DEVIATE: The Unglamorous Foundations of Machine Learning

Students: Rahmy Salman, Xin Ye, Juhyun Lee, Nuobei Zhang, Hriday Chhabria Researcher: Kathleen Klinich, PhD MDP Group: DEVIATE

UMTRI Fall 2022

MICHIGAN ENGINEERING UNIVERSITY OF MICHIGAN

Abstract

Our project is developed a ground truth labeled dataset from videos recorded of vehicle passengers with and without motion sickness. With this ground truth, we can train algorithms that could be run in an autonomous vehicle to be able to detect the passenger's motion sickness level and respond accordingly. To achieve this goal, we establish the level of deviation of the passenger from the standard posture, for example the head tilt and eye motion. In a process of iterative development, we hand-labeled the data with our rubric and refined the rubric depending on the Fleiss Kappa score. The challenge is how we could improve the rubric to eliminate the human biases in the first step of machine learning.

Introduction

Recent research has investigated how vehicle dynamics are associated with the motion sickness of passengers in a vehicle. In particular, some research is interested in the design of autonomous driving patterns that avoid motion sickness in passengers of autonomous vehicles. These studies have amassed great video data of passengers in vehicles, along with quantification of their motion sickness levels in the vehicles over a trip.

Our work is in the identification of key primitives in these videos that may be indicators of motion sickness in the participants of the study. To enable human coders to accurately identify primitives, we developed a rubric of key primitives in body posture, gaze, device use, and emotion. The relevance and accuracy of this rubric is of great foundational importance to the training of machine learning models in predicting motion sickness.

Objectives

- To generate ground-truth primitives to train a machine learning model to predict motion sickness from video elements.
- To code video-frames using our rubric to create a set of training data for the aforementioned Machine Learning Model.
- To use Fleiss Kappa score to evaluate the rubric we coded and improve the rubric to eliminate as much human biases as possible.

Methods

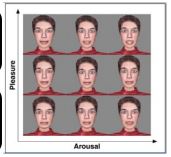
1.Rubric development and human labeling process

- Step1: Extract valuable information from video
- Step2: Decide Six primitives
- Step3: Created operational definitions and labels for each primitive
- Step4: Discussion and integration of labels
- Step5: Label videos by current rubric version
- Step6: Calculate Fleiss Kappa scores and update rubrics

2. An example: how did we develop Emotion rubric?

Emotion rubric is one of several rubrics in which progress resulted from several iterations of testing and discussions. At first, we reviewd relevant literature on emotion and gathered existing popular emotion models on expression and voice such as FACES and PADS. We also gathered existing emotion coding scales and definitions used in psychology to develop the first version emotion rubrics.

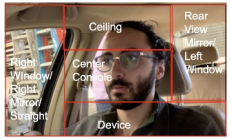
After discussion and tests, it was too complicated and redundant to include three emotion scales. Rubric 2.0 was generated by selecting emotion models which includes three dimensions and eight categories.



Rubric 3.0 narrowed down the scale of pleasure, arousal, and dominance from 5 point scale to 3 point scale based on further tests. Precise descriptions and examples are also added to reduce discrepancies amongst the labelers.

3. Gaze Calibration

Due to differences in participant height, the identification of gaze becomes a task of calibration. After collecting images of people of different heights in the test car, we created this gaze grid as a rubric aid.



Results

To test the fidelity of a coding rubric, we gathered the results of several human coders using that rubric and measured their level of agreement. If the coding task is well-defined and the rubric is well-made, high agreement is to be expected. The Fleiss Kappa test measures agreement in this setting. Thus, we employed the Fleiss Kappa test to be used as a proxy for the applicability of our rubric.

Results

Label: Manipulation

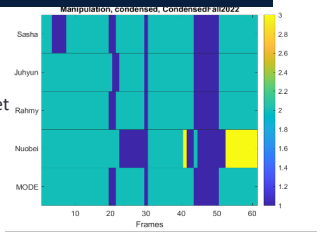
1: Yes: passenger is typing on tablet.

2: No: passenger is not typing on tablet

3: Indiscernible

Percent Overall Agreement: 0.8066

Overall Fleiss Kappa Score: =0.4828



Reject null hypothesis: Observed agreement is not accidental.

Different colors represent the different values for this label.

The X-axis represents frame numbers and the Y-axis represents labers. Each row is one labeler's result. From the visualization, we can see that different labelers have similar values for this label. Based on Fleiss Kappa, our rubric has a high agreement between labelers, which means great inter-user reliability.

FleissKappa	StdError	CI	z	p_value
4.82e-1	3.57e-2	4.64e-1	5.01e-1	1.34e+1

Conclusions

The interrater agreement achieved through usage of our rubric as measured by the Fleiss Kappa score supports the conclusion that we successfully developed our rubric to identify meaningful ground truth primitives in body posture, eye gaze, emotion, and device use in the vehicle passenger setting. Further research can now use this rubric to allow human coders to generate ground truth labels on video datasets, which can then be used to train machine learning models to predict signs of motion sickness from the videos.