

Data Processing and Extraction

Joseph Blubaugh

15 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 simulation files are excel files and all except 2 have identical layouts. 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.

Table 1: Simulation Example: T001-001

Subject	Trial	Frame	Anger	Contempt	Disgust	Fear	Joy	Sad	Surprise	Neutral
T001	001	0	0.0101	0.0218	0.0043	0.0541	0.526	0.0959	0.001	0.2868
T001	001	1	0.0101	0.0218	0.0043	0.0541	0.526	0.0959	0.001	0.2868
T001	001	2	0.0101	0.0218	0.0043	0.0541	0.526	0.0959	0.001	0.2868
T001	001	3	0.008	0.0187	0.0032	0.0375	0.5353	0.105	0.0011	0.2911
T001	001	4	0.0091	0.038	0.0158	0.0036	0.6902	0.0177	4e-04	0.2252
T001	001	5	0.0104	0.045	0.0139	0.003	0.7157	0.0162	3e-04	0.1955

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 Example: T001-007.stm

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

The two python programs used to extract the data and a third file which contained basic demographics data were all read into R. An R program was created to

Table 3: Final Combined Dataset (continued below)

Subject	Trial	Age	Gender	Frame	Time	Event.Switch	Event	Anger
T001	001	Y	M	0	0	0		0.0101
T001	001	Y	M	1	0.0333	0		0.0101
T001	001	Y	M	2	0.0667	0		0.0101
T001	001	Y	M	3	0.1	0		0.008
T001	001	Y	M	4	0.1333	0		0.0091
T001	001	Y	M	5	0.1667	0		0.0104

Contempt	Disgust	Fear	Joy	Sad	Surprise	Neutral
0.0218	0.0043	0.0541	0.526	0.0959	0.001	0.2868
0.0218	0.0043	0.0541	0.526	0.0959	0.001	0.2868
0.0218	0.0043	0.0541	0.526	0.0959	0.001	0.2868
0.0187	0.0032	0.0375	0.5353	0.105	0.0011	0.2911
0.038	0.0158	0.0036	0.6902	0.0177	4e-04	0.2252
0.045	0.0139	0.003	0.7157	0.0162	3e-04	0.1955