drivers’ multi-modal/driving status data in 11 dimensions, including facial videos, vehicle acceleration, steering wheel coordinates and so on. In particular, the relationship between vehicle speed and facial feature points is explored since vehicle speed is considered as one of the most intuitive external environmental change factors. So, we collect images from facial videos along with driving status of speed data by time frame from Brook. Specifically, we divide speeds ranging from 0 to 100 uniformly into 10 intervals and drivers into 2 group. for each image, we locate the position of the face in the input image, and then transform it into a 400 × 400 RGB image in size.

2.2 Step 2: Feature Extraction

During this step, the 468 scattered points of the whole facial expression obtained from step 1 are classified into 6 components by a customized neural network. In the shallow layer of this model, we adopt the LBF to realize Normalized Pixel Difference, which can be simply explained as the gray scale difference between two pixels. The reasons for this design are as follows: By taking the whole face as the coordinate system, the coordinates of the 468 points of the face conform to the probability distribution, and the residuals between the points predicted by the coordinate points can be calculated. Then, by minimizing the residuals, a random forest can be obtained to quickly realize Face Alignment which achieves that each face components information can be accurately located and analyzed based on the key points. Furthermore, because points’ relative position at the edge of the image can implicitly provide valuable information, when such a receptive field reaches the image boundary, it implicitly depends on the model in the input image. So, to eliminate the influences of those unrelated points, the key points are re-positioned so that a fixed aspect ratio is maintained between the span of x coordinates and the span of z coordinates, i.e., a face that is scaled to half its size has its depth range (the nearest to farthest) scaled down by the same multiplier. More specifically, the x and y coordinates of the vertices correspond to the point locations in the 2D plane as given by the image pixel coordinates. The z coordinates are interpreted as the depth relative to a reference plane passing through the mesh’s center of mass [4]. In other words, this model, a customized neural network architecture, uses a more aggressive sub-sampling network in the early layers and uses most of the calculations in its shallow parts. As a result, the
neuron receptive field begins to cover a large area of the input image relatively early to achieve face alignment and deeper neurons can specifically distinguish different features of the image through CNN [1].

In the deeper layer, based on the vector of 3D landmark coordinates, which subsequently gets mapped back into the original image coordinate system, CNN will limit the range of x, y and z coordinates to group all feature points into 6 preset categories (forehead, eyes, eyebrows, nose, mouth and jaw).

2.3 Step 3: Output

This step aims to unify each category feature points through linear regression and obtain the final feature points of each type. We divide the number of each type of feature points by that of the total feature points to get percentage, which is used to identify the degree of change of each type feature points with the speed. Figure 2 shows a comparative example between the original figure and the estimated mesh topogy.

Figure 2: The original image and visualization of the estimated mesh topology.

3 PRELIMINARY RESULTS

We report the preliminary results of our study. We focus on the speed statistics, and use BROOK [10] as our source dataset. We apply the above methodology to carry out this study, and classify the drivers into two groups based on their gender.

The above figure 3 shows the results of the standard deviation of 6 different face components in terms of men and women. As we divided the speed into 10 groups and each group will output different percentages of the number of each category’s feature points, the standard deviation of 10 sets of data can symbolize stability of each type (forehead, eyes, eyebrows, nose, mouth and jaw) under the change of the vehicle speed, which helps us to better understand to what extend the number of each type’s feature points change with the speed. Therefore, the greater the standard deviation of a facial component, the more sensitive it is to changes in vehicle speed, which is more suitable to be used as the main data source for predicting.

Based on figure 3, we make the following two key observations:

- When choosing gender as the angle of observation, for men, the standard deviation of nose is the highest, which is 0.03610; while for women, the highest standard deviation is from eye, which is 0.03701. Therefore, in practice, applying different types of facial components for different groups (men and women) during data analyzing may lead to a better performance.
- There will be more or less differences in the six face components of men and women, especially nose. So, when using the entire facial expressions as the common input for both men and women, these differences will have some negative effects on data analysis and predicting.

4 CONCLUSIONS AND FUTURE WORKS

In this paper, we explore how the distributions of facial expressions vary over time. Based on our preliminary findings from the experimental studies, we summarize our work from two aspects:

- There is no one-size-fits-all solutions to all types of data streams. In this experiment, we pay attention to normal cases, so we limit the speeds of vehicles within the range between 0 and 100, which is suitable for most driving scenarios. However, if the specific scenario only focuses on speeds which are over 100 (i.e. driving on the highway), the dataset needs to be modified and corresponding facial points will also be changed, leading to different results.
- Through our observation, there are distinct variations of correlations in terms of the speeds of vehicles. In detail, the whole face contains various organs, which can have great impacts in terms of validation accuracy, and it’s essential to select a suitable one in driving scenarios to avoid the influence of unrelated factors. And our work can bring benefits that the number of feature points of new predictors in the future does not have obvious fluctuations due to changes in external factors such as light, vehicle speed, etc., which improves the accuracy of prediction efficiently.

We believe increasing the expressiveness of facial expressions have both opportunities and challenges. For opportunities, our future works lie on extending and deriving new insights from data-driven techniques, using detailed facial expressions: for instance, customization supports are essential for practical deployments, and there are more relations to be formed for guiding future efforts under other surrounding variables, like heart rate, skin conductivity and so on [5]. Therefore, with the application of such a method,
more various scenes can take good advantages of the particular chosen type of facial expressions to make prediction. Our study may also potentially help with more high-level user demands (e.g. user trust [12]). However, our works also reveal several technical challenges. One is that the study is based on a simulation-based in-lab study, and the results may need to be vetted in real vehicles or applying more advanced simulation techniques (e.g. in-lab view synthesis [11]). Also, applying detailed facial expressions may also increase the computational burdens, and how to leverage such findings in practice may need to rethink the overall system design and implementations (e.g. Internet-of-Vehicles [6]).

ACKNOWLEDGMENTS

We thank the anonymous reviewers from AutomotiveUI’21 Work-in-Progress Track and all members of User-Centric Computing Group for their valuable and insightful feedbacks.

REFERENCES


