Data Extraction and Processing

Joseph Blubaugh

20 October 2016

Dataset Overview

The data in this project are of 8 driving simulations for 66 individuals ranging from 3,000 to 30,000 observations per simulation. There are over 6.7 million observations in the entire dataset. The data from each simulation includes likelihood scores for 8 facial expressions recorded at a fixed interval of .03 seconds. Stimuli data which records targetted events that were introduced into each simulation and basic demographic data on each subject are also available.

Data Extraction

In order to prepare the data for analysis each of the 509 simulation files needed to be extracted and combined into a single file. A python program was witten to accomplish this. All of the simulations were stored as excel files. The python program extracted data from each simulation file beginning on the 9th (header) row. Two of the simulation files did not start on the 9th row (8th & 10th) and had to be corrected manually (see extract_faces.py on github repo).

Table 1: Simulation Sample: T001-001 (first 6 rows)

Subject	Trial	Frame	Time	Anger	Contempt	Disgust	Fear	Joy	Sad	Surprise	Neutral
T001	001	0	0	0.01	0.022	0.004	0.054	0.526	0.096	0.001	0.287
T001	001	1	0.033	0.01	0.022	0.004	0.054	0.526	0.096	0.001	0.287
T001	001	2	0.067	0.01	0.022	0.004	0.054	0.526	0.096	0.001	0.287
T001	001	3	0.1	0.008	0.019	0.003	0.038	0.535	0.105	0.001	0.291
T001	001	4	0.133	0.009	0.038	0.016	0.004	0.69	0.018	0	0.225
T001	001	5	0.167	0.01	0.045	0.014	0.003	0.716	0.016	0	0.195

There are 267 stimuli files (.stm) that record events that occur in each simulation. They have identical formats and were able to be extracted in the same manner as the simulation data. Python was also used to extract the data.

Table 2: Stimuli Sample: T001-007.stm

StartTime	EndTime	Event Switch	Action Type	Question Number
81	240	1	4	Texting
510	671	1	4	Texting

Data Quality

In addition to the two simulation files with differing formats, several misspellings occur in the event descriptions.

Table 3: Record Count by Event (original .stm data)

Action Type	Question Number	Count
1	Analytical Questinos	1
1	Analytical Questions	66
1	Annalytical Questions	1
2	Emotional Questions	2
2	Mathematcal Question	25
2	Mathematical Question	1
2	Mathematical Questions	42
3	Emotional Equestions	2
3	Emotional Quesitons	4
3	Emotional Questions	128
4	Failure Event	1
4	Texting	128
4	Texting and Talking	1
5	Testing and Talking	1
5	Texting and Emotional Questions	11
5	Texting and Talking	22
6	Failure	3
6	Failure Evemt	1
6	Failure event	5
6	Failure Event	95

It was assumed that Action Type and Question Number are directly related and that Question Number is supposed to reflect the name of the event. The Question Number was renamed Event and the descriptions were renamed based on the most frequent event per action type.

Table 4: Record Count by Event (Corrected data)

Action	Event	Count
0	No Event	509
1	Analytical Questions	65
2	Mathematical Questions	66
3	Emotional Questions	62
4	Texting	61
5	Texting and Talking	31
6	Failure Event	62

Final Dataset

An R program was written to combine the three datasets (2 .csv, 1 .xlsx). The stimuli data was joined with the simulation data based on the file name and the start and end times of the events. For each frame in a particular subject's simulation, if the frame time were between the start and end time in the stimuli (.stm) file, then every record in that interval is assumed to have the corresponding event. The stimuli files only record designed events, so records that fall outside of the event time intervals were coded as "No Event".

Table 5: Final Combined Dataset (continued below)

Subject	Trial	Age	Gender	Frame	Time	Event.Switch	Event	Action
T001	007	Y	M	2427	80.9	0	No Event	0
T001	007	Y	M	2428	80.93	0	No Event	0
T001	007	Y	M	2429	80.97	0	No Event	0
T001	007	Y	M	2430	81	1	Texting	4
T001	007	Y	M	2431	81.03	1	Texting	4
T001	007	Y	M	2432	81.07	1	Texting	4

Anger	Contempt	Disgust	Fear	Joy	Sad	Surprise
0.09088	0.05749	0.4205	0.00034	0.0011	0.1343	2e-05
0.06125	0.03971	0.4293	0.00039	0.00109	0.163	2e-05
0.1034	0.09633	0.3186	0.00018	0.0013	0.08558	1e-05
0.03631	0.4976	0.01708	6e-05	0.00235	0.00694	1e-05
0.00591	0.7285	0.00273	4e-04	0.00679	0.00635	2e-05
0.00576	0.689	0.00348	0.00038	0.00774	0.00682	3e-05

Possible Misscoding

A possible event misscoding appears when you view the simulation trials by event. Trial 006 has 62 out of 63 trials where the only event involves emotional questions, however one trial shows as having mathematical question events. This appears to be a data error based on the experimental design where mathematical and anlaytical questions are only supposed to occur in trial 005. The subject who this possible misscoding is associated with is T018-006.

Table 7: Count of Simulations by Event Type

Event	001	002	003	004	005	006	007	008
No Event	63	63	64	64	65	63	62	65
Analytical Questions	0	0	0	0	65	0	0	0
Emotional Questions	0	0	0	0	0	62	0	0
Failure Event	0	0	0	0	0	0	0	62
Mathematical Questions	0	0	0	0	65	1	0	0
Texting	0	0	0	0	0	0	61	0
Texting and Talking	0	0	0	0	0	0	0	31

Additional Comment

The final dataset does not include the possible correction to T018-006 because it is does not interfere with my proposed analysis. Anyone who uses this subject and trial should analyze further or consider omitting this simulation.

Reproducible Code

All of the findings and code used in this project can be reproduced by cloning my project repository on github: https://github.com/JestonBlu/driving